Methods in Molecular Biology 2509

Springer Protocols

Nicholas F. Parrish
Yuka W. Iwasaki *Editors*

piRNA

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, UK

For further volumes:
http://www.springer.com/series/7651

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

piRNA

Methods and Protocols

Edited by

Nicholas F. Parrish

Center for Integrative Medical Sciences, RIKEN, Yokohama, Kanagawa, Japan

Yuka W. Iwasaki

Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan

Humana Press

Editors
Nicholas F. Parrish
Center for Integrative Medical Sciences
RIKEN
Yokohama, Kanagawa, Japan

Yuka W. Iwasaki
Department of Molecular Biology
Keio University School of Medicine
Tokyo, Japan

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-2379-4 ISBN 978-1-0716-2380-0 (eBook)
https://doi.org/10.1007/978-1-0716-2380-0

Cover Illustration Caption: A photomicrograph of mouse testis, provided by the lab of Dr. Nicholas F. Parrish.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

PIWI-interacting RNAs (piRNAs) are complexed with PIWI-clade Argonaute proteins and direct the activity of their protein partner against complementary nucleic acids. While an RNA-guided nucleic acid modification system from prokaryotes has led to more capital gains, studies of piRNAs in diverse eukaryotes over the past 15 years have approached a central mystery of multicellular biology, the immortality of the germline. Abundant high-quality evidence now confirms the role of piRNAs in protecting germlines from the ill-effects of transposons. Yet, as a recently published collection of papers using golden hamsters has demonstrated, knowledge based on the classical model organisms is inevitably partial, and studies using novel organisms provide new insight. Descriptions of "the" piRNA pathway, as if it were a monolith, ignore the diversity inherent in life's myriad forms. For example, substantial evidence supports a role for piRNAs in silencing virus infection in arthropod somatic tissues. This allows conceptualization of piRNA-mediated silencing as one of the mechanisms that germlines have, on occasion, loaned to their somatic "vehicles" for use in the soma's (losing) battle against invasive mobile genetic elements, defined more broadly than transposons to include exogenous viruses. Even for those organisms in which piRNAs' function to retrain transposons is most clearly tied to staving off germline mortality, we struggle to clearly answer why; insertional mutagenesis is often invoked, but it has been noted that error catastrophe in a single generation is implausible. *C. elegans* piRNA deficiency has recently been characterized as a reproductive arrest phenotype triggered by an unknown transcriptional stress signal, rather than DNA damage *per se*, that warns of piRNA dysfunction. Various innate immune pathways known for their role in recognizing virus-associated (and potentially TE-associated) patterns could fit this bill, raising the prospect of rescuing, at least temporarily, the fertility of piRNA pathway mutants in various species. However, if such an uncoupling of piRNAs from fertility is on the horizon, it seems distant. More immediately relevant—for those of us who judge relevance using our somatic tissues—is the evidence presented in the past two years that clearly shows the activity of PIWI proteins in transformed cells, albeit thus far independently of piRNAs. In summary, while the canonical biochemical functions of PIWIs and piRNAs in the germline of model organisms are extensively studied, further molecular biological investigations of piRNAs will advance several frontiers, including cancer biology, antiviral immunity, and the diverse mechanisms guarding and perpetuating germline genomes, especially those of non-model organisms.

Many foundational biochemical methods used to study piRNAs were compiled in an earlier volume of this series (*PIWI-Interacting RNAs: Methods and Protocols.* ed. Siomi M. C., 2014). Thanks in large part to the work of contributors to that volume, the enzymatic pathways and intracellular traffic involved in biogenesis of piRNAs, especially cluster-derived piRNAs, have now been mapped to impressive resolution. That is not to suggest that piRNA biogenesis is a closed book; far from it. For example, the *raison d'etre* linking piRNA biogenesis to mitochondria, and recently ribosomes, remains unclear, as does the biological meaning of tRNA-derived piRNAs. However, in considering methods to solicit for this volume, we aimed to include newly developed methods, methods currently applied to other ncRNAs involved in nuclear regulation which can be used to study piRNAs, and piRNA methods applied in nonclassical organisms. We have also included several bioinformatic and biophysical methods related to piRNA studies, consistent with the increasing importance of

high-throughput sequencing and computational methods in piRNA analysis as with other areas of biology. While the final volume fails to exhaustively cover the diversity of species in which piRNAs are now studied rigorously, we were pleased to receive several contributions from the active community studying piRNAs in mosquitos, where they are conceivably linked to vector competence and thus enormous human suffering and societal cost. We have learned a lot in the process of compiling these diverse methods, and we offer this volume hoping that they will prove useful to other workers in this field.

Yokohama, Kanagawa, Japan — *Nicholas F. Parrish*
Tokyo, Japan — *Yuka W. Iwasaki*

Contents

Contributors

Omar S. Akbari • *Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, La Jolla, CA, USA*

Masaru Ariura • *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan*

Wan Kin Au Yeung • *Division of Epigenomics and Development, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan*

Mariangela Bonizzoni • *Department of Biology and Biotechnology, University of Pavia, Pavia, Italy*

Katia Bulekova • *Boston University Research Computing Services, Information Services and Technology, Boston, MA, USA*

Mariko Dacher • *Laboratory of Chromatin Structure and Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan*

Gargi Dayama • *Boston University School of Medicine, Department of Biochemistry, Boston University Bioinformatics Program, Boston, MA, USA*

Mahmoud Dondeti • *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA*

Risa Fujita • *Laboratory of Chromatin Structure and Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan*

Tsukasa Fukunaga • *Waseda Institute for Advanced Study, Waseda University, Tokyo, Japan*

Rebecca Halbach • *Department of Medical Microbiology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands*

Michiaki Hamada • *Faculty of Science and Engineering, Waseda University, Tokyo, Japan; AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), Tokyo, Japan*

Rippei Hayashi • *The John Curtin School of Medical Research, Australian National University, Acton, ACT, Australia*

Mitsuru Ishikawa • *Department of Physiology, Keio University School of Medicine, Tokyo, Japan*

Yuka W. Iwasaki • *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan; Japan Science and Technology Agency (JST), Precursory Research for Embryonic Science and Technology (PRESTO), Saitama, Japan*

Alla Kalmykova • *Institute of Molecular Genetics of National Research Centre, "Kurchatov Institute", Moscow, Russia*

Makoto Kashima • *College of Science and Engineering, Aoyama Gakuin University, Sagamihara Chuo Ku, Kanagawa, Japan*

Koji Koizumi • *Scientific Imaging Section, Okinawa Institute of Science and Technology Graduate University (OIST), Okinawa, Japan*

Shohei Kojima • *Genome Immunobiology RIKEN Hakubi Research Team, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan*

Kaoru R. Komatsu • *xFOREST Therapeutics Co., Ltd., Kyoto, Japan*

Reina Komiya • *Science and Technology Group, OIST, Okinawa, Japan; PRESTO, Japan Science and Technology Agency (JST), Saitama, Japan*
Tomoya Kujirai • *Laboratory of Chromatin Structure and Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan*
Hitoshi Kurumizaka • *Laboratory of Chromatin Structure and Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan*
Nelson C. Lau • *Boston University School of Medicine, Department of Biochemistry, Boston University Bioinformatics Program, Boston, MA, USA*
Ming Li • *Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, La Jolla, CA, USA*
Conor J. McMeniman • *W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA; The Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
Pascal Miesen • *Department of Medical Microbiology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands*
Emi Miyashita • *Department of Life Science Frontiers, Center for iPS Cell Research and Application, Kyoto University, Kyoto, Japan; xFOREST Therapeutics Co., Ltd., Kyoto, Japan*
Atsumi Miyata • *Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto, Japan*
Kazufumi Mochizuki • *Institute of Human Genetics (IGH), CNRS and University of Montpellier, Montpellier, France*
Valeriya Morgunova • *Institute of Molecular Genetics of National Research Centre, "Kurchatov Institute", Moscow, Russia*
Zissimos Mourelatos • *Division of Neuropathology, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA*
Kensaku Murano • *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan*
Masahiro Naganuma • *Laboratory of RNA Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan; RIKEN Center for Biosystems Dynamics Research, Yokohama, Japan*
Shinichi Nakagawa • *RNA Biology Laboratory, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan; Global Station for Biosurfaces and Drug Discovery, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, Sapporo, Japan*
Satoshi H. Namekawa • *Department of Microbiology and Molecular Genetics, University of California, Davis, Davis, CA, USA*
Hideyuki Okano • *Department of Physiology, Keio University School of Medicine, Tokyo, Japan*
Naoki Osato • *Faculty of Science and Engineering, Waseda University, Tokyo, Japan*
Umberto Palatini • *Department of Biology and Biotechnology, University of Pavia, Pavia, Italy*

Elisa Pischedda • *Department of Biology and Biotechnology, University of Pavia, Pavia, Italy*
Elke F. Roovers • *Hubrecht Institute for Developmental Biology and Stem Cell Research, Royal Netherlands Academy of Arts and Sciences, Utrecht, The Netherlands*
Hirohide Saito • *Department of Life Science Frontiers, Center for iPS Cell Research and Application, Kyoto University, Kyoto, Japan*
Kuniaki Saito • *Invertebrate Genetics Laboratory, Department of Chromosome Science, National Institute of Genetics, Research Organization of Information and Systems (ROIS), Mishima, Shizuoka, Japan; Division of Invertebrate Genetics, Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Mishima, Shizuoka, Japan*
Akihiko Sakashita • *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan*
Hiroyuki Sasaki • *Division of Epigenomics and Development, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan*
Kotaro Sekine • *Faculty of Science and Engineering, Waseda University, Tokyo, Japan*
Abd-El Monsif Shawky • *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA*
Salman Shehzada • *Institute of Human Genetics (IGH), CNRS and University of Montpellier, Montpellier, France*
Shuting Shen • *School of Life Science and Technology & Gene Editing Center, ShanghaiTech University, Shanghai, China*
Norito Shibata • *Department of Integrated Science and Technology, National Institute of Technology, Tsuyama College, Tsuyama-City, Okayama, Japan*
Maria M. Sukhova • *Institute of Molecular Genetics of National Research Centre, "Kurchatov Institute", Moscow, Russia; Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia*
Ruichen Sun • *Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, La Jolla, CA, USA*
Hisashi Tadakuma • *School of Life Science and Technology & Gene Editing Center, ShanghaiTech University, Shanghai, China; Laboratory of RNA Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan*
Atsushi Takeda • *Faculty of Science and Engineering, Waseda University, Tokyo, Japan*
Chikara Takeuchi • *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan*
Minjie Tan • *Shenzhen Bay Laboratory, Shenzhen, China*
Yukihide Tomari • *Laboratory of RNA Function, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan; Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan*
Helena T. A. van Tol • *Farm Animal Health, Population Health Sciences, Faculty of Veterinary Medicine, Utrecht, The Netherlands*
Anastasios Vourekas • *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA*
Tetsuya Yamamoto • *Institute for Chemical Reaction Design and Discovery, Hokkaido University, Sapporo, Japan; PRESTO, Japan Science and Technology Agency (JST), Saitama, Japan*

Tomohiro Yamazaki • *Graduate School of Frontier Biosciences, Osaka University, Suita, Japan*
Saori Yokoi • *RNA Biology Laboratory, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*
Chao Zeng • *Faculty of Science and Engineering, Waseda University, Tokyo, Japan; AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), Tokyo, Japan*

Part I

Methods to Study the Nature and Function of piRNAs in Non-classical Organisms

Chapter 1

Functional Analysis of Individual piRNAs in *Aedes aegypti* Cells and Embryos Using Antisense Oligonucleotides

Rebecca Halbach and Pascal Miesen

Abstract

In insects, PIWI-interacting (pi)RNAs fulfill versatile regulatory functions inside and outside the germline, including posttranscriptional repression of transposable elements and regulation of gene expression. Canonically, piRNAs act—and have been studied—as a conglomerate of several thousand sequences that cooperatively silence target RNAs. Interestingly, however, an increasing number of studies have demonstrated that individual piRNAs can have profound biological activity as a *unique* piRNA sequence. Prime examples are the tapiR1 and 2 piRNAs, which mediate target RNA degradation in the developing embryo of *Aedes* mosquitoes. To study such outstanding individual piRNA species, we describe here a method to interfere with RNA target silencing using antisense oligonucleotides in cell culture as well as in mosquito pre-blastoderm embryos. Although the method has been established for *Aedes* mosquitoes, it can likely be adapted for use in other invertebrate species as well.

Key words Antisense oligonucleotides, piRNA, Target site, Gene silencing, Mosquito cell culture, Embryo injection, Luciferase reporter assay, RT-qPCR

1 Introduction

piRNAs are a class of small RNAs in the size range of 25–30 nt that associate with PIWI-type Argonaute proteins. They are extensively studied in the fly germline where they are key to repressing the detrimental activity of transposable elements [1]. However, *Aedes (Ae.) aegypti* mosquitoes, as well as many other arthropod species, express piRNAs also in somatic tissues [2, 3], and mosquitoes encode an expanded repertoire of PIWI proteins, suggesting a broader functionality [4, 5]. Interestingly, the majority of piRNAs in *Ae. aegypti* are not processed from transposon sequences, but instead, exert additional functions beyond transposon control, like antiviral defense [6, 7], or posttranscriptional gene silencing during development [8, 9].

Many studies use knockdown or knockout of PIWI proteins to decrease global piRNA levels and deduce piRNA functions from

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_1,

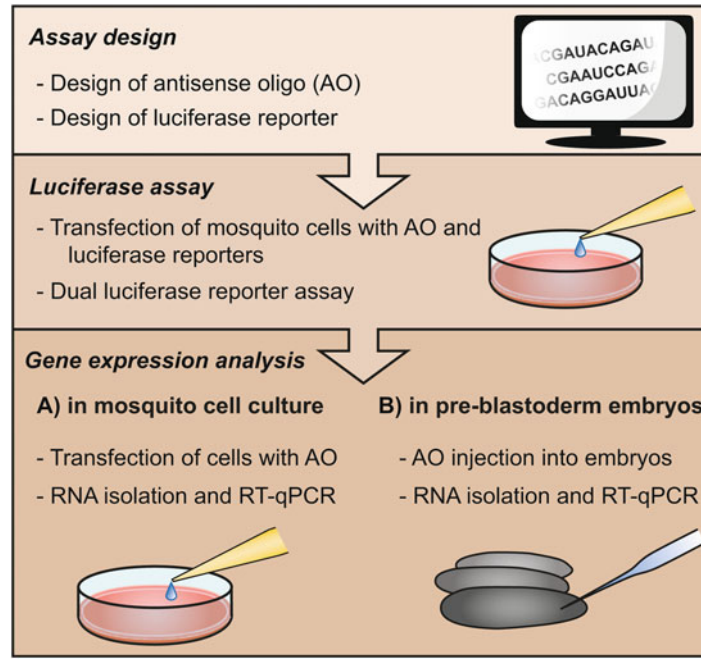


Fig. 1 Graphic overview of the experimental procedure

the resulting phenotypes (e.g., [10, 11]). This approach is suitable to dissect the regulatory potential of all piRNAs that associate with the PIWI protein that was silenced. Yet, increasing evidence indicates that individual and highly abundant piRNA species act in regulatory networks independently of the remaining piRNA population in the cell. For instance, unique piRNA sequences are central to sex determination in the silkworm *Bombyx mori* and nematodes [12, 13] and to the regulation of gene expression in the developing embryo in *Ae. aegypti* mosquitoes [8, 9]. Dissecting such regulatory networks requires specific manipulation of the function of individual piRNAs in contrast to the depletion of entire piRNA populations.

This chapter describes a detailed protocol to inhibit single piRNA sequences, making use of antisense oligonucleotides (AOs) (Fig. 1). This method is a powerful tool, in particular when combined with a targeted or global gene expression analysis for instance by RT-qPCR or deep sequencing, respectively. In such assays, genes that are robustly upregulated upon AO treatment are putative target genes of the piRNA of interest. The presence of a predicted target site in the upregulated transcripts further indicates that these genes are silenced due to *direct* piRNA–target site interactions. This can be experimentally validated by placing putative target sites from their endogenous sequence context into a luciferase reporter, making it possible to study the silencing potential of an isolated piRNA target site.

We describe here a straightforward RT-qPCR approach that can be used when a piRNA target gene is known or at least suspected a priori, and is a valuable tool for optimization of treatment conditions (e.g., AO concentrations, time points) prior to a large-scale RNA-sequencing experiment. Yet, since possible target genes of the piRNA of interest are often unknown at the start of a project, we also describe an assay based on a luciferase reporter harboring an artificial, perfectly complementary piRNA target site that can indicate whether the AO efficiently blocks piRNA function in general. As illustration of these methods, we discuss AO-mediated inhibition of tapiR1, a piRNA that is abundantly expressed in *Aedes* mosquitoes and critical for embryonic development. We describe the use of AOs in mosquito cell culture, and we will outline the application of AOs to manipulate individual piRNAs in vivo, using mosquito pre-blastoderm embryos as a model.

2 Materials

2.1 Cell Culture

1. *Aedes aegypti* Aag2 cells (*see* **Note 1**).
2. Aag2 cell culture medium, fully supplemented: 500 ml Leibovitz's L-15 medium without phenol red (Gibco) with 50 ml heat-inactivated Fetal calf serum (FCS), 5 ml penicillin/streptomycin, 5 ml non-essential amino acids (100× stock), and 10 ml tryptose phosphate broth solution (stock concentration: 29.5 g/l). Sterile-filter all medium supplements and store the prepared medium at 4 °C for up to 2 months. Pre-warm to 28 °C before use.
3. 50-ml syringe.
4. 0.2-μm syringe filter.
5. 25 cm^2 or 75 cm^2 cell culture flasks (*see* **Note 2**).
6. 24-well and 96-well cell culture plates.

2.2 Transfection of AOs in Aag2 Cells

1. 10 μM antisense oligos: Dilute lyophilized RNA oligonucleotides in nuclease-free water and store in smaller aliquots at −70 °C to avoid frequent freeze–thaw cycles.
2. Aag2 culture medium, fully supplemented (see Subheading 2.1).
3. Leibovitz's L-15 medium, unsupplemented. Store at 4 °C for several months.
4. X-tremeGENE HP DNA transfection reagent (Roche) (*see* **Note 3**).

2.3 Luciferase Reporter Assay

1. pMT-firefly luciferase (FLuc)-based piRNA target reporter and corresponding control (for design considerations, *see* Subheading 3.2).
2. pMT-*Renilla* luciferase (RLuc) plasmid for internal transfection control (*see* **Note 4**).
3. 96-well flat bottom cell culture plate.
4. 50 mM $CuSO_4$: Dissolve 250 mg $CuSO_4 \cdot 5H_2O$ (Sigma Aldrich) in 10 ml water. Mix well, sterilize by filtration, aliquot, and store at −20 °C for up to several years.
5. Dual-luciferase assay kit (Promega).
   (a) Resuspend lyophilized Luciferase Assay Substrate in Luciferase Assay Buffer II (LARII). Aliquot and store at −70 °C for several months.
   (b) *Just before use*, dilute the 5× Passive Lysis Buffer to 1× in water.
   (c) *Just before use*, add 50× Stop&Glo Substrate to a final concentration of 1× in Stop&Glo Buffer (*see* **Note 5**).
6. Rocking platform or orbital shaker.
7. Modulus Single-Tube Luminometer (Turner BioSystems) or equivalent luminescence reader (*see* **Note 6**).
8. Round-bottom screw tubes (required tubes depend on luminescence reader used).

2.4 RNA Isolation

1. RNA-Solv Reagent (Omega Bio-Tek), or similar means of RNA isolation (*see* **Note 7**).
2. Chloroform.
3. Isopropanol.
4. 80% ethanol: mix 20 ml nuclease-free water with 80 ml absolute ethanol.
5. Nuclease-free water.

2.5 Reverse Transcription and Quantitative PCR

1. DNaseI with 10× DNaseI buffer (Ambion).
2. 25 mM EDTA: Dilute 50 µl 0.5 M nuclease-free EDTA pH 8.0 (Invitrogen) in 950 µl nuclease-free water. Store at room temperature for several months.
3. TaqMan Reverse transcription kit (Applied Biosystems). The kit includes reverse transcriptase buffer, $MgCl_2$, dNTPs, reverse transcriptase, oligo dT primers, random hexamers, and RNase inhibitors.
4. GoTaq qPCR master mix (Promega).
5. Gene-specific qPCR primers for a piRNA target gene, and one or several reference genes. The primers can be designed with a qPCR design tool, for example, primer3 https://primer3.ut.ee/. As

reference genes, we routinely use lysosomal aspartic protease (LAP) and RpL5 for Aag2 cells [8], and RpL8, Actin and eukaryotic translation elongation factor 1A (eEF1A) for early embryos [14]. The primers for the piRNA target gene should not span the piRNA target site itself. Prepare a 10 μM dilution of the primer stock.

6. LightCycler 96-well multi-well plate white and sealing foil (Roche).
7. LightCycler 480 instrument (Roche) or equivalent qPCR machine.

2.6 Mosquito Husbandry

1. *Aedes aegypti* mosquitoes, Liverpool strain (BEI resources, MRA-735) (*see* **Note 8**).
2. Vacuum machine.
3. 2–3 l trays to rear mosquito larvae. We use containers with the following proportions (L × W × H): 25 × 16 × 8 cm.
4. Tetramin Baby Bio Active Fish food (Tetra).
5. 250–500 ml plastic cup to collect mosquito pupae (e.g., autoclavable Nalgene containers).
6. BugDorm-1 insect rearing cage (30 × 30 × 30 cm) (BugDorm).
7. Human or animal blood (for example from rabbit or chicken). We routinely use human whole blood in heparin tubes provided by the Sanquin blood bank, Netherlands.
8. Membrane feeding system (Hemotek Ltd.).
9. 10% sucrose solution: Dissolve 50 g of sucrose or household sugar in 500 ml drinking-quality tap water. For storing sugar water, use a clean bottle that is not used for preparation of other chemicals or buffers. Store in the fridge for up to a week.
10. 50-ml Erlenmeyer flasks.
11. Parotisroll dental cotton sticks (Coltene).

2.7 Injection of Embryos

1. 50 μM antisense oligos: dilute lyophilized RNA oligonucleotides in nuclease-free water. Centrifuge for 20 min at maximum speed and 4 °C. Store in smaller aliquots at −70 °C to avoid frequent freeze–thaw cycles (*see* **Note 9**).
2. P-2000 needle puller (Sutter Instruments).
3. Quartz needles with filament (10 cm, O.D 1.0 mm, ID 0.7 mm) (Sutter Instruments, QF100-70-10). Pull the needles on a P-2000 needle puller with Heat: 750; Filament: 4; Velocity: 40; Delay: 150; Pull: 150 (*see* **Notes 10** and **11**).
4. FemtoJet (Eppendorf), Pneumatic PicoPump (World Precision Instruments), or uPUMP (World Precision Instruments) microinjector.

5. Inverted microscope with 4× and 10× objectives (Leica), with micromanipulator attached.
6. Stereomicroscope.
7. *Drosophila* vials 25 × 95 mm (Dutscher).
8. Cotton balls.
9. Round Whatman paper circles, 2.5 cm diameter, grade 1 (*see* **Note 12**).
10. Halocarbon oil 700 (Sigma Aldrich).
11. Two fine paint brushes (keep one of the brushes for use with halocarbon oil exclusively).
12. Fine tweezers.
13. Square cover slips 22 × 22 mm.
14. Transparent, double-sided sticky tape (3 M double-coated tape 415, 1/2 in. wide).
15. "Embryo coverslips": Stick a small patch of double-sided sticky tape on a square coverslip. Embryos will eventually be immobilized on the sticky tape during injection, without obscuring the light (*see* **Note 13**).
16. Elevated microscope slide podium: Using double-sided sticky tape, glue two stacks of four cover slips each on a microscope slide. The two stacks should have a distance of ~1 cm. Place a final layer of double-sided adhesive tape on each stack to eventually attach the coverslip with embryos on top. The podium facilitates immobilization and handling of the embryo coverslip during and after injection.
17. Whatman folded filters, 110 mm (GE Healthcare Life Sciences).
18. 100–200 ml plastic beaker.
19. Petri dish.
20. 1 mm Zirconia beads (Biospec) and bead-beating homogenizer, or plastic pestle for manual homogenization of embryos.

3 Methods

3.1 Design of Antisense Oligonucleotides (AOs)

In theory, every piRNA of interest can be investigated. However, the silencing effect of lowly abundant, individual piRNAs is likely hard to measure. Therefore, we recommend to design AOs only for piRNAs that are abundant enough to robustly silence a luciferase piRNA reporter (*see* Subheading 3.2), which immediately provides a tool to optimize and validate the experimental setup of the AO treatment.

AOs should be complementary to the full-length sequence of the piRNA species in question. Lyophilized RNA oligonucleotides can be purchased from standard suppliers and should be of high purity (e.g., HPLC-purified). We have good experience with using RNA oligonucleotides that are fully 2′*O*-methylated which increases the stability of the AOs and, additionally, prevents slicing by PIWI protein. We noticed that the use of 5′-Cy5-labeled AOs resulted in ~10-fold stronger effects in cell culture experiments, probably by increasing transfection efficiency, allowing for the use of lower AO concentrations. The Cy5-label does not interfere with luminescence assays. A control AO should be designed similarly but must not have extensive complementarity to endogenous small RNAs or transcripts. You can check this in advance by aligning the sequence of the control AO to the transcriptome of the species of interest using for instance the BLAST algorithm.

3.2 Design of Luciferase Reporter

A luciferase-based reporter assay can serve two important purposes. It can be used to optimize the AO treatment with a reporter containing an artificial, fully complementary piRNA target site, or to experimentally validate the targeting potential of an actual predicted piRNA target site taken from a putative target gene and placed out of its endogenous sequence context. The design considerations of the assay as well as the experimental procedures are the same for these two objectives.

To set up a luciferase reporter assay (Fig. 2a), clone a target site for the piRNA of interest in the 3′ untranslated region (UTR) of the firefly luciferase (FLuc) gene. The target site must be in antisense orientation to the full-length sequence of the piRNA. Alternatively, an endogenous piRNA target site from a validated or predicted target gene can be introduced in the reporter (*see* **Note 14**). Such target sites can be cloned by annealing and phosphorylation of short DNA oligonucleotides that are ligated into the digested pMT-FLuc vector. To express the reporter in Aag2 cells, we use the pMT vector in which luciferase expression is under the control of the inducible metallothionein promoter (*see* **Note 15**); however, any other promoter that drives sufficient expression of the luciferase gene (e.g., *Polyubiquitin* or *Actin* promoter) will work as well. A firefly reporter with a nonfunctional target site (e.g., a scrambled sequence or seed mutant) is used as negative control (*see* **Note 16**).

In addition to the piRNA reporters, a construct with *Renilla* luciferase expressed from the same promoter is needed as internal control. This reporter must not be targeted by the piRNA of interest (*see* **Note 4**) and serves as internal control to account for well-to-well and condition-to-condition differences in transfection efficiencies and can be used to normalize the data.

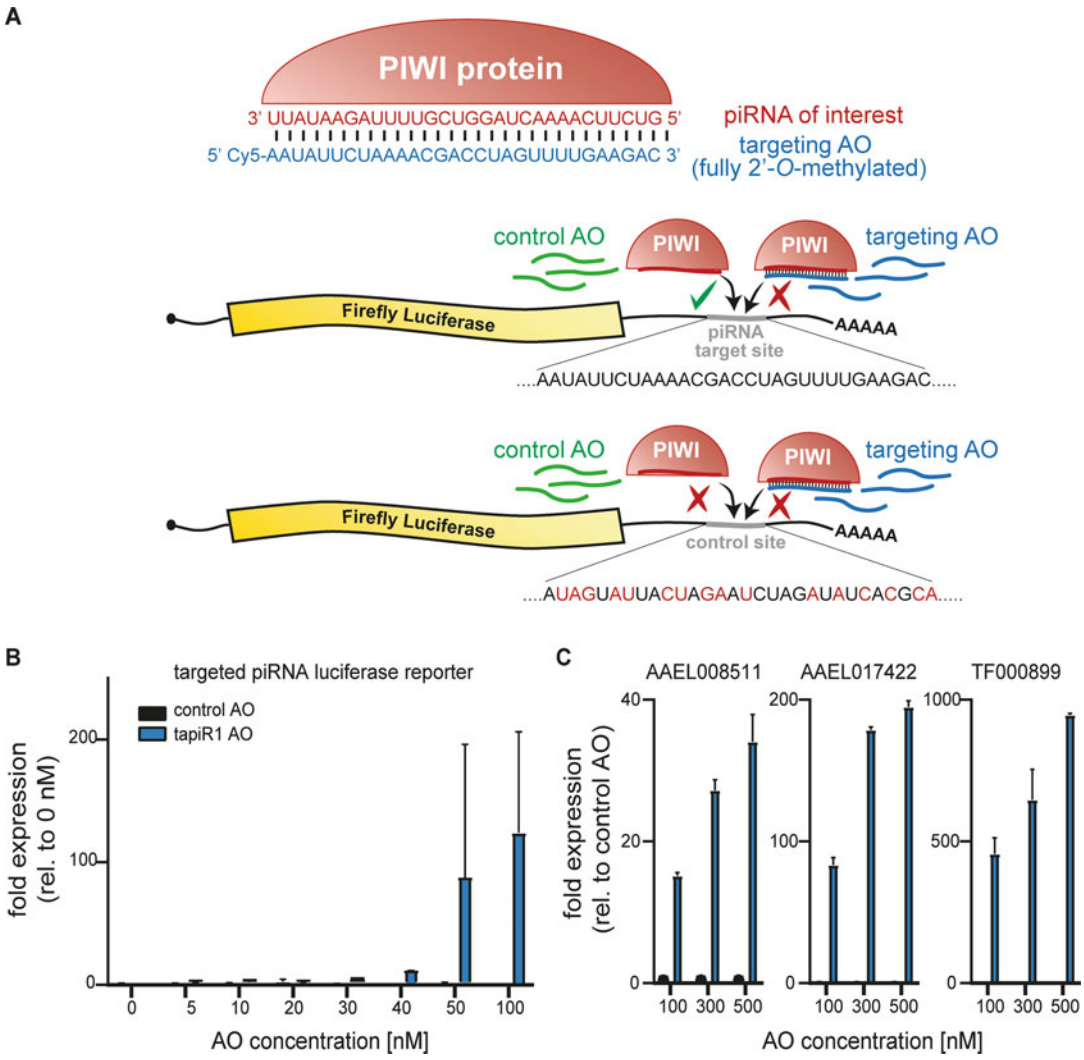


Fig. 2 Gene expression analyses after AO treatment. (**a**) Design of a luciferase reporter assay to measure posttranscriptional gene silencing mediated by a single piRNA. The piRNA of interest (indicated in red) is loaded in a PIWI protein and is physically blocked from targeting a target-site containing luciferase gene by a fully methylated RNA AO (blue). Co-transfection of the AO, but not a control AO (green) together with the piRNA reporter, relieves reporter silencing. A non-targeted luciferase reporter with a scrambled target site is not silenced by the piRNA of interest and serves as additional targeting control. (**b**) Dual-luciferase assay to optimize AO concentrations to achieve de-silencing of a piRNA-targeted reporter construct in Aag2 cells. Only the fully complementary AO but not a control sequence relieves silencing. (**c**) RT-qPCR assay to optimize AO concentration to achieve robust de-silencing of three endogenous tapiR1 target genes in Aag2 cells. Only a fully complementary AO, but not a control AO relieves the silencing. Bars and error bars indicate mean and standard deviation of a biological triplicate (**b**) or technical duplicate (**c**)

3.3 Aag2 Cell Culture

1. Culture Aag2 cells in 25-cm^2 or 75-cm^2 flasks at 28 °C without CO_2, and passage when approaching confluency.
2. To split cells, gently dissociate by flushing the cells in their cell culture medium using a serological pipet until they detach from the cell culture flask.
3. Transfer the desired amount of cell suspension to a new clean cell culture flask with fresh medium. We generally split cells in a 1:3 to 1:5 dilution twice per week.

3.4 Luciferase Reporter Assay

3.4.1 Transfection of Aag2 Cells

1. One day before the transfection, resuspend cells by gently flushing the bottom of the flask with the cell culture medium several times or by using a cell scraper. Dilute the cells 1:3 to 1:5 in cell culture medium, and transfer 100 µl of cell suspension to a well in a 96-well cell culture plate. The cells should be 60–80% confluent at the day of transfection. Seed three wells per condition.
2. Prepare a master mix with the following components (per condition):
   (a) 100 µl unsupplemented Leibovitz's L-15 medium.
   (b) 300 ng pMT-RLuc plasmid.
   (c) 300 ng pMT-FLuc plasmid (with target site or control).
3. Mix well by vortexing and distribute into one microcentrifugation tube per condition.
4. For each condition, add increasing amounts of AO (piRNA-specific or control) (*see* **Note 17**).
5. Vortex gently, and briefly spin down.
6. Add the required amount of X-tremeGENE HP DNA transfection reagent (*see* **Notes 18** and **19**), depending on the AO concentration: For transfection of plasmid DNA only, use 1.2 µl of X-tremeGENE HP DNA transfection reagent (2 µl reagent per µg of plasmid), and additional 2 µl of X-tremeGENE HP DNA transfection reagent per µg of AO (*see* **Note 20**).
7. Incubate for 15–20 min at room temperature.
8. Mix carefully by slowly pipetting up and down twice, and distribute the reaction mix among three wells in a dropwise manner.
9. Incubate the cells at 28 °C for 3–4 h.
10. In the meantime, dilute the $CuSO_4$ stock to 0.5 mM in fully supplemented cell culture medium.

11. To induce firefly and *Renilla* luciferase expression, replace the transfection medium with 100 µl of fresh $CuSO_4$-supplemented cell culture medium.
12. Incubate the cells for 24 h at 28 °C.

3.4.2 Luciferase Assay

This protocol follows the general instructions for the Promega Dual-Luciferase Assay System provided by the manufacturer. We have, however, adapted the volumes of the required reagents to decrease material costs.

1. Let all reagents equilibrate to room temperature.
2. Take transfected cells from the incubator and remove the cell culture medium.
3. Add 30 µl of freshly prepared 1× Passive Lysis buffer to each well.
4. Incubate the plate for at least 15 min with gentle shaking on a rocking platform or orbital shaker (*see* **Note 21**).
5. For each well, pipet 25 µl LARII into a round-bottom screw tube.
6. Add 10 µl lysate to a luminometer tube and mix with the LARII Reagent by pipetting up and down.
7. Measure firefly luciferase activity in the luminometer and measure with the following settings: 10 measurements, measurement frequency: 1 s, integration time: 1 s.
8. Remove tube from luminometer, add 25 µl freshly prepared Stop&Glo Reagent and mix by brief vortexing (*see* **Note 22**).
9. Place the tube into the luminometer and measure *Renilla* luciferase activity with the settings described in **step 7**.
10. Discard the tube and proceed with the next measurement.
11. Analyze the data: For each individual lysate, divide the firefly luciferase counts by the corresponding *Renilla* luciferase counts. Calculate the mean and standard deviation of the three FLuc/RLuc ratios determined for each condition. The following results are expected when validating AOs: (1) Without AO treatment, the piRNA-targeted firefly luciferase reporter is efficiently silenced compared to a control reporter. (2) Firefly luciferase activity of the targeted reporter increases in an AO concentration-dependent manner while *Renilla* luciferase activity is unaffected, and (3) the control AO does not change firefly luciferase activity of the reporter (Fig. 2b). If these expectations are not met, further optimization may be required.

3.5 Gene Expression Analysis in Aag2 Cells by RT-qPCR

3.5.1 Transfection of Aag2 Cells

1. One day before the transfection, resuspend cells by gently flushing the bottom of the flask with the cell culture medium several times, or use a cell scraper. Dilute the cells 1:3 to 1:5 with fully supplemented cell culture medium and transfer the cells to a well in a 24-well cell culture plate. The cells should be 60–80% confluent at the day of transfection. Seed three wells per conditions.
2. In a microcentrifugation tube, mix 300 μl of transfection medium with the desired amount of AO required for the transfection of three wells (*see* **Note 23**). Vortex gently, and briefly spin down.
3. Add 2 μl of X-tremeGENE HP DNA transfection reagent per μg of AO (*see* **Notes 19** and **23**).
4. Incubate for 15–20 min at room temperature.
5. Carefully mix by pipetting up and down twice and add 100 μl of reaction mix to each well of the triplicates in a dropwise manner.
6. Incubate the cells at 28 °C for 3–4 h.
7. Replace the transfection medium with 500 μl of fresh, fully supplemented cell culture medium.
8. Incubate the cells for 24 h at 28 °C (*see* **Note 24**).

3.5.2 RNA Isolation

1. Remove the culture medium from the cells.
2. Add 1 ml of RNA-Solv Reagent and incubate for several minutes (*see* **Note 25**).
3. Transfer the reagent to a 1.5-ml microcentrifugation tube.
4. Add 200 μl of chloroform. Vortex for 20 s and leave at room temperature for a couple of minutes until phases start to separate (*see* **Note 26**).
5. Centrifuge for 20 min at 16,000 × *g* and 4 °C. Transfer 80% of the upper, aqueous phase to a new 1.5-ml microcentrifugation tube. Make sure to not disturb the intermediate or organic phase.
6. Add 1 volume of ice-cold isopropanol and vortex for 20 s (*see* **Note 27**).
7. Incubate for 1 h on ice or −20 °C.
8. Pellet the precipitated RNA by centrifugation for 20 min at 16,000 × *g* and 4 °C. Discard all isopropanol without disturbing the pellet.
9. Add 1 ml of 80% ethanol. Centrifuge for 5 min at 8500 × *g* and 4 °C and discard ethanol without disturbing the pellet.
10. Repeat the ethanol wash twice.

11. Carefully remove all residual ethanol and air-dry the pellet until all ethanol is evaporated (do not dry for >10 min).
12. Resuspend the pellet in 20 µl of nuclease-free water. Store at −70 °C for several months.

The RNA can be used for downstream analysis like RT-qPCR for the analysis of individual target RNAs (Subheading 3.5.3) or used to prepare deep sequencing libraries for global transcriptome analysis. Additionally, stability of the piRNA of interest upon AO treatment can be assessed by small RNA northern blotting. In our laboratory, we have good experience with the northern blot protocol described in Ref. 15.

3.5.3 cDNA Synthesis and Quantitative PCR

We describe here an RT-qPCR protocol making use of the Taqman reverse transcription kit and the GoTaq qPCR mastermix. Different reagents for cDNA synthesis and quantitative PCR should be assembled according to the manufacturer's instructions or based on empiric optimization in the laboratory.

1. Set up the DNaseI reaction (total volume 7 µl).
   (a) 0.2–1 µg of RNA purified in Subheading 3.5.2.
   (b) 0.7 µl 10× DNaseI buffer.
   (c) 0.7 µl DNaseI.
   (d) Add nuclease-free water to 7 µl.
2. Incubate for 45 min at 37 °C.
3. Add 0.7 µl 25 mM EDTA to prevent chemical scission of the RNA.
4. Heat-inactivate the DNaseI for 10 min at 70 °C.
5. Set up the reverse transcription reaction (total volume 20 µl):
   (a) 7.7 µl DNaseI-treated RNA.
   (b) 2 µl 10× Taqman RT buffer.
   (c) 4.4 µl 25 mM $MgCl_2$.
   (d) 4 µl 2.5 mM dNTPs each.
   (e) 1 µl Random hexamers/oligo dT primers.
   (f) 0.4 µl RNase inhibitors.
   (g) 0.5 µl Reverse transcriptase.
6. Incubate the reaction for 10 min at 25 °C, 60 min at 48 °C, 5 min at 95 °C, hold at 12 °C.
7. Dilute cDNA 1:5 with nuclease-free water.
8. Prepare the qPCR mix on ice:
   (a) 5 µl diluted cDNA.
   (b) 10 µl 2× SYBRGreen Master Mix.

(c) 0.6 µl 10 µM forward primer.

(d) 0.6 µl 10 µM reverse primer.

(e) 3.8 µl nuclease-free water.

Seal the plate and briefly centrifuge.

9. Run the qPCR with the following settings: 5 min initial denaturation at 95 °C, 40 cycles of 5 s denaturation at 95 °C, 10 s annealing at 60 °C, 20 s extension at 72 °C. Melting curve: 5 s denaturation at 95 °C, cooling to 65 °C for 1 min, continuously heating to 95 °C.
10. Analyze the data: Determine the relative expression of the piRNA target gene in AO-treated cells compared to control AO treatment with the $2^{-\Delta\Delta Ct}$ method [16], linear regression on the Log(fluorescence) per cycle number data method [17], or an equivalent method. The following results are expected: (1) expression of the piRNA target gene is unaffected by increasing amounts of transfected control AOs and comparable to untreated cells, (2) expression is higher in the targeting AO treatment compared to the control AO treatment, and (3) expression increases with increasing amounts of transfected AOs (Fig. 2c)

3.6 Gene Expression Analysis in Pre-Blastoderm Embryos

We have provided a brief protocol for mosquito husbandry, which can be used as guidelines, however, will not be sufficient to set up mosquito experiments from scratch. Training in an experienced laboratory is highly recommended.

3.6.1 Mosquito Husbandry

1. Hatch mosquito eggs, usually stored on filter paper or cotton pads, by submerging them in a small cup with tap water (drinking quality).
2. Place the cup with eggs in a vacuum machine and apply negative pressure for 30–60 min to stimulate hatching. Hatched larvae should now be visible in the container.
3. Transfer 200 hatched larvae to a larval tray filled with 1.5 l of drinking water.
4. Feed the larvae with fish food powder every other day until they start to pupate.
5. Transfer pupae into a small, water-filled pupal container and place it into a BugDorm Insect rearing cage for the mosquitoes to emerge.
6. Keep mosquitoes in a climate room with 28 °C and 80% humidity, with constant access to sugar water. The sugar water can be provided in a 50-ml Erlenmeyer flask with two dental cotton sticks placed into the flask making contact with the sugar solution and providing a landing platform for mosquitoes.

7. Blood-feed mosquitoes with a membrane feeding system. After 2 days, provide an egg laying cup with a moist Whatman paper to provide a surface for the females to lay their eggs. The surface should not dry for at least 2 days after egg laying. Hereafter, the eggs can be dried and stored for up to 3 months.

3.6.2 Injection into Pre-Blastoderm Embryos

A detailed protocol for injecting *Ae. aegypti* embryos can be found in Ref. 18. Embryo injection requires extensive training. It is advisable to learn the procedure from someone who is experienced in microinjections.

To reduce the time between the injection of AOs and control AOs, it is easiest to work together with two people, one already lining up the embryos while the other one performs the injections. Since there can be variation between different injections and experiments (e.g., sharpness of the needle, time between injections of the different AOs), it is absolutely necessary to include sufficient repetition not only of the experiment itself but also of the injections within one experiment. We inject the same AOs two to three times within one experiment (each round of injection using a new needle) and change the order of the injected conditions between experiments to obtain robust data.

1. Blood-feed female mosquitoes at least 72 h before the injection. After feeding, do **not** provide an egg-laying cup in the mosquito cage to avoid premature oviposition.
2. At the beginning of the experiment, prepare egg laying vials: Put a wet cotton ball at the bottom of a *Drosophila* vial and place a round Whatman paper circle on top, so that the surface is moist but not covered with water. Close the vial with a cotton ball or similar plug.
3. Place 10–12 gravid females into the egg-laying vial. Place the vial into the dark (e.g., under a cardboard box) for 30–60 min to enforce egg laying. The females will deposit the eggs onto the moist Whatman paper (*see* **Note 28**).
4. Meanwhile, prepare the embryo recovery cups: Place a damp Whatman filter to the wall of a small plastic beaker, and place a wet cotton ball at the bottom of the cup to prevent drying of the paper. Place the cup at 28 °C.
5. Back-fill the quartz needle with the AOs. Do this only shortly before injection and use gloves to avoid contamination with RNases. Once the needle is filled, place it in the pipet holder and attach the holder to the micromanipulator. Be careful not to break the needle.
6. Turn on the microinjector beforehand so that the pressure has time to build up.

7. When the females have laid eggs, remove them from the *Drosophila* vial (*see* **Note 29**), and place the Whatman paper in a petri dish. Keep the paper moist at all times.
8. Use a stereomicroscope to line up 50–80 embryos with comparably dark-gray color on a new moist Whatman paper circle using a fine brush. The embryos should be aligned with the anterior pole to one side. It is important that the paper stays moist all the time; otherwise, the embryos will start to desiccate.
9. Once all embryos are aligned, dry the paper by pressing it on a tissue. Transfer the embryos to a coverslip by gently pressing the adhesive tape onto the embryos.
10. Cover the embryos with halocarbon oil once the embryos have reached the desired level of desiccation (small dimples appear) (*see* **Note 30**).
11. Transfer the coverslip to the elevated microscope slide podium, and place under the inverted microscope with the micromanipulator and needle attached. Position the needle in front of and in the same focus plane as the first embryos in line by adjusting the position of the needle with the micromanipulator. It is easiest to bring the needle into position using only low (e.g., 4×) magnification and then to switch to higher magnification (e.g., 10×) for the injections.
12. Set the injection pressure of the microinjector to 30 psi, and the compensation pressure to 10 psi.
13. Break the tip of the needle by gently touching the first embryo. The opening should be wide enough to inject a sufficient amount of liquid with one injection (we inject approximately 5% of the embryo volume). The injected volume can be checked by injecting into the halocarbon oil and can be adjusted by increasing or decreasing the injection pressure and time of injection (*see* **Note 31**).
14. Inject the line of embryos: move the embryos against the stationary needle by moving the microscope stage back and forth. The needle needs to penetrate both chorion as well as serosa before injecting AOs into the embryo.
15. After injecting the entire row of embryos, transfer the embryo coverslip to a stereomicroscope, and carefully remove the embryos from the adhesive tape with a fine paint brush without damaging them, and transfer to the moist Whatman paper in the recovery cup. Remove as much of the oil as possible by gently rolling the embryos over the paper using a brush.
16. Leave the embryos to develop for the desired time at 28 °C and 80% humidity.

17. For RNA extraction, transfer the embryos of one injected row of 50–80 embryos to an Eppendorf tube and crush in RNA-Solv Reagent with a pestle or by bead-beating, and isolate RNA as described in Subheading 3.5.2.
18. Analyze the expression of piRNA target genes by RT-qPCR as described in Subheading 3.5.3 or by deep sequencing. The analyzed target genes should be elevated upon AO injection compared to the control AO. We normally take along non-injected embryos as well to compare them to the control AO-injected embryos. Expression of the target gene should be similar in those two conditions. As an alternative to or in addition to analyzing gene expression, functional experiments to assess for instance developmental progression or embryo survival can be performed.

4 Notes

1. The Aag2 cell line is a highly heterogenous, non-clonal line that was originally isolated from embryonic tissues [19]; however, clonal sub-clones are also available, e.g., Aag2-AF5 [20], or the Aag2-C3 clone that is cleared from two persistently infecting insect-specific viruses [21]. Depending on the clone, culture conditions, or passage history, further optimization of the protocol might be required. The Aag2-AF5 clone can be purchased from the European Collection of Authenticated Cell Cultures.
2. In our experience, the choice of cell culture flask can widely influence Aag2 cell quality and survival. We have good experience with Corning TC-treated PE flasks (#430725U).
3. Other transfection reagents can be used to transfect Aag2 cells as well but require further optimization.
4. We noticed that *Renilla* luciferase harbors a potent target site for tapiR1, a piRNA that is highly expressed in Aag2 cells [8]. After mutating this target site, we observed an increase of *Renilla* luciferase expression by up to ten-fold.
5. Before performing the dual luciferase assay, let the temperature of all reagents and materials equilibrate to room temperature. Temperature largely influences luciferase activity.
6. It is possible to assay luciferase activity in a 96-well format with a luminescence plate reader, however, with reduced sensitivity.
7. RNA isolation with RNA-Solv Reagent is labor-intensive and involves handling of toxic chemicals. However, in contrast to commercial column-based purification methods, this method recovers both longer and small RNAs and gives higher yields and is therefore suitable to study both the piRNA and the target RNA(s) isolated from the same material.

8. Other *Ae. aegypti* strains can easily be injected as well, however, might require some optimization regarding mosquito husbandry and handling. We noticed, for example, that different strains might require longer or shorter times to lay eggs before injection experiments.
9. It is crucial to centrifuge the AOs before injections to remove all impurities from the solution, as these may otherwise clog the needle during injection. After centrifugation, transfer most of the AO solution to a new tube and discard the remaining few microliters that contains the impurities. For convenience, the AOs can be aliquoted in 1–2 μl, which is enough for one round of injection.
10. Settings might vary between instruments and need to be adjusted accordingly. Recommendations can be found in the Sutter instrument needle puller guide available at https://www.sutter.com/PDFs/cookbook.pdf.
11. At the beginning of an injection round, the tip of the needle will break when touching the first embryo, resulting in an open needle. However, if a beveller is available, it is best to bevel the needle tip to ~20–25° before the experiment to produce a sharp tip that can easily penetrate the chorion.
12. It is important to only use high-grade (grade 1) Whatman paper to ensure that as little fiber as possible is transferred to the tape. Fibers on the adhesive tape can easily damage the needle during injections.
13. Avoid touching the coverslip or the tape with bare fingers, since fingerprints will hamper injections. Prepare a couple of coverslips in advance.
14. A putative endogenous piRNA target site can be a sequence in a transcript that is either fully complementary or only has few mismatches to the piRNA, especially when outside of the seed. We predict putative target sites in transcripts with the microRNA: target prediction algorithm of RNAHybrid available online at https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid/ [22].
15. The metallothionine promoter is inducible by copper sulfate in *Drosophila* S2 cells; however, we noticed that transcription from this promoter is leaky in Aag2 cells; induction by copper sulfate increases expression by only two- to threefold. Therefore, the promoter needs to be considered constitutively active in Aag2 cells.
16. The control site should be as similar as possible to the piRNA target site with regard to length and GC content. We often use scrambled sequences, or seed mutants in which the bases opposite of nt 2–6 of the piRNA are mutated.

17. As a good starting point, we recommend using concentrations ranging from 25 to 300 nM.
18. Do not let undiluted X-tremeGENE HP come into contact with plastic surfaces and do not use siliconized tips when pipetting the transfection reagent. For additional recommendations concerning the handling of X-tremeGENE HP, consult the manufacturer's instructions.
19. High amounts of transfection reagent can be cytotoxic to the cells and induce considerable stress. If this is the case, it is possible to reduce the ratio to 1:1 without major reduction of transfection efficiencies.
20. Molar quantities of AOs can be easily converted to weight with online conversion calculators, for example nebiocalculator.neb.com.
21. Lysates can be stored at −20 °C for up to a month. When working with frozen lysates, make sure that the plate and lysates reach room temperature before use.
22. Properly mixing by vortexing is essential to ensure that all firefly luciferase signal is quenched by the Stop&Glo buffer. Otherwise, FLuc activity will bleed into the *Renilla* luciferase measurement.
23. The amount of AO needed for the experiment needs to be experimentally optimized. In our experience, 200–300 nM of AO final concentration is a good starting point.
24. The incubation time can be adjusted depending on the purpose of the experiment, the cellular processes studied, and the mechanism of target gene silencing. We found that 24 h post transfection is a good starting point for further optimization; however, we have observed strong effects for tapiR1 AO treatment as early as 4 h post transfection.
25. RNA-Solv Reagent contains phenol, which is highly hazardous, suspected to cause cancer, and is harmful to aquatic life. Always handle with care, and make sure that materials contaminated with RNA-Solv Reagent are disposed in the appropriate waste stream.
26. Chloroform is neurotoxic. Always work in a chemical flow cabinet and handle with care.
27. If very little RNA is expected (RNA isolation from ~50 embryos yields approximately 1 μg RNA), add 20 μg glycogen to the isopropanol before vortexing. Glycogen will increase RNA recovery without interfering with most downstream analyses.

28. The time frame should be long enough to allow for sufficient egg laying; however, not too long as the embryos will become too old for injection (eggs should not be older than 120 min or show a strong dark-gray or black-colored chorion).
29. Females can be re-used once or twice since not all females will have laid their eggs in the short period of time given.
30. The embryos should be covered by halocarbon oil for the shortest time possible, and not longer than 5–10 min. Halocarbon oil is toxic, and prolonged exposure can significantly impact survival. We normally do not inject more than 50–80 eggs at once to reduce the time needed for the injections.
31. Injection pressure depends greatly on the width of the tip. Adjust the injection pressure to yield an injected volume that is approximately 5% of the embryo volume, and the compensation pressure to a level that the injection mix is just leaking out of the needle ("bleeding") to prevent clogging.

Acknowledgments

We would like to thank Valerie Betting and Ronald van Rij for critical reading of the manuscript. This work was funded by a VENI project grant from the Dutch Research Council (NWO; grant number: VI.Veni.202.035) to P.M. and by an institutional fellowship of the Radboudumc Nijmegen.

References

1. Czech B, Hannon GJ (2016) One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. Trends Biochem Sci 41(4): 324–337
2. Akbari OS et al (2013) The developmental transcriptome of the mosquito Aedes aegypti, an invasive species and major arbovirus vector. G3 3(9):1493–1509
3. Lewis SH et al (2018) Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol 2(1):174–181
4. Campbell CL et al (2008) Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. BMC Genomics 9: 425
5. Lewis SH, Salmela H, Obbard DJ (2016) Duplication and diversification of dipteran Argonaute genes, and the evolutionary divergence of Piwi and Aubergine. Genome Biol Evol 8(3):507–518
6. Suzuki Y et al (2020) Non-retroviral endogenous viral element limits cognate virus replication in Aedes aegypti ovaries. Curr Biol 30(18): 3495–3506.e6
7. Schnettler E et al (2013) Knockdown of piRNA pathway proteins results in enhanced Semliki Forest virus production in mosquito cells. J Gen Virol 94(7):1680–1689
8. Halbach R et al (2020) A satellite repeat-derived piRNA controls embryonic development of Aedes. Nature 580(7802):274–277
9. Betting V et al (2021) A piRNA-lncRNA regulatory network initiates responder and trailer piRNA formation during mosquito embryonic development. RNA 27(10):1155–1172
10. Malone CD et al (2009) Specialized piRNA pathways act in germline and somatic tissues of the drosophila ovary. Cell 137(3):522–535
11. Li C et al (2009) Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell 137(3):509–521

12. Kiuchi T et al (2014) A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature 509(7502):633–636
13. Tang W et al (2018) A sex chromosome piRNA promotes robust dosage compensation and sex determination in C. elegans. Dev Cell 44(6): 762–770.e3
14. Dzaki N et al (2017) Evaluation of reference genes at different developmental stages for quantitative real-time PCR in Aedes aegypti. Sci Rep 7(1):43618
15. Pall GS, Hamilton AJ (2008) Improved northern blot method for enhanced detection of small RNA. Nat Protoc 3(6):1077–1084
16. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2-\Delta\Delta CT$ method. Methods 25(4):402–408
17. Ramakers C et al (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339(1): 62–66
18. Aryan A, Myles KM, Adelman ZN (2014) Targeted genome editing in Aedes aegypti using TALENs. Methods 69(1):38–45
19. Lan Q, Fallon AM (1990) Small heat shock proteins distinguish between two mosquito species and confirm identity of their cell lines. Am J Trop Med Hyg 43(6):669–676
20. Fredericks AC et al (2019) Aedes aegypti (Aag2)-derived clonal mosquito cell lines reveal the effects of pre-existing persistent infection with the insect-specific bunyavirus Phasi Charoen-like virus on arbovirus replication. PLoS Negl Trop Dis 13(11):e0007346
21. Göertz G et al (2019) Mosquito small RNA responses to West Nile and insect-specific virus infections in Aedes and Culex Mosquito cells. Viruses 11(3):271
22. Rehmsmeier M et al (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10(10):1507–1517

Chapter 2

CRISPR-Mediated Genome Engineering in *Aedes aegypti*

Ruichen Sun, Ming Li, Conor J. McMeniman, and Omar S. Akbari

Abstract

CRISPR-mediated genome engineering technologies have been adapted to a wide variety of organisms with high efficiency and specificity. The yellow fever mosquito, *Aedes aegypti*, is one such organism. It is also responsible for transmitting a wide variety of deadly viruses including Dengue, Zika, Yellow fever, and Chikungunya. The key to successful CRISPR-mediated gene editing applications is the delivery of both Cas9 ribonuclease and single-guide RNA (sgRNA) to the nucleus of desired cells. Various methods have been developed for supplying the Cas9 endonuclease, sgRNA, and donor DNA to *Ae. aegypti*. In this chapter, we focus on methods of direct embryo delivery of editing components, presenting detailed step-by-step CRISPR/Cas9-based genome-editing protocols for inducing desired heritable edits in mosquitoes as well as insights into successful application of these protocols. We also highlight potential opportunities for customizing these protocols to manipulate the mosquito genome for innovative in vivo gene function studies.

Key words *Aedes aegypti*, CRISPR/Cas9, sgRNA, Gene editing, HDR

1 Introduction

The discovery of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems has revolutionized the field of molecular biology, enabling transformative applications in both medicine and biotechnology [1]. In the past decade, the CRISPR/Cas gene editing technologies have been successfully applied in a wide variety of organisms with high efficiency and specificity. This has included several mosquito vectors [2–4], one of which is the yellow fever mosquito, *Aedes aegypti* [5, 6]. *Ae. aegypti* are notorious for their ability to transmit many blood-borne human pathogens, such as yellow fever virus, dengue virus, Zika virus, chikungunya virus, and others [7]. CRISPR/Cas technologies offer potential genome engineering solutions that may alleviate the significant public health burden of diseases spread by *Aedes aegypti* [8, 9].

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_2,

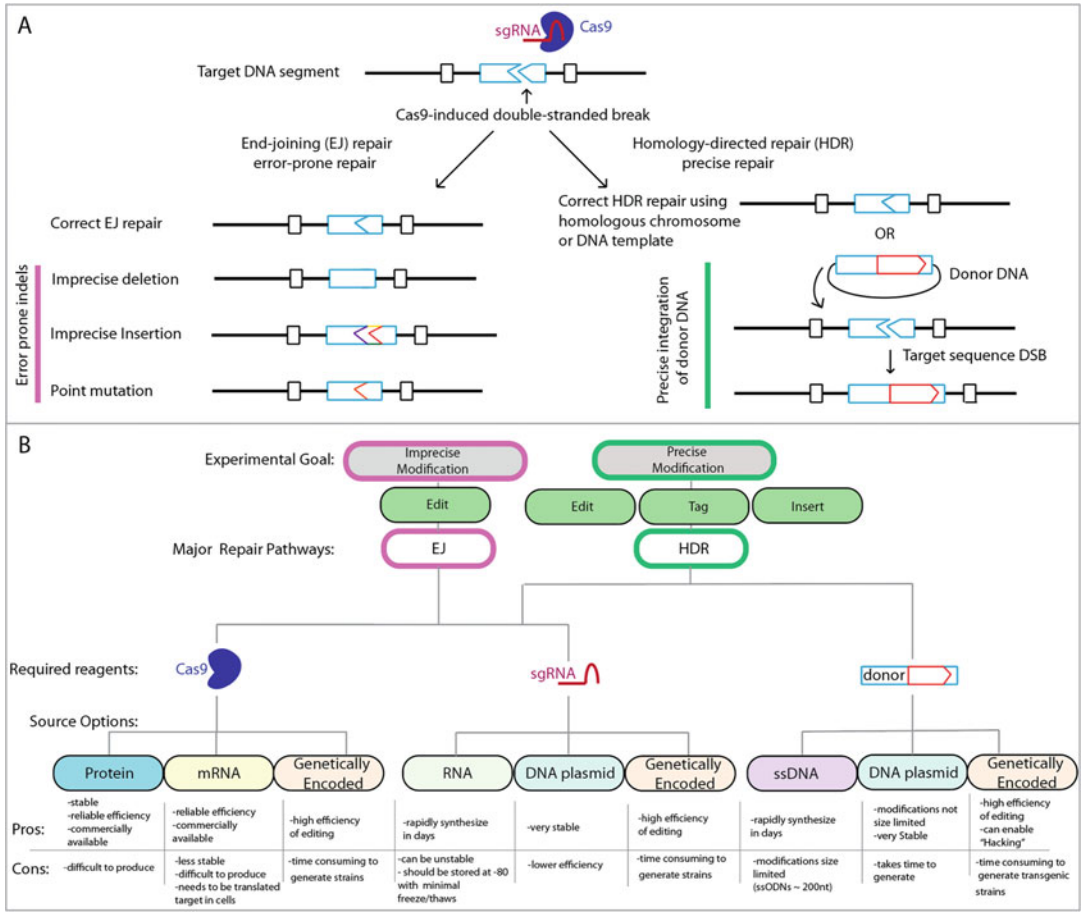


Fig. 1 CRISPR/Cas9-mediated gene editing mechanisms and delivery options. (**a**) A CRISPR/Cas9-mediated double-stranded DNA break can be repaired by either error-prone end joining (EJ) resulting in imprecise mutations in the target site or homology-directed repair (HDR) which can result in the precise insertion of desired stretches of DNA. (**b**) Multiple decisions and reagent delivery options are required for planning a CRISPR/Cas9 project with the goal of inducing an imprecise or precise modification

CRISPR/Cas9-based gene editing works as follows: Cas9 endonuclease (Cas9), originating from *Streptococcus pyogenes*, forms a complex with a synthetic single guide RNA (sgRNA)—an RNA segment engineered to encode complementary to a specific 20 bp-long genomic target DNA sequence [1]. This ribonucleoprotein complex induces a double-stranded DNA break at the genomic target locus complementary to the sgRNA sequence. The DNA break activates the endogenous cellular repair machinery which repairs the break via two predominant mechanisms: end-joining (EJ) or homology-directed repair (HDR) (Fig. 1). The EJ mechanism is error-prone and tends to incur small base-pair insertions or deletions (indels), while HDR uses templates for repair allowing insertion of desired exogenous DNA fragments into the genome.

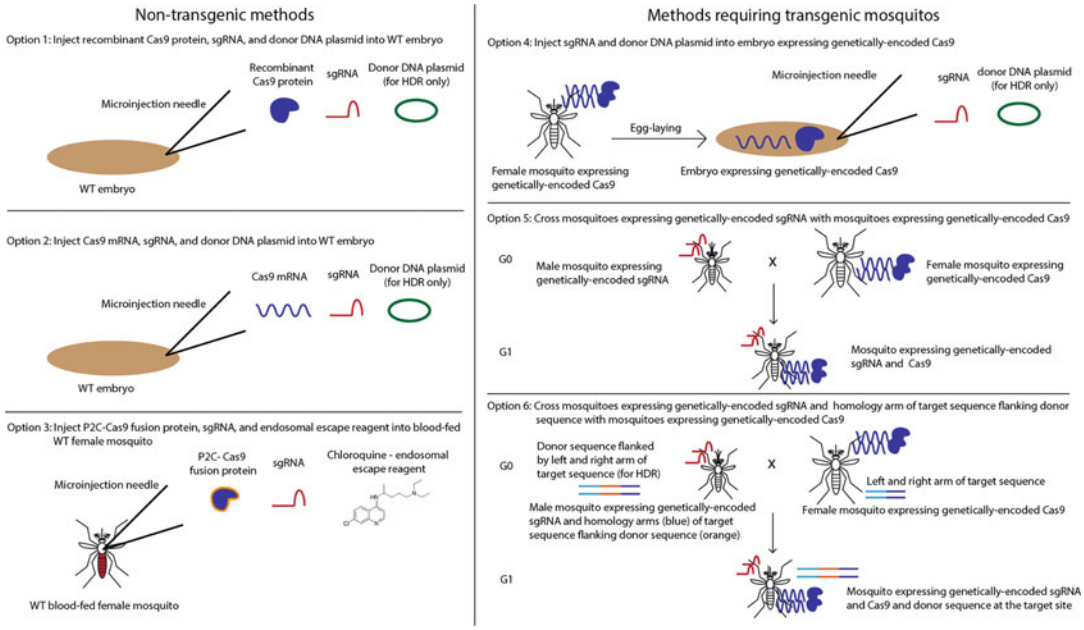


Fig. 2 Delivery methods for CRISPR DNA editing components. Option 1: Inject recombinant Cas9 protein, sgRNA, and donor DNA plasmid (for HDR only) into wild-type embryos. Option 2: Inject Cas9 mRNA, sgRNA, and donor DNA plasmid (for HDR only). Option 3: Inject P2C-Cas9 fusion protein, sgRNA, and endosomal escape reagent into the thorax of the blood-fed wild-type adult females. Option 4: Inject sgRNA and donor DNA plasmid into embryos expressing genetically encoded Cas9. Option 5: Cross mosquitoes expressing genetically encoded sgRNA with mosquitoes expressing genetically encoded Cas9. Option 6: Cross mosquitoes expressing genetically encoded sgRNA with homology arms of target sequence flanking donor sequence with mosquitoes expressing genetically encoded Cas9 to enable hacking

Spatially delivering both Cas9 endonuclease and sgRNAs to the correct tissue at the correct time is the key to the success for any CRISPR gene editing application. To this end, several methods have been experimented for supplying the Cas9 endonuclease and sgRNA to mosquitoes including: direct delivery of the sgRNA (as in vitro transcribed RNA or as a DNA plasmid) and Cas9 (as either mRNA or protein) by injection, as well as crossing genetically encoded sgRNA-expressing mosquitoes with Cas9-expressing mates, to provide offspring inheriting both components the capacity of self-editing (Fig. 2).

In the direct delivery method, Cas9/sgRNA components are injected into pre-syncytial blastoderm stage embryos, during which the embryo is one large multinucleated cell. Specifically, either recombinant Cas9 protein or purified Cas9 mRNA can be used for this method (Fig. 1b). If donor DNA (e.g., short single-stranded DNA oligonucleotides [ssODN] or plasmid DNA) containing homology to the sequences adjacent to the cut site is also supplied, insertion of the DNA donor into the cut site by HDR can also be achieved [5](Figs. 1b and 2, option 1–2). Simply injecting

the thorax of adult mothers with Cas9/sgRNA can also enable germline editing, termed Receptor-Mediated Ovary Transduction of Cargo (ReMOT Control) [10]. ReMOT exploits a mixture consisting of a endosomal escape reagent combined with a modified Cas9 protein with a fused peptide (P2C) that mediates uptake of the Cas9/gRNA complex from the female hemolymph (insect blood) to the developing oocytes (developing eggs). This results in heritable gene editing of the resulting offspring (Fig. 2, option 3). ReMOT has been used to generate heritable EJ mutations in several mosquito species, however has not yet been adapted for inserting DNA via HDR [10].

In the methods requiring transgenic mosquitoes, Cas9 and sgRNA are genetically encoded in the mosquito genome and expressed under the control of promoters which express in the appropriate target tissue such as the germline or soma. In these methods, editing can also occur if the mother maternally deposits Cas9 protein into embryos that are then injected with only sgRNAs, with the additional option of knock-in by HDR if donor DNA is supplied (Fig. 2, option 4) [6, 9]. Alternatively, when both the Cas9 and the sgRNA are genetically encoded, very high rates of editing can occur in the cells of target tissues when these strains are genetically crossed together (Figs. 1b and 2, option 5). In this case, a donor DNA cassette with homology arms can even be encoded elsewhere in the genome and can facilitate targeted knock-in by a HDR-like mechanism termed Homology-Assisted CRISPR Knock-in (HACK) [11, 12] (Fig. 2, option 6).

In this chapter, we focus on methods involving embryonic injections (Option 1, 2, and 4 in Fig. 2) and present detailed CRISPR/Cas9-based genome editing protocols for either inducing desired end-joining mutations (Subheading 3.1) or HDR insertions using plasmid DNA (Subheading 3.2). Overall, delivering Cas9 (as mRNA and as protein) into early embryos via injection results in lower and more stochastic editing than that of the genetically encoded expression methods. As a result, the editing efficiency of the germline expression method (Option 4) is often higher than that of direct delivery (Option 1 and 2) [6]. In addition to the step-by-step standard protocol, we also provide some creative examples of how HDR can be used to manipulate the genome of mosquitoes providing versatile tools to study gene function. It should be noted that while this chapter focuses on genome engineering of *Ae. aegypti*, many aspects of these protocols could be adapted to other mosquito vectors, in addition to other insects and animals in general.

2 Materials (a List of all Materials)

2.1 CRISPR Target Site Selection and Validation

1. Reagents for genomic DNA extraction. We use a commercial DNeasy blood and tissue kit.
2. Custom-designed oligonucleotides.
3. Thermal cycler.
4. Reagents for PCR product purification. We use a commercial Gel DNA Recovery Kit.

2.2 Injection Mixture Preparation

1. sgRNA designed by the user. We have used the sgRNA synthesis service offered by Synthego Corporation.
2. NanoDrop 2000 spectrophotometer or similar instrument capable of DNA/RNA concentration measurement.
3. Q5 high-fidelity DNA polymerase (New England BioLabs Inc.).
4. MegaClear columns (Life Technologies).
5. Nuclease-free dH_2O.
6. Ampure XP magnetic beads (Beckman Coulter).
7. MegaScript T7 (Ambion) or similar.
8. Primers for generating DNA template of sgRNA using example target sequence (Figs. 3 and 4):

   sgRNA-F (example target sequence underlined. Replace it with your target sequence): 5′- GAAATTAATACGACTCAC TATAGGATCCTCATGACATTACATGTTTTAGAGCTA GAAATAGC-3′.

   sgRNA-R (universal): 5′- AAAAGCACCGACTCGGTGC CACTTTTTCAAGTTGATAACGGACTAGCCTTATTT TAACTTGCTATTTCTAGCTCTAAAAC-3′.
9. Agilent Bioanalyzer 2100 with RNA 6000 Nano Kit.
10. pMLM3613 (Addgene, 42251).
11. mMESSAGE mMACHINE T7 Ultra Transcription Kit (Life Technologies) or similar.
12. Cas9 mRNA finished product SpCas9 mRNA (ARCA, 5mCTP, ψUTP) (ApexBio, R1006).
13. Recombinant Cas9 protein. Several commercial sources are available, we have used PNA Bio, CP01).
14. Cas9-expressing transgenic mosquito strain. We recommend Exu-Cas9 strain expressing Cas9 under the *Exuperentia* promoter and marked with Opie2-dsRed from the Akbari Lab [6].
15. Plasmid Maxiprep Kit.
16. *piggyBac* plasmid pBac-3xP3-dsRed (AAEL010097-Cas9) (Addgene, 100707).

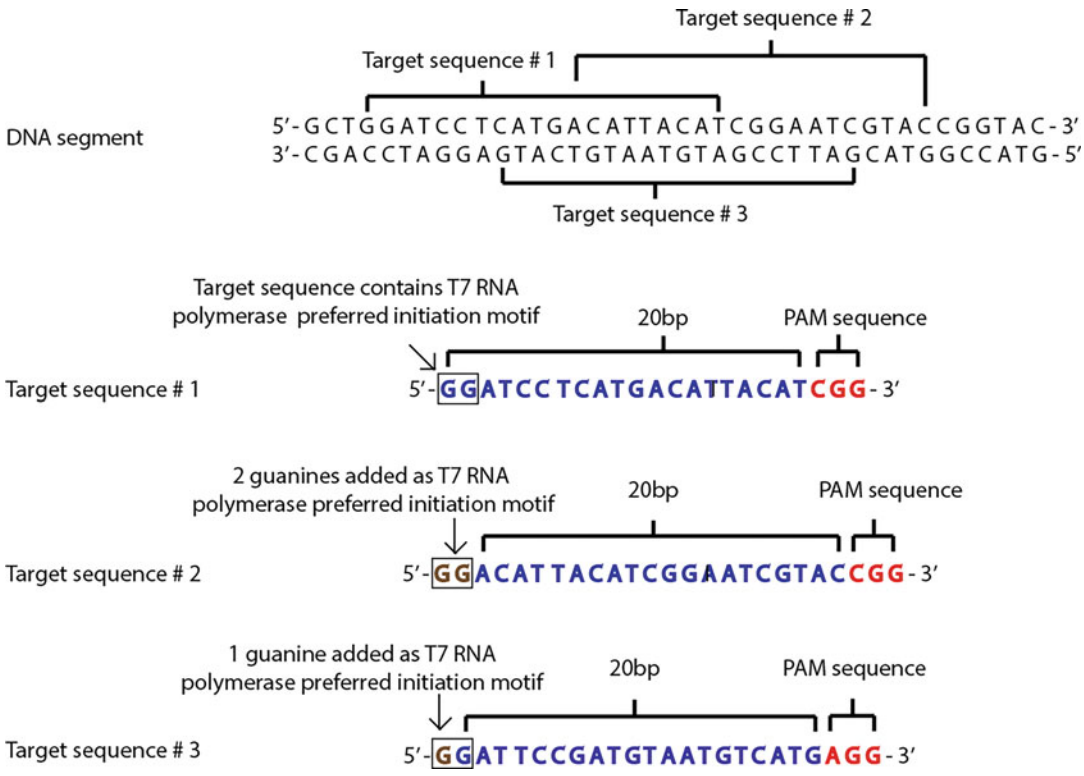


Fig. 3 Target sequence selections of an example DNA segment from the *Aedes aegypti* genome

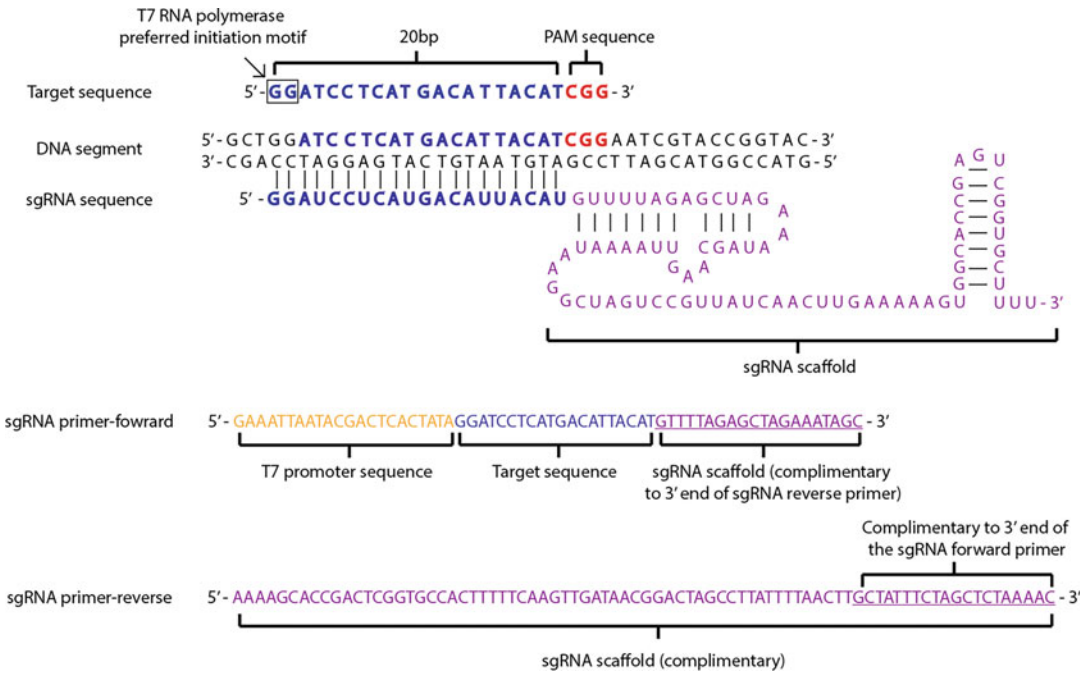


Fig. 4 sgRNA primer design. An example of a sgRNA designed using target sequence #1 from Fig. 3. The primers for generating the sgRNA are depicted and annotated

17. Gibson Assembly Master Mix (New England BioLabs Inc.).
18. RNaseZap (Ambion).
19. Primers for amplifying 3xP3-DsRed fragment:
    3xP3-DsRedF: 5′- AAAGTGGCACCGAGTCGGTGCTTTTGCTAGCAATTCGAGCTCGCCCGGGGATCT-3′.
    3xP3-DsRedR: 5′- AACATTGTCAGATCCGAGATCGGCCGGCCTAGAAGCTTTAAGATACATTGATGAGTTTGGAC-3′.
20. BmrI.
21. BsrI.
22. Plasmid pSL1180-HR-PUbECFP (*see* **Note 13**) (Addgene, 47917).
23. Plasmid PSL1180polyUBdsRED (*see* **Note 13**) (Addgene, 49327).

2.3 Mosquito Husbandry and Embryonic Microinjection

1. *Aedes aegypti* Liverpool strain (BEI resources, #735).
2. Plastic containers (34.6 cm × 21 cm × 12.4 cm) for larvae rearing (Sterilite).
3. Crushed TetraMin Tropical Flakes (Tetra Werke).
4. Mosquito rearing cages: BugDorm-1 Insect Rearing Cage (24.5 cm × 24.5 cm × 24.5 cm) (Bugdorm, DP1000).
5. Sucrose.
6. Mice.
7. Ketamine/Xylazine/sterile saline.
8. Membrane feeding system (Hemotek, SP6W1-3).
9. Plastic cup.
10. Fine tip paintbrush for positioning embryos (ZEM, #2595).
11. Filter paper (Whatman), #1825-021.
12. Double-sided sticky tape.
13. Disposable cover slip.
14. Microscope Slides.
15. Halocarbon oil 700 (Sigma-Aldrich).
16. Halocarbon oil 27 (Sigma-Aldrich).
17. Quartz capillary glass tube (Sutter Instrument, QF100-70-10).
18. Micropipette pullers (Sutter Instrument, P-1000 and P-2000).
19. Microloader tips (Eppendorf, 930001007).
20. Micromanipulator (Sutter Instrument, MM-33R).
21. FemtoJet 4× (Eppendorf).

22. Needle Beveler (Sutter Instrument, BV-10).
23. Vacuum Drying Oven (Across International, AT19e).
24. Nutating Shaker (Benchmark, B3D1320-E).
25. Nalgene transparent polycarbonate classic design desiccator (Thermo Scientific, 5311-0250).
26. 16 oz. plastic cups.
27. Nitrogen gas (research purity 99.9999%; Matheson, G2173112).
28. Nitrogen gas regulator (Miller Electric, 210-4109).
29. Drierite granules (Carolina Biological Supply Company, 858963).

2.4 Genetic Crosses, Screening for Mutations and Generation of Homozygous Mutant Lines

1. Polystyrene vial (Genesee Scientific, 32-116).
2. Fluorescent stereo microscope (Leica M165FC).
3. Fluorescent Filters stereo microscope.
   (a) ECFP—Leica Part #10447409; Excitation ET436/20×, Emission ET480/40 m wavelengths/band-pass.
   (b) EGFP—Leica Part #10447408; Excitation ET470/40×, Emission ET525/50 m wavelengths/band-pass.
   (c) EYFP—Leica Part #10447410; Excitation ET500/20×, Emission ET535/30 m wavelengths/band-pass.
   (d) RFP/DsRed/tdTomato/mCherry Leica Part #10447410; Excitation ET500/20×, Emission ET535/30 m wavelengths/band-pass.
   (e) ECFP/EGFP/RFP -Leica Part #10450611; ET434.5/21, 501.5/19, 574.5/23, ET469.5/25, 536.5/29, 635.5/69 wavelengths.
4. LED light source for the Leica M165FC (Lumencor, SOLA-III).
5. Microscope camera (Leica, DMC2900).

3 Methods

3.1 CRISPR/Cas9-Mediated Knockout

3.1.1 Selection and Validation of CRISPR Target Sites

1. Select target genes for editing based on project requirements, and determine the target region based on the gene structures, transcripts, and exon–intron junction boundaries (Fig. 3, *see* **Note 1**). If producing sgRNA in house using T7 in vitro transcription, then the 5′ target sequences need to be considered as well as the 3′ PAM sequence. The 3′ PAM sequence has a specified pattern of "NGG," where N represents any of the four nucleotides (A,T,C,G), while G refers to guanine. There are multiple online programs to aid in the design of effective

Table 1
PCR reagents and program setting for target site validation from genomic DNA

PCR reagents		PCR program
5× Q5 reaction buffer	20 µl	**Initial denaturation step** 98 °C - 30 s **Amplification step (35 cycles)**: 98 °C, 10 s (denaturation) 58 °C, 10 s (primer anneal) 72 °C, 30 s (extension) **Final extension** 72 °C, 2 min **Storage** 4 °C—∞
10 mM dNTPs	2 µl	
10 µM forward primer	5 µl	
10 µM reverse primer	5 µl	
Genomic DNA extracted from *ae. Aegypti* (50 ng)	2 µl	
Q5 high-fidelity DNA polymerase	1 µl	
Nuclease-free water	67 µl	
Total	**100 µl**	

sgRNAs with these parameters including CHOPCHOP V3.0.0 or CRISPR software to minimize potential genomic off-target events (Fig. 4) (*see* **Note 1**). T7 RNA polymerase starts transcription most efficiently if the first two nucleotides to be transcribed are 5′-GG (T7 RNA polymerase preferred initiation motif). A common recommendation is to add the prefix GG- if the guide does not start with G (5′-N20-(NGG)-3′), to add G- if your guide starts with a single G (5′-GN19-(NGG)-3′) and to not add anything if your guide starts with GG already (5′-GGN18-(NGG)-3′) (Figs. 3 and 4, *see* **Notes 1** and **2**).

2. Validate the target DNA sequence in the *Ae. aegypti* genetic background to be used by polymerase chain reaction (PCR) amplification and Sanger sequencing (Table 1). Briefly, pool 5–10 *Ae. aegypti* individuals in one sample (three replicates) and extract genomic DNA using the DNeasy blood and tissue kit. Design PCR primers to amplify the target region via PCR with approximately 200–300 bp on each side of the target (desired PCR size of 500–650 bp). Then purify the PCR product and send the purified PCR product for Sanger sequencing using primers on each end of the amplified DNA. This is an important step because the reference genome sequence of *Ae. aegypti*, and many other insect species, is imperfect. There may be single-nucleotide polymorphisms, insertions and deletions in the target regions of in-house laboratory strains relative to the genome assembly strains which would render sgRNA designs based on the reference genome ineffective [13].

3.1.2 Preparation of the Injection Mixture

1. Generate sgRNAs via one of the three options: (1) use commercial synthesis service providers; (2) generate in-house using a synthesis protocol for T7-mediated in vitro transcription [14]

(*see* **Note 3**); or 3) generate a sgRNA-expressing plasmid (*see* **Note 4**) [15]. Prepare in vitro sgRNAs and sgRNA-expressing plasmids in ultrapure, molecular grade, nuclease-free dH_2O. Determine the concentration using a spectrophotometer, aliquot, and store at −80 °C.

2. Prepare the Cas9 component via one of the three options: (1) use Cas9 mRNA which could be generated from plasmid pMLM3613 using the mMessage mMachine T7 Ultra Transcription kit [5] or purchased from an external vendor; (2) use recombinant *Streptococcus pyogenes* Cas9 protein which could be purchased from an external vendor; or (3) use embryos derived from Cas9-expressing transgenic mosquitoes (e.g., Exu-Cas9 line marked with opie2-dsRed) for injection. These transgenic mosquito lines are readily available in the mosquito research community [6] (*see* **Note 5**) and work quite efficiently [16]. For options 1 and 2, prepare the Cas9 source in nuclease-free dH_2O. Determine the concentration using a spectrophotometer, aliquot, and store at −80 °C.
3. Prepare sgRNA/Cas9 mixtures for microinjection as in Table 2. Prepare injection mixtures with nuclease-free dH_2O and store mixture in aliquots at −80 °C until use.

3.1.3 Mosquito Husbandry and Embryonic Microinjection

1. Larvae from the wild-type *Ae. aegypti* Liverpool strain (wild-type [WT]) and Exu-Cas9 strains are reared separately at a set density (~200 larvae in 3 l of deionized H_2O (dH_2O)) in plastic containers. Larvae are fed fish food daily until pupation using approximately 50–1300 mg per day. Less food per day (~50 mg/200 larvae) is required while larvae are in the first few instars (L1–L2) and more food (~1300 mg/200 larvae) is needed in the later instar stages (L2–L4) until pupation. If the containers are overcrowded above this recommended density, the larvae will take longer to pupate, and the adults will be

Table 2
sgRNA/Cas9 final concentration for microinjection (* most efficient methods)

	Final concentration (ng/ul)			
Injection mixture type	**sgRNA plasmid**	**sgRNA**	**Cas9 mRNA**	**Cas9 protein**
sgRNA + Cas9 mRNA	–	100	300	–
sgRNA + Cas9 protein	–	100	–	100
*sgRNA + Exu-Cas9 line	–	100	–	–
sgRNA plasmid + Cas9 mRNA	300	–	300	–
sgRNA plasmid + Cas9 protein	300	–	–	100
*sgRNA plasmid + Exu-Cas9 line	300	–	–	–

smaller and weaker than in a less crowded container. Overfeeding the larvae may also cause them to die due to bacterial and yeast blooms. As such, we strongly recommend standardizing the rearing procedure and feeding regimen according to larval density to achieve synchronous pupation and uniform adult size.

2. Collect *Ae. aegypti* pupae (n = 200-300) from the target genetic background to be used as a source of eggs for embryonic microinjection into cups containing ~100 ml water. Place pupal cups for adult eclosion into a holding cage maintained at 27 ± 1 °C, with 50–80% humidity, and 12:12 (light:dark) photoperiod. Provide the adults with 0.3 M aqueous sucrose or glucose in a container with a filter paper or cotton roll capable of saturation by capillary action for ad libitum adult feeding, and allow adults to mate freely in the cage for 4–5 days. Blood-feed the females from each cage with anesthetized mice until satiated (approximately 15 min at a time) for two consecutive days (*see* **Note 6**). Verify that no sources of moisture other than the sugar feeder are present in the cage, as adults may prematurely lay on this surface, preventing timed oviposition required for injections.
3. On the day of embryonic microinjections (72 h post second blood meal), create an oviposition cup (Fig. 5b) and place it into the cage. Move the holding cage to a dark location (~20 min) such as a closed cupboard or cardboard box to collect eggs (Fig. 5a–c) (*see* **Note 7**). Blowing on, or agitating the cage, in addition to heating the oviposition cup to just above room temperature may assist in stimulating laying.
4. Align individual eggs side-by-side on ddH_2O moist filter paper (sufficiently wet to keep the eggs moist without a meniscus forming) using a fine-tip paintbrush. Ensure the narrow posterior poles of the embryos are all oriented in the same direction for microinjection. After approximately 50 embryos that are white to light gray in coloration are aligned, absorb the water from the wet filter paper used as an alignment substrate with fresh filter paper (to soak up residual moisture to help the embryos stick better to the double sided tape in the next step).
5. To transfer embryos onto a glass substrate for microinjection, stick a piece of double-sided sticky tape to a glass coverslip, and gently press the coverslip against the surface of the aligned embryos. Invert such that the eggs are on top, and adhere the cover slip to a microscope slide with double-sided sticky tape to fix the cover slip.
6. Immediately cover the embryos with water-saturated halocarbon oil, consisting of a mixture of 9 ml halocarbon oil 700, 1 ml halocarbon oil 27, and 20 ml ddH_2O (Fig. 5d).

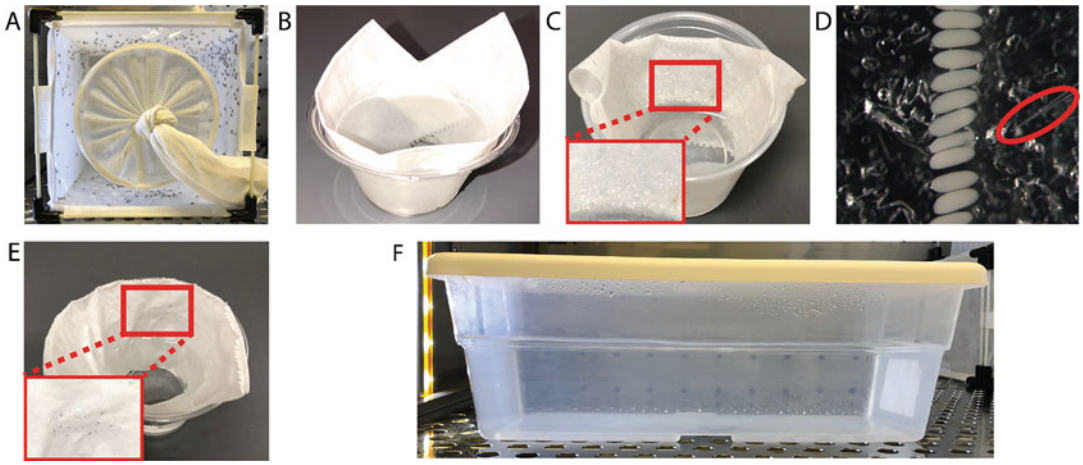


Fig. 5 Mosquito embryonic microinjection procedure. (**a**) Mosquito rearing cage holding blood-fed females. (**b**) An oviposition cup is introduced into the rearing cage. (**c**) Freshly laid embryos (white when initially laid) are collected using an oviposition cup. (**d**) A row of embryos (still white/light gray in color) is then aligned in the same orientation and microinjected in their posterior poles to target the developing germline (the red circle indicates the microinjection needle). (**e**) Post-injection, embryos recover for 4 days in a moist oviposition cup placed in the incubator under normal mosquito rearing conditions (27 ± 1 °C, with 50–80% humidity, and 12:12 (light:dark) photoperiod). (**f**) Embryos are then hatched in deionized H_2O (dH_2O) in a vacuum chamber (27 °C, 20 psi, 1 h)

7. Thaw the injection mix on ice. Prepare glass needles for injection (*see* **Note 8**). When embryos are white/light gray in color (about 40 min old), load the needle with 2 μl of injection mixture using Microloader Tips while avoiding bubble formation, and then mount it on the micromanipulator. Carefully insert the needle into the posterior pole of the embryo and inject the mixture using an Eppendorf Femtojet 4× with a typical injection pressure (P_I) of 30–50 psi and injection time (T_I) of 0.1–0.3 s resulting in the injection of a quantity of about 10% of the volume of the embryo (*see* **Note 9**). To minimize capillary back-flow/blockage, set the Femtojet to a constant compensation pressure (P_c) of ~1 psi. For best results, inject the embryos with the needle at an angle slightly greater than 45° relative to the embryo surface in the posterior quarter of the embryo.
8. After injection, carefully remove the halocarbon oil within 20 min post injection using a kimwipe to absorb the oil. Eggs are then removed gently from the coverslip with a clean paintbrush and transferred to wet filter paper to recover and develop for 4 days at 27 °C (Fig. 5e).
9. Four days after injection, hatch injected G0 eggs in ddH_2O in a vacuum drying oven (27 °C, 20 psi, 1 h) (Fig. 5f). Then follow normal larval rearing procedures, with special care to prevent inter-larval competition due to underfeeding or high density. For non-lethal gene targets, we typically expect 40–80% of embryos to hatch.

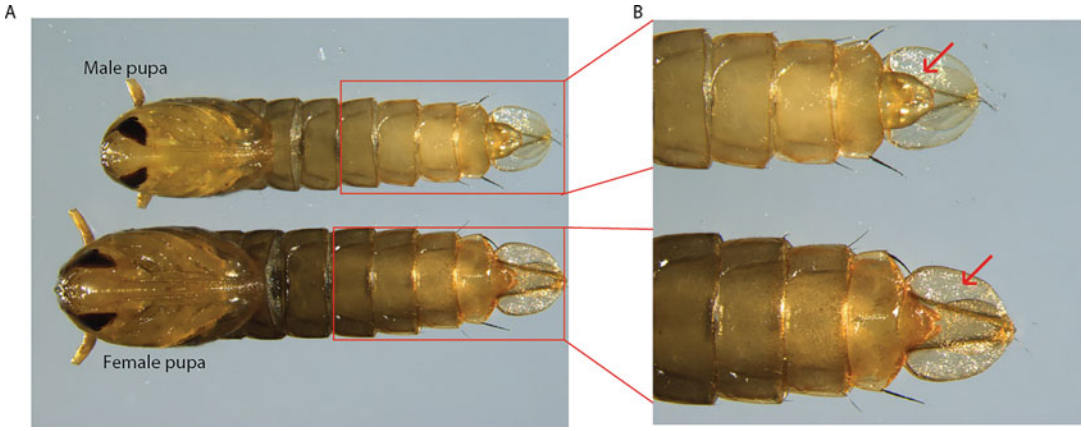


Fig. 6 Sex difference between male and female *Aedes aegypti* pupa. (**a**) A male and female pupa (ventral view). (**b**) Magnified view of pupal genitalia. *Ae. aegypti* males tend to be smaller than females and pupae can be reliably sexed by the morphological differences in the genital lobe shape (*see* the red arrows at the end of the pupal abdominal segments just below the paddles)

3.1.4 Genetic Crosses, Screening for Mutations, and Generation of Homozygous Mutant Lines

1. Rear larvae to the pupal stage (Fig. 6). Sort and separate surviving pupae (G0) into ♀ or ♂ cups (1 pupa/cup to enable isolation of each end-joining event) (Fig. 6). Add opposite sex of WT pupae into each cup at 7:1 ratios (WT:G0). If disrupting a visual marker gene (e.g., *white*), then a somatic mosaic mutant phenotype should be readily visible at G0 larval, pupal and adult stages [9]. The smallest cup size mosquitoes will reliably mate in for single pair or small group crosses is 16 oz. solo cups. A common mistake is trying to mate mosquitoes in smaller containers and vials (e.g., *Drosophila* vials) which is not recommended. If identifying/isolating independent end-joining events is not a priority, then batch mating can also be done here where G0's can be mated in larger cages as pools of males/females with excess virgin wild-type females/males, respectively.
2. Four days post-eclosion (enabling sufficient time for development and mating), provide a blood meal to the adult females.
3. Three days post-blood meal, individually transfer each blood fed ♀ to a single narrow polystyrene vial with a wet paper towel or a wet cone of filter paper to attract females to lay eggs (Fig. 7).
4. Four days later, hatch G1 eggs in deoxygenated deionized H_2O (ddH_2O) using a vacuum drying oven (27 °C, 20 psi, 1 h). Then follow normal larval rearing procedures. If a vacuum drying oven is not available, 2 days post-oviposition condition the wet egg papers for synchronous hatching by removing them individually from their vials and drying them on fresh paper towels to soak up residual moisture for 5 min at room

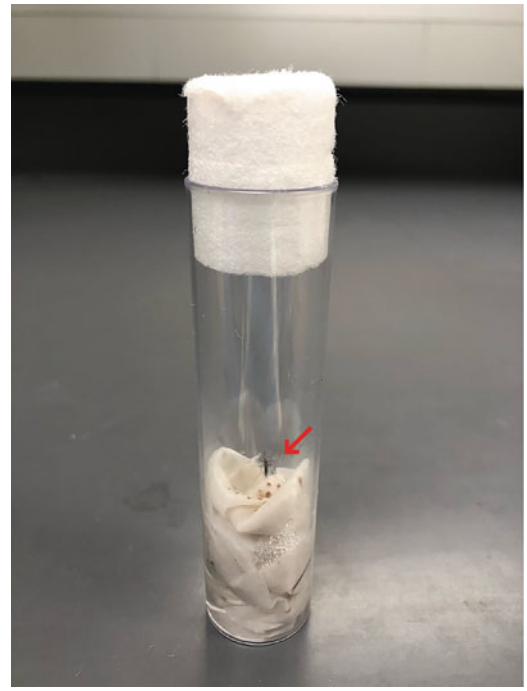


Fig. 7 A G0 female in a vial. After mating and obtaining a blood meal, individual females (red arrow) are isolated in a single narrow polystyrene vial with a wet paper towel or a wet cone of filter paper for egg oviposition

temperature. Dry the inside of the original oviposition vials with kimwipes to remove any residual water, and place the moist oviposition papers that have been drying back into their respective polystyrene vials and plug using cotton stoppers. Further condition these egg papers for 2 additional days in the insectary to allow the papers in each vial to dry fully. Hatch conditioned eggs in ddH_2O that has been boiled, capped in a bottle, and cooled to RT, and rear as specified in Subheading 3.1.3.

5. Select 10 larvae (G1) randomly from each isolated colony (should have up to 40–80 larvae per female), pool and extract DNA from these larvae, and conduct PCR and Sanger sequencing to check for CRISPR/Cas9-induced mutations in the target site. Only keep the colony with positive mutant genotypes. In addition to validating the target site mutations, we recommend checking for off-target effects (*see* **Note 10**).
6. For colonies with evidence of CRISPR-induced mutations, use the remaining larvae to set up multiple single pair crosses of G1s (1 G1♂ × 1G1♀) using sexed pupae to ensure virginity of females (Fig. 6). Alternatively, outcross G1 individuals to wild-type mosquitoes. After blood feeding these crosses and obtaining eggs, genotype the candidate G1 mutant parents with PCR

and Sanger sequencing or restriction fragment analysis. Hatch the G2 eggs from families containing mutant alleles, rear to adulthood and sex at the pupal stage to ensure virginity.

7. Repeat **step 6** as desired, until a homozygous mutant line is generated (can take up to 5–10 generations).

3.2 CRISPR/Cas9-Mediated Knock-in

3.2.1 Selection and Validation of CRISPR Target Site

1. Same as Subheading 3.1.1.

3.2.2 Preparation of the Injection Mixture

1. Design and assemble the HDR donor plasmid using Gibson assembly (*see* **Note 11**) [15]. An assembled HDR plasmid generally contains the following elements:
   (a) A fluorescent marker for transgenesis. We often use a 3xP3-DsRed marker, which is amplified from a pBac-3xP3-dsRed vector using primers 3xP3-DsRedF and 3xP3-DsRedR. Depending on the genomic site of integration, this cassette may express the DsRed-dominant fluorescent marker in larval photoreceptors, ventral nerve cord, and anal papillae, as well as the pupal and adult eyes with varying fluorescence intensity (*see* **Note 12**). If future removal of the chosen fluorescent marker from mosquitoes is conceivably desirable for your application, add two LoxP sites in the same orientation to flank the marker transgene to facilitate Cre-LoxP-mediated excision after line establishment [17].
   (b) Two homology arms (500–1000 bp) flanking the target cleavage site that will be used to facilitate HDR insertion of the donor DNA fragment. HDR arms are amplified from *Ae. aegypti* genomic DNA with designed primer pairs (a pair for the 5′ arm and a pair for the 3′ arm). One HDR arm may proceed up until the cleavage site (i.e., 3 bp 5′ to NGG), and the other arm may exclude the six nucleotides 3′ to this cleavage site inclusive of the PAM. It is common for donor arms to also completely exclude the sgRNA sequence completely.
   (c) Donor DNA fragment of interest (Fig. 8). The fragment for insertion is at the discretion of the user and may serve as a visual marker for gene knockout or also be useful from a functional perspective. Potential applications of HDR in *Ae. aegypti* include: (1) inserting a fluorescent protein into the ORF of a target gene for gene knockout and simplifying the goal of following the mutation during stock maintenance, crossing, and generating homozygous mutants

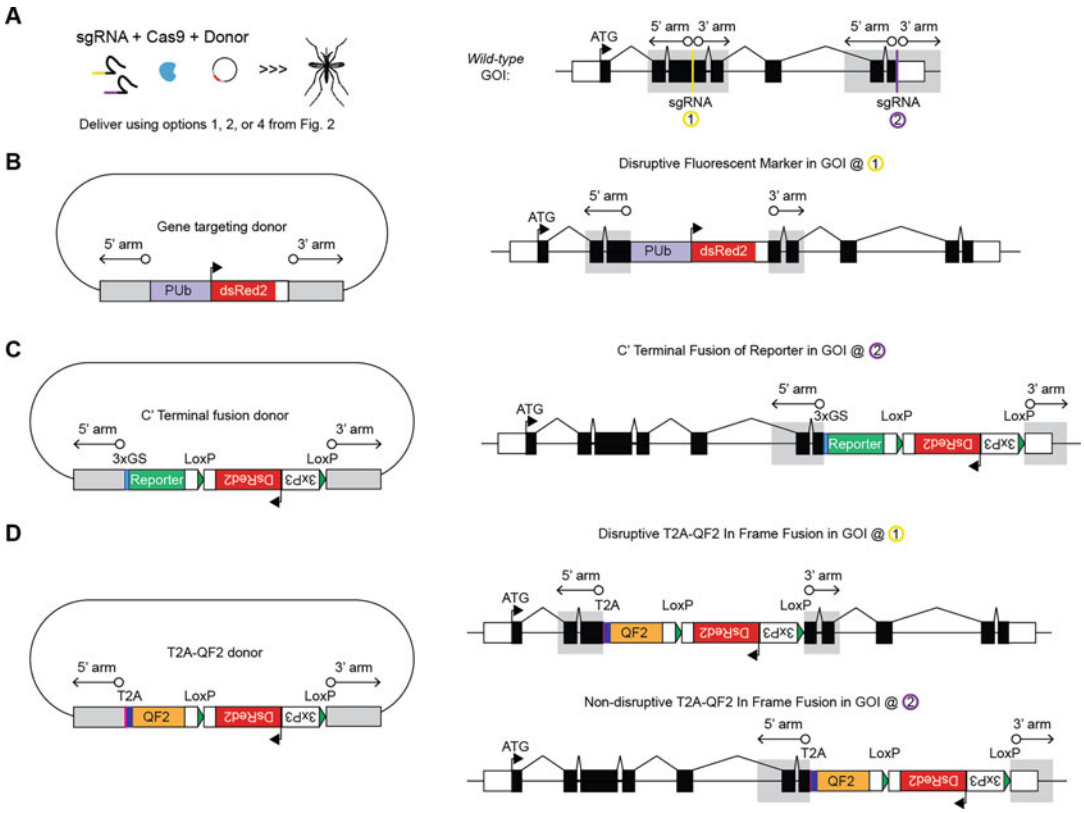


Fig. 8 CRISPR-Cas9 HDR in *Aedes aegypti*. (**a**) A Gene of Interest (GOI) is targeted with CRISPR/Cas9-mediated HDR for insertion of a user-specified donor DNA fragment using a mixture of sgRNA, Cas9, and donor plasmid (left). These three editing components may be delivered to embryos using multiple options (Fig. 2). Numerous sgRNA target regions in a GOI (right) may be chosen as sites for fragment insertion depending on the envisaged application with HDR donor plasmids (example configurations are provided in **b**–**d**); (**b**) Gene targeting donor plasmid (left) for insertion of a disruptive constitutive fluorescent marker in a GOI at hypothetical sgRNA site 1 (right); (**c**) C′ terminal fusion donor plasmid (left) for tagging the C′ terminal end of the GOI protein with a reporter of choice. HDR is targeted to hypothetical sgRNA site 2 which is immediately upstream of the stop codon (right); (**d**) T2A-QF2 donor plasmid (left) for generation of a T2A-QF2 In Frame Fusion in a GOI. The T2A-QF2 cassette may be inserted by HDR at hypothetical sgRNA site 1, or alternatively hypothetical sgRNA site 2 to potentially disrupt (right top) or conserve (right bottom) functionality of the targeted allele, respectively; while facilitating genetic access to the specific cell type expressing the GOI (*see* **Note 13**)

[5, 6, 18]; (2) tagging a gene on either the 5′ or 3′ end with a fluorescent protein [19]; and (3) performing T2A-driver In Frame Fusions to place the expression of the driver component of binary expression systems such as QF2 [20] and others [21], under the endogenous regulatory control of the target gene's enhancer and promoter elements to faithfully mimic its expression pattern [17, 22]. These driver lines can then be crossed with

Table 3
sgRNA/Cas9/donor plasmid final concentration for microinjection (* denotes the most efficient methods in our experience)

Injection mixture type	Final concentration (ng/ul)				
	sgRNA plasmid	sgRNA	Cas9 mRNA	Cas9 protein	Donor plasmid DNA
sgRNA + Cas9 mRNA + donor plasmid	–	100	300	–	100
sgRNA + Cas9 protein + donor plasmid	–	100	–	100	100
*sgRNA + Exu-Cas9 line + donor plasmid	–	100	–	–	100
gRNA plasmid + Cas9 mRNA + donor plasmid	300	–	300	–	100
*gRNA plasmid + Cas9 protein + donor plasmid	300	–	–	100	100
gRNA plasmid + Exu-Cas9 line + donor plasmid	300	–	–	–	100

various engineered responder lines for functional and genetic studies in *Ae. aegypti* (*see* **Note 13**).

For examples of base donor constructs suitable for user-specific insertion of HDR arms to generate null mutations or T2A-QF2 in frame fusions in target genes of choice, *see* **Note 14**.

2. To obtain high-quality, ultrapure, and endotoxin-free donor plasmid for microinjection, maxi-prep the assembled HDR plasmid using a Plasmid Maxiprep kit. Confirm the plasmid sequence with Sanger sequencing. Plasmids can be stored at −80 °C more than a year before being used for embryonic microinjection.
3. Prepare sgRNA/Cas9 mixtures for microinjection as in Table 3. Prepare injection mixtures with nuclease-free dH_2O and store mixture in aliquots at −80 °C until use.

3.2.3 Mosquito Husbandry and Embryonic Microinjection

1. Same as Subheading 3.1.3.

3.2.4 Genetic Crosses, Screening for HDR Insertions, and Generation of Homozygous Mutant Lines

1. Rear injected G0 larvae to the pupal stage, sort and separate surviving pupae into two cages (♀ or ♂ cage) every day until the last larva develop to pupa. Add WT virgin adults of the opposite sex into each cage at 7:1 ratios (WT:G0) (*see* **Note 15**).

2. Same as Subheading 3.1.4, **step 2**.
3. Three days post-blood meal, provide an oviposition cup for each cage.
4. Same as Subheading 3.1.4, **step 4**.
5. Screen G1 larvae (stage 3 or 4) with a fluorescent stereomicroscope and isolate the larvae with positive fluorescent markers.
6. Rear all larvae to the pupal stage, sex and separate all G1 pupae into ♀ or ♂ cups (1 pupa/solo cup or batch mate according to user specifications). Add opposite sex of WT virgin adults into each cage at 7:1 ratios (WT:G1).
7. Same as Subheading 3.1.4, **step 2**.
8. Same as **step 3**.
9. Same as Subheading 3.1.4, **step 4**.
10. Select 10 larvae (G2) with markers randomly from each colony and extract pooled DNA from these larvae. Perform a diagnostic PCR with a primer anchored within the donor DNA fragment and another primer anchored in genomic DNA outside the HDR arms. Sanger sequence the resulting PCR products to validate precise insertion of the HDR donor cassette. Only maintain the mosquito colonies with the correct insertion.
11. Set up multiple single pair crosses of G2s (1 G2♂ × 1 G2♀) from the colony that have been identified with a precise insertion in the above step. Alternatively, outcross these positive individuals to wild-type mosquitoes for five or more generations to purge potential off-target mutations (*see* **Note 10**) before generating homozygous lines.
12. Repeat **step 11** until generating the homozygous insertion lines.

4 Notes

1. To define putative sgRNA genomic target sites, we suggest considering several factors. First, confirming transcriptional expression of the target region, and looking for conservation between other species, will help define putative *Ae. aegypti* genomic target regions that have a higher probability of being necessary for gene function (assuming the goal is to disrupt gene function). To do this, we recommend using both available *Ae. aegypti* transcriptional databases and simple NCBI-BLAST searches (www.vector.caltech.edu) [23, 24]. Second, once general target regions are defined, the putative sgRNA target sites can be identified by simply scanning both the sense and antisense strands for the presence of the NGG-PAMs manually either by eye or by utilizing available software such as

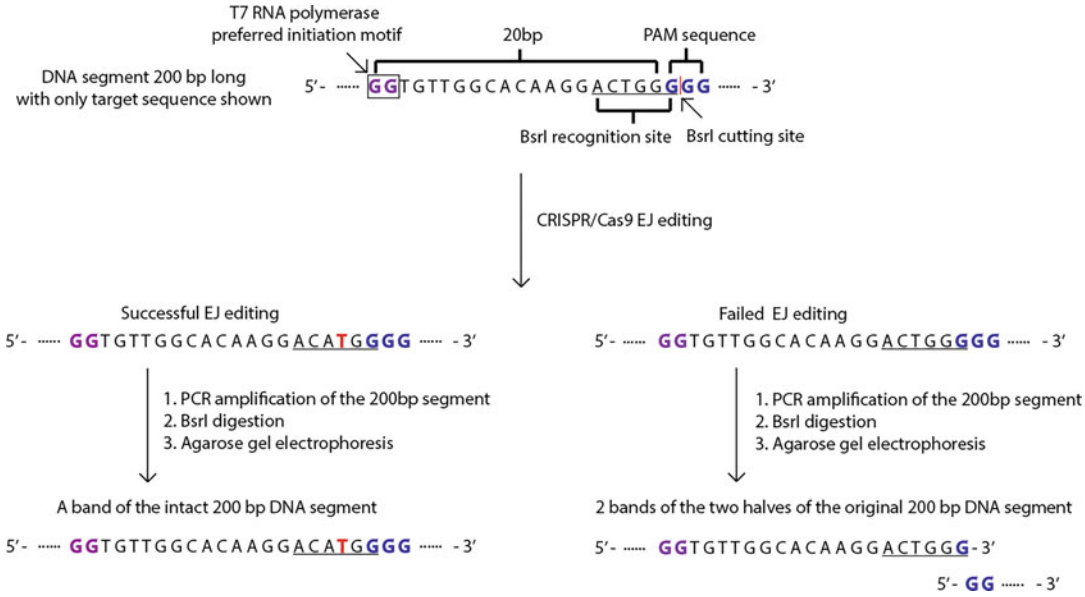


Fig. 9 Alternative post-editing mutational screening method by leveraging restriction enzymes. An example of incorporating the BsrI restriction enzyme recognition pattern in the target sequence design, along with an expedited genotyping workflow is shown

CHOPCHOP v3 [25] , and/or local sgRNA Cas9 packages [26]. Finally, to minimize potential off-target effects, we recommend confirming specificity of the sgRNAs using publicly available bioinformatic tools, for example, NCBI-Blast [27], Blat [27, 28] and selecting the most specific sgRNAs within the specified target regions with the least potential off-target binding sites. It should be noted that even if the chosen sgRNA target sequences fulfill all of the above requirements, sgRNA specificity and activity is not guaranteed. Therefore, we recommend designing multiple different sgRNAs to target the exonic coding sequences and co-injecting them in order to increase the chance of successful editing.

2. To expedite downstream genotyping after gene editing, we recommend selecting a target site with a PAM sequence that is near/within a restriction enzyme cutting site to ensure Cas9-induced NHEJ events can destroy the restriction enzyme cut site. Cas9 typically cleaves DNA 3 bp 5′ of the PAM site, and due to error-prone EJ, the nucleotide sequence spanning this genomic region may be disrupted (Fig. 9, Table 4). Therefore, many restriction enzymes could be chosen, and care should be taken to analyze/choose Cas9 target sequences with restriction sites near the PAM sequence. For example, for the hypothetical sgRNA target sequence listed in Table 4, BmrI which recognizes 5′-ACTGGG$(N)_5$-3′ or BsrI which recognizes 5′-ACTGG(N)-3′ is suitable as Cas9 cutting would likely

Table 4
Hypothetical sgRNA target site with overlapping BmrI and BsrI restriction enzyme sites (purple: T7 preferred motif; underlined/bold: restriction site; blue: PAM)

Enzyme	Pattern	Guide DNA target site with Restriction Site	Suppliers
BmrI		5'- GGACATGCCACAAG G**ACTGGGGG**-3'	NEB (Cat # R0600S)
BsrI		5'- GGTGTTGGCACAAG G**ACTGGGGG**-3'	NEB (Cat # R0527S)

disrupt these recognition sequences (Fig. 9, Table 4). This way, when the target sequence is successfully edited by CRISPR/Cas9 to yield EJ alleles, it would no longer be recognized by restriction enzymes. As a result, rather than sequencing every single edited individual's DNA samples, genotyping can be done by simply PCR amplifying the target region with a few hundred base pairs on each side, then cutting the PCR product with the dedicated restriction enzyme and checking the size of the resulting products on agarose gels. Restriction enzymes chosen for genotyping using this method therefore need to be uniquely selected based on the nucleotide sequence of each particular sgRNA target region. Other methods to screen for potential EJ events include High-Resolution Melting Analysis (HRMA) [18] and T7E1/Surveyor Nuclease assays [29] that detect nucleotide mismatches in a PCR amplicon from the target region. However, we caution that when using these alternative methods, wild-type nucleotide polymorphisms present in the mosquito genetic background used for injections or outcrossing may yield false positives for EJ events. Knock-in of a single-stranded oligo donor (ssODN) containing a unique restriction enzyme recognition sequence into the CRISPR cut site can also be used as a rapid method for mosquito genotyping [5].

3. For in-house in vitro sgRNA synthesis, first linear double-stranded DNA templates for sgRNAs are generated via template-free PCR with Q5 high-fidelity DNA polymerase

Table 5
PCR reagents and program setting for generating DNA templates for sgRNAs

PCR reagents		PCR program
5× Q5 reaction buffer	20 μl	**Initial denaturation step** 98 °C, 30 s **Amplification step (35 cycles)**: 98 °C, 10 s (denaturation) 58 °C, 10 s (primer anneal) 72 °C, 10 s (extension) **Final extension** 72 °C, 2 min **Storage** 4 °C—∞
10 mM dNTPs	2 μl	
10 μM forward primer (sgRNA-F)	5 μl	
10 μM reverse primer (sgRNA-R)	5 μl	
Q5 high-fidelity DNA polymerase	1 μl	
Nuclease-free water	67 μl	
Total	**100 μl**	

(NEB) using a forward primer for the sgRNA (sgRNA-F) and universal reverse primer that can be used for all targets (sgRNA-R) (Table 5). The conditions for template PCR are programmed as detailed in Table 5. PCR products (the DNA templates for sgRNAs) are confirmed on a 1% agarose gel by loading 2 μl and visualizing a single band of approximately 125 bp in size. The resulting PCR product is purified with Ampure XP magnetic beads using standard protocols and sgRNAs generated by in vitro transcription using the Ambion MegaScript T7 kit (AM1334; Life Technologies) using 300 ng purified DNA as template in an overnight reaction incubated at 37 °C (Table 6). The resulting RNA is purified using MegaClear columns following the manufacturer's protocols (AM1908; Life Technologies). The concentration of the sgRNA can be confirmed on a spectrophotometer (expect anywhere from 5 to 100 μg). Verify the size and concentration using an Agilent Bioanlyzer with the RNA 6000 nano kit or Tapestation. The sgRNA should appear as a single band without obvious degradation products. Due to the secondary structure of the sgRNA, the sizing may not be accurate, but do expect a sharp band at the approximate size with little evidence of degradation. Purified sgRNAs are then diluted to 1 μg/μl in nuclease-free water, aliquoted, and stored at −80 °C until use. It is very important to work under RNase-free conditions when producing or working with RNA. Be sure to use nuclease-free consumables including filter tips and microfuge tubes. Also, thoroughly clean the work area including the microinjection apparatus, gloves, and pipettes with RNaseZap before conducting experiments. Given that not all sgRNAs function efficiently, and specificity and activity are unpredictable, we recommend designing multiple sgRNAs for each target gene to increase probability of generating desired modifications of target

Table 6
In vitro transcription of sgRNAs

IVT reaction components	Volume (μl)	Incubation period
Nuclease-free water	To 20 μl	2 h (minimum) to overnight (12–16 h maximum) at 37 °C
Free ribonucleotides (ATP, CTP,GTP,UTP)	2 μl each (8 μl total)	Then add 1ul turbo DNAse and incubate for an additional 15 min at 37 °C
10× reaction buffer	2 μl	Then purify using MegaClear column
sgRNA PCR template	300 ng	Confirm concentration with spectrophotometer—Expect 5-100 μg of total RNA
T7 enzyme	2 μl	
Total	**20 μl**	

genes. Make small aliquots (1–10 μl) of each sgRNAs to avoid excess freeze–thaw cycles of editing components.

4. Synthesizing sgRNA plasmids according to Li et al [15], which are as follows: (1) generate double-stranded DNA templates for the designed sgRNAs; (2) choose a U6 plasmid of interest to be used as backbone (e.g., Addgene # 117209-117212)—for best editing results use either the U6b (AAEL017774) or the U6c (AAEL017763) promoter to drive expression of the gRNA [9]; (3) digest the backbone plasmid using appropriate restriction enzymes; and (4) assemble/ligate the DNA templates of the sgRNAs and the digested plasmid backbone (use Gibson assembly or other methods).
5. Germline encoded Cas9 lines result in the most efficient editing (via EJ repair or HDR) [6]. At the same concentration, recombinant Cas9 protein is more efficient in inducing mutagenesis compared to Cas9 mRNA [5, 6]. To obtain NHEJ mutations, we typically inject into 100 WT embryos, or 25 Exu-Cas9-derived embryos. To obtain a desired HDR insertion, we typically inject 500–1000 WT embryos or 200–400 Exu-Cas9-derived embryos. It is important to note that the Exu-Cas9 and other generated Cas9 lines [6] are marked with Opie2-DsRed. Following editing via NHEJ, this marked Cas9 transgene can easily be sorted away by selecting non-DsRed progeny. For transgenesis/HDR, other markers should be chosen to distinguish the Cas9 transgene from the desired insertion. Importantly, there are several commonly used promoters (Table 7) and fluorescent markers (Table 8) that can be used for this purpose, but care should be taken as some of these promoters have similar expression patterns (e.g., Polyubiquitin and HR5IE1).
6. Hemotek membrane feeding system loaded with fresh animal blood (e.g. human, cattle, chicken) provides an alternative option for mosquito blood-feeding [33]. Based on our experience however, blood-feeding with live anesthetized mice

Table 7
Commonly used promoters used to express fluorescent proteins as robust and reliable genetic markers for transformation in *Ae. aegypti*

Promoter	Expression pattern	Stage of visible expression	An example citation of promoter being used in *Ae. aegypti*
Ubiquitin-L40 (AAEL006511)	Robust ubiquitous expression including the germline	Larva/pupa/adult	[6]
Polyubiquitin (AAEL003877)	Robust ubiquitous expression excluding the germline	Larva/pupa/adult	[6]
3xP3	Eyes/neural	Larva/pupa/adult	[6, 8]
HR5IE1	Robust ubiquitous expression excluding the germline	Larva/pupa/adult	[16, 30]
OPIE2	Robust ubiquitous expression excluding the germline	Larva/pupa/adult	[6, 8]

Table 8
Fluorescent markers commonly used in *Ae. aegypi* for transgenic applications

Fluorescent marker	Excitation max (nm)/ emission max (nm)	Leica single-filter set for m165FC	Leica multi-filter set for m165FC	An example citation of marker being used in Ae. aegypti
ECFP	439/476	Leica part #10447409; excitation ET436/20x, emission ET480/40 m wavelengths / band-pass	Leica part #10450611; ET434.5/21, 501.5/19, 574.5/23, ET469.5/25, 536.5/29, 635.5/69 wavelengths	[8]
EGFP	484 /507	Leica part #10447408; excitation ET470/40x, emission ET525/50 m wavelengths / band-pass	Leica part #10450611; ET434.5/21, 501.5/19, 574.5/23, ET469.5/25, 536.5/29, 635.5/69 wavelengths	[8]
EYFP	514/527	Leica part #10447410; excitation ET500/20x, emission ET535/30 m wavelengths/band-pass	N/A	[31]
RFP / DsRed/ tdTomato / mCherry	558/583	Leica part #10450195; excitation ET560/40×, emission ET630/75 m wavelengths/band-pass	Leica part #10450611; ET434.5/21, 501.5/19, 574.5/23, ET469.5/25, 536.5/29, 635.5/69 wavelengths	[8, 32]

Table 9
Microinjection needle preparation protocol for different glass types

Capillary glass needle type	Sutter needle puller model	Heat	Filament	Velocity	Delay	Pull	Pressure
Quartz	P-2000	750	4	40	150	165	–
Aluminosilicate	P-1000	605	–	130	80	70	500
Borosilicate	P-1000	450	–	130	80	70	500

results in better egg lays and survival rates of injected eggs. Mice are anesthetized intraperitoneally using Ketamine/Xylazine/sterile saline solution (82.5 mg/kg/10 mg/kg) for mosquito blood-feeding abiding institutionally approved IACUC protocols to ensure proper animal handling and well-being.

7. We do not recommend collecting embryos from adult females older than 10 days or from gravid females more than 7 days post-blood meal. We also do not recommend collecting eggs from females during their second or later gonotrophic cycles for microinjection. Based on our experience, the quality and hatching rates of embryos from these females are reduced.
8. For effective microinjection into *Ae. aegypti* embryos, we experimented with several types of capillary glass needles with filaments including quartz, aluminosilicate, and borosilicate types. The quality of needles is critical for avoiding breakage/clogging during injection and to enhance embryo survival and CRISPR editing efficiency. For each of these glass types, we developed optimal protocols for generating needles on different Sutter micropipette puller models (P-1000 and P-2000) to enable the needles to have a desired hypodermic-like long tip, which is most effective for embryo microinjection. The parameters (filament, velocity, delay, pull, pressure) for the different types of capillary glass needles are listed in Table 9. While all three types of needles were effective, we prefer Quartz capillary glass needles, because the aluminosilicate and borosilicate capillary glass needles were a bit too soft and break and clog easily. Prior to loading the injection needle with injection mixture, the tip of the needle will need to be beveled using a Sutter needle beveler to generate a sharp open point (Fig. 10).
9. Due to the inherent variability in the injection procedure, three biological replicates, each replicate having injections of 50 embryos from three separate batches of mothers on separate days, are usually required. We usually do not desiccate the embryos prior to injection. However, if the embryo cytoplasm is observed to be leaking/oozing out of the embryos following needle penetration, then a light desiccation may be required.

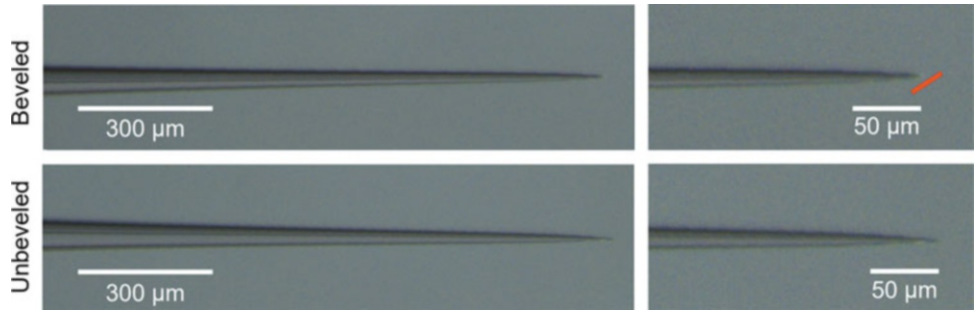


Fig. 10 Tips of a beveled and unbeveled quartz capillary glass needles used for microinjection. These glass needles are pulled with a laser needle puller and then gently opened and refined using a beveler. A good beveled needle has a very sharp tip (red line indicates the optimal angle of the tip aperture following beveling)

To lightly desiccate embryos for microinjection, place the slide containing embryos briefly under the illumination of a dissecting microscope and constantly monitor the embryos under high magnification until the slightest sign of a dimple or indentation is observed in the side of the embryos. Alternatively, the slide containing aligned embryos can be placed in a dehydration chamber consisting of a sealed Nalgene jar with desiccant (e.g., Drierite); the time in the jar required for optimal desiccation for microinjection needs to be adjusted according to the user's laboratory conditions (i.e., background temperature and humidity).

10. To search for potential off-target effects in the *Ae. aegypti* genome, we use the CHOP-CHOP web tool. If this tool predicts off-target sites of concern (i.e., high degree of homology), then the following method can be used to screen and select against these mutations. Design pairs of primers for each putative off-target site to cover the entire off-target site sequence. Then extract the genomic DNA from generated transgenic mosquitoes and PCR amplify the putative off-target site sequences. Purify the amplified PCR product and send for Sanger sequencing. Compare the sequenced results with the reference *Ae. aegypti* genome. If off-target effects are detected on a separate chromosome, back cross the transgenic mosquitoes with a wild-type strain for at least three generations (or until the off-target effects are removed). Use the above method to check for off-target site sequence at every generation and select against the off-target edited individuals while selecting for the desired on-target edited individuals. Alternatively, to exclude the possibility of closely linked off-target effects, generate a separate allele using another unique sgRNA.
11. For synthesizing a donor DNA plasmid from several DNA fragments, we use the Gibson assembly method. The Gibson

Table 10
Gibson assembly protocol for donor DNA plasmid synthesis [34]

Step	Conditions and reagents
1. Set up reactions on ice	Each reaction (20 μl) contains the following reagents 1. DNA fragments to be assembled of the following concentration (total volume *x* μl): 0.02–0.5 pmols (when assembling 2–3 fragments) 0.2–1 pmols (when assembling 4–6 fragments) 2. Gibson assembly master mix 2×, 10 μl 3. dH_2O (10 − *x* μl)
2. Incubation in thermocycler	50 °C 60 min (could be shortened to 15 min if only 2–3 fragments are being assembled)
3. Storage	−20 °C, until use

assembly protocol can be found in Table 10. Three types of reagents are needed: DNA fragments (0.02–0.5 pmols when assembling 2–3 fragments, and 0.2–1 pmols when assembling 4–6 fragments), Gibson assembly master mix, and dH_2O.

12. If future removal of the chosen fluorescent marker from mosquitoes is conceivably desirable for your application, add two LoxP sites in the same orientation to flank the marker transgene to facilitate Cre-LoxP-mediated excision after line establishment [17].

13. T2A is a viral sequence that causes ribosome skipping and allows two separate proteins to be expressed from the same mRNA transcript [35, 36]. T2A-driver cassettes are commonly inserted in-frame at one of two different points in a target ORF, yielding non-functional or wild-type alleles of the target gene depending on the region targeted for insertion: (1) If the T2A-driver is inserted in-frame into a coding exon early in the ORF (e.g., in the first one to two thirds of the ORF), the targeted gene may give rise to a truncated and potentially non-functional target protein, as well as a full-length functional driver protein [17, 22]. If the T2A-driver cassette is inserted into the last part of the ORF just before the stop codon, the targeted gene will likely yield a full-length functional target protein and driver protein [19, 37, 38]. Such gene-specific T2A-driver lines can then be crossed with various responder lines for cell-type-specific control of reporter gene expression for anatomical and functional studies in *Ae. aegypti* [17, 19, 22, 37, 38]. As most *Ae. aegypti* genes are haplosufficient, T2A-drivers inserted early in target gene ORF can normally be used in the heterozygous state while still maintaining wild-type functions of the targeted gene, as there is another wild-type and undisrupted allele of the gene to compensate [17, 22].

14. Base donor plasmids are publicly available for insertion of HDR arms to generate user-specified targeting vectors for gene knockout and cell-type-specific labeling in *Ae. aegypti*. These include plasmid pSL1180-HR-PUbECFP and plasmid PSL1180polyUBdsRED which can be used for the integration of a constitutive fluorescent cassette into the CRISPR-Cas9 cut site for gene knockout. The base plasmid (pBB) can further be used to clone target-specific homology arms for the generation of T2A-QF2 in-frame fusions in *Ae. aegypti* [17].
15. When using 3xP3-DsRed as a marker transgene in the donor HDR plasmid, transient somatic expression of DsRed can be observed in some of the hatching G0 larvae. We use this to confirm plasmid was indeed injected. Moreover, if significant embryo lethality is observed following microinjection, but target lethality is unknown, injection success can be verified by observation of "mosaic" episomal fluorescence from the injected transgene in the posterior end of young G0 larvae.

Acknowledgments

We thank Andie Smidler for providing feedback on this protocol. This work was supported by funding from a DARPA Safe Genes Program Grant (HR0011-17-2-0047) and a NIH award (R01AI151004) awarded to O.S.A, and NIH award (R21AI14645001) to C.J.M. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the U.S. government.

Ethical Conduct of Research All animals were handled in accordance with the Guide for the Care and Use of Laboratory Animals as recommended by the National Institutes of Health and approved by the UCSD Institutional Animal Care and Use Committee (IACUC, Animal Use Protocol #S17187) and UCSD Biological Use Authorization (BUA #R2401).

Disclosures O.S.A. is a founder of Agragene, Inc. The terms of this arrangement have been reviewed and approved by the University of California, San Diego, in accordance with its conflict of interest policies. All other authors declare no competing interests.

References

1. Jinek M et al (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821
2. Bui M, Li M, Raban RR, Liu N, Akbari OS (2020) Embryo Microinjection Techniques for Efficient Site-Specific Mutagenesis in Culex quinquefasciatus. J Vis Exp. https://doi.org/10.3791/61375
3. Li M et al (2020) Methods for the generation of heritable germline mutations in the disease vector Culex quinquefasciatus using clustered regularly interspaced short palindrome repeats-associated protein 9. Insect Mol Biol 29:214–220
4. Li M, Akbari OS, White BJ (2018) Highly efficient site-specific mutagenesis in malaria mosquitoes using CRISPR. G3 (Bethesda) 8:653–658
5. Kistler KE, Vosshall LB, Matthews BJ (2015) Genome engineering with CRISPR-Cas9 in the mosquito Aedes aegypti. Cell Rep 11:51–60
6. Li M et al (2017) Germline Cas9 expression yields highly efficient genome engineering in a major worldwide disease vector, Aedes aegypti. Proc Natl Acad Sci U S A 114:E10540–E10549
7. Powell JR (2018) Mosquito-borne human viral diseases: why Aedes aegypti? Am J Trop Med Hyg 98:1563–1565
8. Li M, Yang T, Bui M et al (2021) Suppressing mosquito populations with precision guided sterile males. Nat Commun 12:5374. https://doi.org/10.1038/s41467-021-25421-w
9. Li M et al (2020) Development of a confinable gene drive system in the human disease vector Aedes aegypti. eLife 9:e51701
10. Chaverra-Rodriguez D et al (2018) Targeted delivery of CRISPR-Cas9 ribonucleoprotein into arthropod ovaries for heritable germline gene editing. Nat Commun 9:3008
11. Gantz VM, Akbari OS (2018) Gene editing technologies and applications for insects. Curr Opin Insect Sci 28:66–72
12. Lin CC, Potter CJ (2016) Editing transgenic DNA components by inducible gene replacement in Drosophila melanogaster. Genetics 203:1613–1628
13. Matthews BJ et al (2018) Improved reference genome of Aedes aegypti informs arbovirus vector control. Nature 563:501–507
14. Li M et al (2017) Generation of heritable germline mutations in the jewel wasp Nasonia vitripennis using CRISPR/Cas9. Sci Rep 7:901
15. Li HH, Li JC, Su MP, Liu KL, Chen CH (2021) Generating mutant Aedes aegypti mosquitoes using the CRISPR/Cas9 system. STAR Protocols 2:100432
16. Zhu GH, Albishi NM, Chen X, Brown RL, Palli SR (2021) Expanding the toolkit for genome editing in a disease vector, Aedes aegypti: transgenic lines expressing Cas9 and single guide RNA induce efficient mutagenesis. CRISPR J 4(6):846–853. https://doi.org/10.1089/crispr.2020.0052
17. Shankar S et al (2021) Synergistic coding of carbon dioxide and a human sweat odorant in the mosquito brain. BioRxiv. https://doi.org/10.1101/2020.11.02.365916
18. Basu S et al (2015) Silencing of end-joining repair for efficient site-specific gene insertion after TALEN/CRISPR mutagenesis in Aedes aegypti. Proc Natl Acad Sci U S A 112:4038–4043
19. Zhao Z, Tian D, McBride CS (2021) Development of a pan-neuronal genetic driver in Aedes aegypti mosquitoes. Cell Rep Methods 1(3):100042. https://doi.org/10.1101/2020.08.22.262527
20. Riabinina O et al (2015) Improved and expanded Q-system reagents for genetic manipulations. Nat Methods 12:219–222. 5 p following 222
21. Gamez S, Vesga LC, Mendez-Sanchez SC, Akbari OS (2021) Spatial control of gene expression in flies using bacterially derived binary transactivation systems. Insect Mol Biol 30(5):461–471. https://doi.org/10.1101/2020.11.24.396325
22. Matthews BJ, Younger MA, Vosshall LB (2019) The ion channel ppk301 controls freshwater egg-laying in the mosquito Aedes aegypti. eLife 8:e43963
23. Akbari OS, Antoshechkin I, Hay BA, Ferree PM (2013) Transcriptome profiling of Nasonia vitripennis testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio. G3 3:1597–1605
24. Ferree PM et al (2015) Identification of Genes Uniquely Expressed in the Germ-Line Tissues of the Jewel Wasp Nasonia vitripennis. G3 5:2647–2653
25. Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E (2016) CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. Nucleic Acids Res 44:W272–W276
26. Xie S, Shen B, Zhang C, Huang X, Zhang Y (2014) sgRNAcas9: a software package for

designing CRISPR sgRNA and evaluating potential off-target cleavage sites. PLoS One 9:e100448

27. Stover NA, Cavalcanti ARO (2014) Using NCBI BLAST. In: Current Protocols Essential Laboratory Techniques, pp 11.1.1–11.1.35
28. Bhagwat M, Young L, Robison RR (2012) Using BLAT to find sequence similarity in closely related genomes. Curr Protoc Bioinformatics. Chapter 10, Unit 10.8
29. Vouillot L, Thélie A, Pollet N (2015) Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. G3 5:407–415
30. Haghighat-Khah RE et al (2019) Engineered action at a distance: blood-meal-inducible paralysis in Aedes aegypti. PLoS Negl Trop Dis 13:e0007579
31. Jové V et al (2020) Sensory discrimination of blood and floral nectar by Aedes aegypti mosquitoes. Neuron 108:1163–1180.e12
32. Buchman A et al (2020) Broad dengue neutralization in mosquitoes expressing an engineered antibody. PLoS Pathog 16:e1008103
33. Gunathilaka N, Ranathunge T, Udayanga L, Abeyewickreme W (2017) Efficacy of blood sources and artificial blood feeding methods in rearing of Aedes aegypti (Diptera: Culicidae) for sterile insect technique and incompatible insect technique approaches in Sri Lanka. Biomed Res Int 2017:1–7
34. New England Biolabs. Gibson Assembly® Protocol (E5510). https://www.neb.com/protocols/2012/12/11/gibson-assembly-protocol-e5510
35. Diao F, White BHA (2012) Novel approach for directing transgene expression in drosophila: T2A-Gal4 in-frame fusion. Genetics 190: 1139–1144
36. González M et al (2011) Generation of stable drosophila cell lines using multicistronic vectors. Sci Rep 1:75
37. Zhao Z et al (2020) Chemical signatures of human odour generate a unique neural code in the brain of Aedes aegypti mosquitoes. bioRxiv. https://doi.org/10.1101/2020.11.01.363861
38. Younger MA et al (2020) Non-canonical odor coding ensures unbreakable mosquito attraction to humans. bioRxiv. https://doi.org/10.1101/2020.11.07.368720

Chapter 3

PIWI-Directed DNA Elimination for Tetrahymena Genetics

Salman Shehzada and Kazufumi Mochizuki

Abstract

Piwi-bound small RNAs induce programmed DNA elimination in the ciliated protozoan *Tetrahymena*. Using the phenomenon called codeletion, this process can be reprogrammed to induce ectopic DNA elimination at basically any given genomic location. Here, we describe the usage of codeletion for genetic studies in *Tetrahymena* and for investigations of the molecular mechanism of Piwi-directed programmed DNA elimination.

Key words PIWI, Small RNA, DNA elimination, Gene knockout, *Tetrahymena*

1 Introduction

Argonaute proteins are a highly conserved family found in species from humans to archaea and comprise the central protein component of the small RNA-mediated regulatory machinery. Argonaute proteins in eukaryotes can be divided into two subfamilies: the Ago and Piwi subfamilies. While the latter is found predominantly in animals, ciliated protozoans also express Piwi subfamily proteins [1–6]. Therefore, ciliates serve as unicellular models to study the molecular actions of Piwi proteins.

Ciliates are a group of eukaryotes characterized by nuclear dimorphism, in which each cell has a polyploid somatic macronucleus (MAC) and a diploid germline micronucleus (MIC). The former controls metabolism, and the latter functions in reproduction. At each sexual cycle in ciliates, the MIC produces new MICs and new MACs, and the parental MAC is degraded. During the development of the new MAC of *Tetrahymena thermophila*, a model ciliate, ~12,000 DNA segments called internal eliminated sequences (IESs) are reproducibly eliminated [7]. IESs are recognized by 26- to 32-nt small RNAs called scnRNAs [8], which are loaded into the Piwi protein Twi1p [9]. Unlike Piwi-loaded small

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_3,

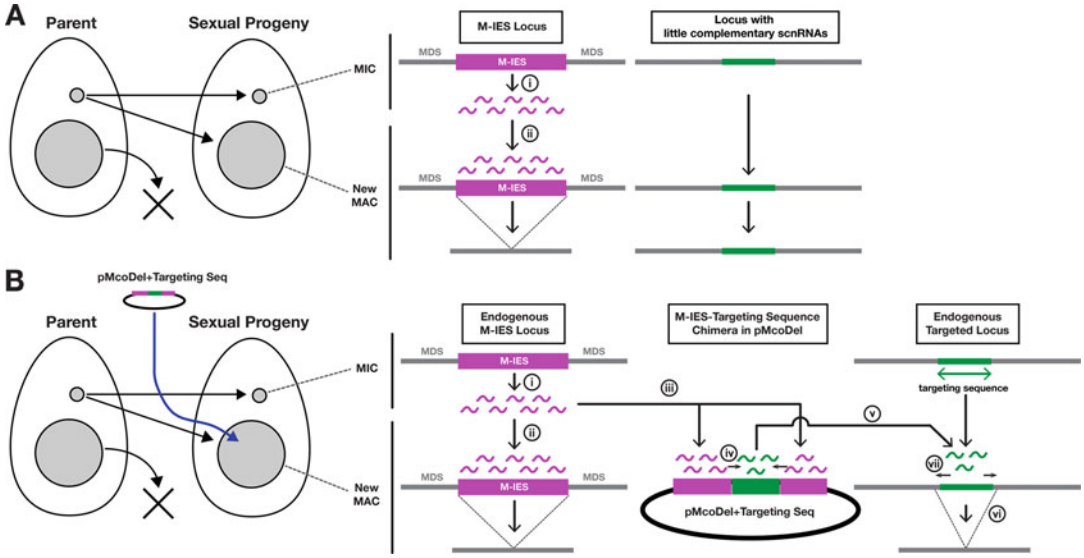


Fig. 1 The proposed actions of scnRNAs in coDel. (**a**) DNA elimination without induction of coDel. Early scnRNAs (magenta wavy lines) produced from Type-A IESs (represented by M-IES in this figure) in the MIC (i) interact with Type-A IESs in the new MAC and induce DNA elimination (ii). Non-IES loci (MAC-destined sequences [MDSs]) either do not produce early scnRNAs or homology-dependent small RNA degradation eliminates early scnRNAs derived from them [8, 12, 20, 26]; thus, DNA elimination is not induced at these loci. (**b**) Induction of coDel. When an IES (magenta box)-targeting sequence (green box) chimeric construct (pMcoDel+Targeting Seq) is introduced into the new MAC, early scnRNAs produced from M-IES in the MIC (i) interact not only with the endogenous M-IES (ii) but also with the M-IES on the introduced chimeric construct (iii). The latter interaction triggers late scnRNA (green wavy lines) production from the adjacent target sequence in cis (iv), which then interacts with the endogenous target locus in trans (v) and induces ectopic DNA elimination (vi). Late scnRNAs likely further induce the production of late scnRNAs in cis at the targeted locus (vii), making deletions longer than the targeted sequence

RNAs in animals, scnRNAs are processed from long double-stranded RNA precursors by a Dicer homolog [10, 11].

IESs in *Tetrahymena* can be classified into two types: Type A IESs, which are recognized in cis by the primary scnRNAs (early scnRNAs) that they produce, and Type B IESs, which are recognized in trans by early scnRNAs produced by Type A IESs [12]. Because secondary scnRNA (late scnRNA) biogenesis occurs concomitantly with the spreading of heterochromatin on IESs [12], the introduction of a chimeric construct of a Type A IES and a fragment of a non-IES locus (targeting sequence) into the new MAC results in the production of secondary scnRNAs from the targeting sequence that induce ectopic DNA elimination at the non-IES locus in trans (Fig. 1). This phenomenon is called codeletion (coDel).

Because coDel can be induced by the simple introduction of an IES-targeting DNA chimeric construct, this approach is suitable as a gene knockout technology in *Tetrahymena* [13–19]. Moreover,

the effects of an array of different targeting sequences can be easily examined [13], and coDel can also be used to analyze how scnRNAs are involved in the transgenerational inheritance of the pattern of DNA elimination [20] and the base-pairing requirement between small RNAs and genomic DNA to induce the downstream effect. In this chapter, we describe the usage of coDel to produce gene knockout strains in *Tetrahymena* and to investigate the length of the seed sequence of scnRNAs in DNA elimination.

2 Materials

2.1 Extraction of Tetrahymena Genomic DNA

1. *Tetrahymena thermophila* strain B2086 (or other wild-type strain of *Tetrahymena thermophila*) (*see* **Note 1**).
2. 1× Super Proteose Peptone (SPP) medium: 2% Bacto Proteose peptone (Bectone Dickinson and Company), 0.4% D-(+)-Glucose (Sigma), 0.2% Bacto Yeast extract (Bectone Dickinson and Company), 0.003% EDTA iron(III) sodium salt (Sigma) (*see* **Note 2**).
3. CHAOS buffer (stored at 4 °C): 4 M guanidine thiocyanate, 0.5% sodium lauroyl sarcosinate, 25 mM Tris (pH 8), 0.1 M 2-mercaptoethanol.
4. Phenol extraction buffer (stored at RT): 100 mM Tris (pH 8), 10 mM EDTA (pH 8), and 1% SDS.
5. 25:24:1 (v/v/v) Phenol/chloroform/isoamyl alcohol, saturated with 10 mM Tris (pH 8) and 1 mM EDTA.
6. 2-Propanol.
7. 70% ethanol.
8. Ethanol.
9. TE (Tris–EDTA) buffer, pH 8.
10. 10 mg/mL RNase A.

2.2 Construction of coDel-Targeting Plasmid

1. pMcoDel plasmid (*see* **Note 1**).
2. FW and RV primers amplifying the ~500–1000 bp region of a targeted locus (*see* **Note 3**). To assemble a construct using the NEBuilder HiFi DNA Assembly system, the primers must have >16 nt overlapping sequence with pMcoDel arms (*see* **Note 4**). For example, to induce deletion at the *TTHERM_00113330* locus (Fig. 2a), we used EMA2-coDel-FW and EMA2-coDel-RV (Table 1). To assemble a construct using T4 DNA ligase, both primers must have the NotI recognition site (*see* **Note 5**). For example, to induce deletion at a noncoding sequence on MIC chromosome 4 with the 661 bp targeting sequence (Fig. 3a), we used C121_FW1 and C121_RV1 (Table 1).
3. Taq DNA Polymerase with Standard Taq Buffer.

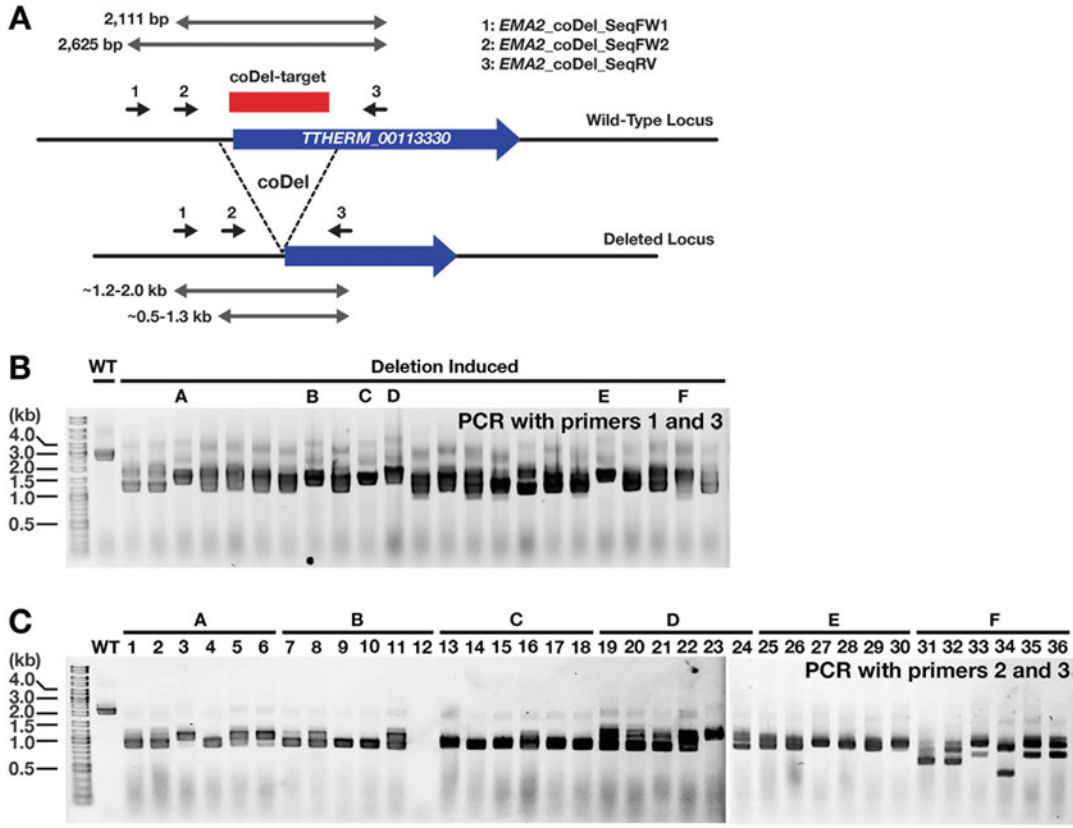


Fig. 2 Production of gene knockout strains using coDel. (**a**) A coDel was induced at the coding sequence of *TTHERM_00113330* (blue boxed arrow) in the MAC by inserting an ~1 kb sequence at the beginning of the *TTHERM_00113330* coding sequence (red box) into the pMcoDel vector. (**b**) Wild-type (WT) and 23 paromomycin-resistant transformants after introduction of the pMcoDel-*TTHERM_00113330* targeting construct (deletion induced) were examined by direct cell PCR using primers #1 and #3 shown in (**a**). The sequences of the primers are listed in Table 1. A 2625 bp product was expected to be obtained without any deletion at the targeted locus (as in WT), and all of the coDel-induced cell lines showed ~1.2–2 kb PCR products, which correspond to the locus with a deletion at the targeted region. (**c**) Six coDel lines (marked with A to F in (**b**)) were selected, six clonal cell lines each were established, and their *TTHERM_00113330* locus was analyzed by PCR using primers #2 and #3 shown in (**a**). The sequences of the primers are listed in Table 1. A 2111 bp product was expected to be obtained without any deletion at the targeted locus (as in WT), and all of the coDel-induced cell lines showed ~0.5–1.2 kb PCR products

4. 2.5 mM each dNTP mix.
5. Column-based Gel and PCR DNA Clean-up Kit.
6. *NotI* restriction enzyme.
7. NEBuilder HiFi DNA Assembly Master Mix (NEB).
8. FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific).
9. DNA Ligation Kit, Mighty Mix (Takara).

Table 1
Sequences of synthetic DNA

Primer name	Sequence (5′--3′)
EMA2_coDel_FW	ATTGTTATCATCTTATGACCGCTTTAATGTATTTGACTAGTGC
EMA2_coDel_RV	TCAAGTTGTAATGCTAAAATAAGCAAAATGAGTTCACTATTTGG
EMA2_coDel_SeqFW1	AATTGGGTACCGGGCCCCCCCTCGAGGTGGATCTGCTGGAAAAAGA TTGTATG
EMA2_coDel_SeqFW2	GCTGGCGGAGATATGCAAGATGGTC
EMA2_coDel_SeqRV	GACTAAGTCCTCTATTGGGTTTGTTC
McoDel_seqFW	ATTGAATAAGGAGACCAGCCTCTC
McoDel_seqRV	AAAACTAATAACTTGGGTCTTTAGATCAC
C121_coDel_cFW	GGTTGATGGTAGGTAGGTACCTC
C121_coDel_cRV	CAGTAGATTTCTTTGAAAGCTAACTC
C121_FW1	GCATGCGGCCGCGAGACATTTCTGGTGTGTTCGG
C121_FW2	GCATGCGGCCGCGAGAGTTAATCTTTATTAAAAGGC
C121_FW3	GCATGCGGCCGCGTATTTTAGTAACATGTATCAAG
C121_FW4	CAAAGTTGATGTTTCTGGTGGTTACAGTTGATAAAAAATTAG
C121_FW5	GCATGCGGCCGCGAGACATTTCTGGTGTGTTGG
C121_FW6	TAAGTACGTCATAATATATTGTTTTCTTGA
C121_FW7	TGACTAGTTTGTAATTATTACTAAATTTAA
C121_FW8	GCATGCGGCCGCGAGACATTTGTGGTGTGTTGG
C121_FW9	TAAGTACGTCTTAATATATTGTTTTCTTGA
C121_FW10	TGACTAGTTTGTAATTTTTACTAAATTTAA
C121_RV1	GCATGCGGCCGCGATTAATTGGCTATTCAAGCTATG
C121_RV2	GCATGCGGCCGCTTGGTGAGGAACCAAATATGTTTG
C121_RV3	GCATGCGGCCGCCTTCATGGGGAGTTAATAATTATC
C121_RV4	GCATGCGGCCGCCCCAACAATTAGATATAATTTCCC
C121_RV5	CTAATTTTTTATCAACTGTAACCACCAGAAACATCAACTTTG
C121_RV6	GCATGCGGCCGCCACCAGAAACATCAACTTTG
C121_RV7	TCAAGAAAACAATATATTATGACGTACTTA
C121_RV8	TTAAATTTAGTAATAATTACAAACTAGTCA
C121_RV9	GCATGCGGCCGCTAGGTGAGGAACCAAATATGTATG
C121_RV10	TCAAGAAAACAATATATTAAGACGTACTTA
C121_RV11	TTAAATTTAGTAAAAATTACAAACTAGTCA
C121_RV12	GCATGCGGCCGCTAGGTGAGGAAGCAAATATGTATG

(continued)

Table 1
(continued)

Primer name	Sequence (5′--3′)
C121_MUT20_1	GAGACATTTCTGGTGTGTTGGGAATTCTATTAAAACTTTCTATGTAA TCTAACTTTTATATTTTTATAAAAATTTAAAATAAAAACAAAG TTGATGTTTGTGGTGATAATTCATAAATATAAGTACGTCATAAT
C121_MUT20_2	ATATTGTTTTCTTGAATATAAGTATAAACTAATTAAGTTAAGTAAC TAAATAAGGTGAAGGATTTCGGAAATTATATCTAATTGTAGGG TAAATTATAATTTTCTACTCTTTTTTATCTTGACTAGTTTGTAAT
C121_MUT20_3	TATTACTAAATTTAATTCAATTTATTGAGAGTGTTAATCTTTA TTAAAAGGGTCAGAAAATAATAGATAATAATTAACTCCCCA TGAAGATTAGATAGAAACAAACAAACATACATATTTGGTTCC TCACCTA
C121_MUT10_1	GAGACATTTGTGGTGTGTTGGGAATTCTAATAAAACTTTCTATGTAA TCAAACTTTTATATTTTTATAATAATTTAAAATAAAAACAAAC TTGATGTTTGTGGTGATAAATCATAAATATAAGTACGTCTTAAT
C121_MUT10_2	ATATTGTTTTCTTGATTATAAGTATAAACTAATTATGTTAAGTAAC TAAATAAGGAGAAGGATTTCGGAAATTATTTCTAATTGTAGGG TAAATTTTAATTTTCTACTCTTTTTTTTCTTGACTAGTTTGTAAT
C121_MUT10_3	TTTTACTAAATTTAATTCAATATATTGAGAGTGTTAATCTTAA TTAAAAGGGTCAGAAAATTATAGATAATAATTAACTCCGCA TGAAGATTAGATAGAAAGAAACAAACATACATATTTGCTTCC TCACCTA

Underlined and double-underlined sequences are for NotI digestion and NEBuilder HiFi assembly to the arms of pMcoDel, respectively

10. NEB 5-alpha Competent *E. coli*: thaw on ice, aliquot 12.5 μL each in screw-capped 1.5-mL tubes, freeze with liquid nitrogen, and store at −80 °C.
11. Primers for colony PCR and sequencing (McoDel_seqFW and McoDel_seqRV [Table 1]).

2.3 Transformation for Inducing coDel

1. Two wild-type *Tetrahymena thermophila* strains having complementary mating types. We typically use B2086 (Mating Type II) and CU428 (Mating Type IIV) (*see* **Note 1**).
2. Biolistic PDS-1000/He Particle Delivery System (Bio-Rad) (*see* **Notes 6** and **7**).
3. 10 mM Tris, pH 7.5.
4. 1× and 2× Super Proteose Peptone (SPP) medium (*see* **Note 2**).
5. 100 mg/mL paromomycin sulfate.
6. 100× Antibiotic/Antimycotic (AB/AM) mix (Millipore).
7. 96-well cell culture plates.
8. 30 °C Incubator.

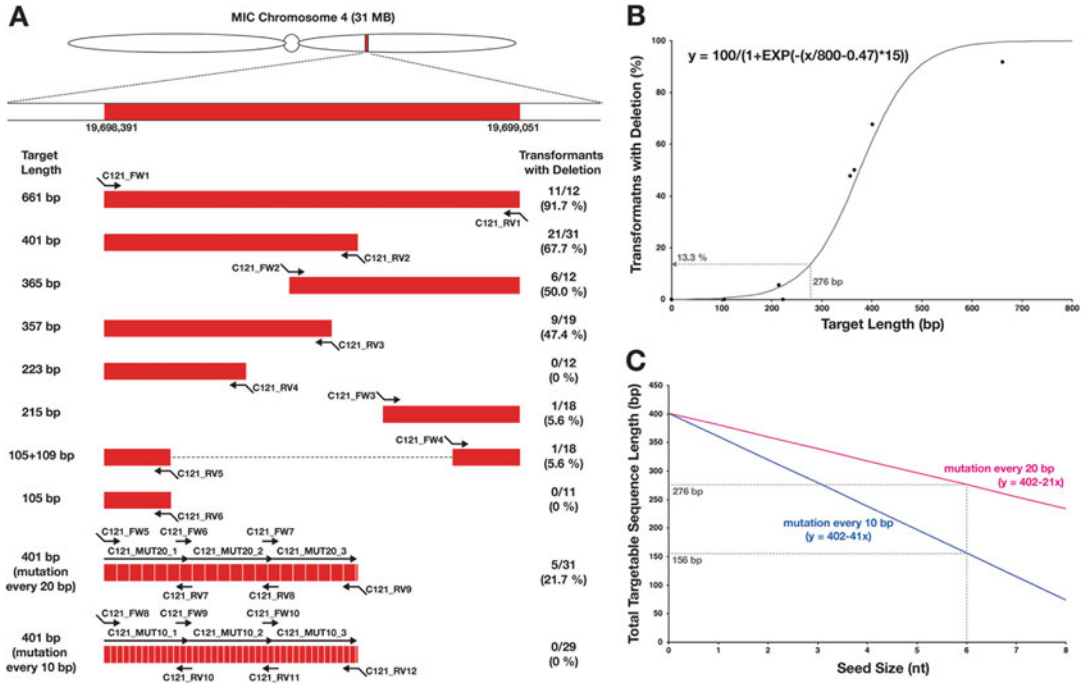


Fig. 3 Examination of the requirement of scnRNA-target interaction in DNA elimination using coDel. (**a**) Various targeting sequences that are designed to target a non-IES region at the indicated position of MIC chromosome 4 (red boxes) were introduced into the new MAC of progeny of B2086 and CU428, and the number of transformants that showed deletion at the targeted locus was counted. Arrows indicate the primers and oligo DNAs (their sequences are listed in Table 1) that were used to amplify and construct the targeting sequences. (**b**) Length of the different targeting sequences shown in (**a**) (except the last two having base replacements) and fraction of transformants that showed deletion by those targeting sequences are plotted. A sigmoidal curve that is close to the plotted experimental data is shown. Dotted lines indicate the expected fraction of deletion-positive transformants by a 276 bp targeting sequence, which corresponds to the total targetable sequence lengths if scnRNAs from the 401 bp targeting sequence having base replacement every 20 nt use 6 nt as seed (see below). (**c**) Total targetable sequence lengths within the 401 bp targeted sequence by the scnRNAs derived from the 401 bp targeting sequence that have base replacement in every 20 (magenta line) or 10 (blue line) bp were predicted (*y*-axis) for different seed sizes to initiate recognition of the targeted sequences by scnRNAs (*x*-axis). Dotted lines indicate length of predicted total targetable sequence when scnRNA uses 6 nt as seed. *See* also **Note 20** for interpretation of the results

2.4 Detecting coDel by Direct PCR

1. Primers amplifying the targeted region and its ~1 kb flanking sequences. To examine the *TTHERM_00113330* locus, we used either EMA2_coDel_SeqFW1 and EMA2_coDel_SeqRV (primers #1 and #3 in Fig. 2a) or EMA2_coDel_SeqFW2 and EMA2_coDel_SeqRV (primers #2 and #3 in Fig. 2a).
2. GoTaq Long PCR Master Mix (Promega).

2.5 Establishing Gene Knockout Cell Lines from Cells with coDel

1. DropMaker (made-to-order, *see* [21] for more explanations).
2. Bacteriological petri dish.
3. Glass capillary.
4. Aspirator tube assemblies for microcapillary pipettes.
5. Glass capillary puller.
6. Dissection Microscope.
7. 24-well and 96-well cell culture plates.
8. 13.3% DMSO in 10 mM Tris, pH 7.5 (*see* **Note 8**).
9. Cryo 1 °C "Mr. Frosty" Freezing Container (Nalgene).

2.6 Using coDel to Study DNA Elimination Mechanism

1. Primers to amplify different regions of target sequences (Fig. 3a) (Table 1).
2. Synthetic double-stranded DNA with base replacements at defined intervals (Integrated DNA Technologies) (*see* **Note 9**).
3. Synthetic single-stranded DNA corresponding to a part of the target sequence and having base replacements at defined intervals (such as Ultramer DNA oligos from Integrated DNA Technologies) (Table 1).
4. Primers amplifying the targeted region and its ~1 kb flanking sequences (Table 1).

3 Methods

3.1 Extraction of Tetrahymena Genomic DNA

1. Culture the B2086 strain of *Tetrahymena* cells in 20 mL 1× SPP at 30 °C overnight (*see* **Note 10**).
2. Place approximately 1 mL of the culture (~1×10^6 cells) in a 1.5-mL Eppendorf tube, centrifuge for ~5 s in a mini table top centrifuge, and quickly remove the supernatant using a pipette.
3. Add 300 μL CHAOS buffer and lyse cells by pipetting (*see* **Note 11**).
4. Add 300 μL of Phenol Extraction Buffer and mix vigorously by inverting the tube (*see* **Note 12**).
5. Add 600 μL of phenol/chloroform/isoamyl alcohol and mix vigorously by inverting the tube.
6. Spin down for 10 min at 14,000 rpm (or maximum speed) at RT and transfer 600 μL of the upper phase to a fresh 1.5-mL tube.
7. Add 600 μL of 2-propanol and mix well by inverting the tube.
8. Spin down for 10 min at 14,000 rpm at 4 °C and discard supernatant.

9. Add 700 μL ice-cold 70% ethanol, leave for 5 min at RT and spin down for 5 min at 14,000 rpm at 4 °C.
10. Discard the supernatant as much as possible and air dry for 5–10 min.
11. Add 200 μL TE buffer and incubate at RT for ~15 min (vortex occasionally) (*see* **Note 13**).
12. Add 1 μL of RNaseA and incubate at 37 °C for 30 min.
13. Add 200 μL TE buffer and 400 μL phenol/chloroform/isoamyl alcohol, mix by inverting tube several times, spin down for 10 min at 14,000 rpm at RT.
14. Transfer the upper phase to a fresh 1.5-mL tube, add 5% vol. of 5 M NaCl, and mix well.
15. Add 2× vol. of ethanol, mix well and incubate the tube at −20 °C for more than 1 h.
16. Spin down for 10 min at 14,000 rpm at 4 °C and discard the supernatant.
17. Add 700 μL ice-cold 70% ethanol and spin down for 5 min at 14,000 rpm at 4 °C.
18. Remove supernatant as much as possible and air dry for 5–10 min.
19. Add 100 μL TE buffer and incubate for ~15 min at RT (vortex occasionally).
20. Measure the DNA concentration, adjust to 0.5 μg/μL by adding TE and store at −20 °C.

3.2 Construction of coDel-Targeting Plasmid

1. Set up a 50 μL PCR containing 1× PCR buffer, 100 μM each dNTP, 0.2 μM FW primer, 0.2 μM RV primer, 10 ng Tetrahymena genomic DNA (Subheading 3.1, **step 20**), and 1.25 Unit Taq polymerase.
2. Perform PCR with the following conditions: 95 °C for 2 min, 35 cycles of 95 °C for 20 s/55 °C for 30 s/68 °C for 60 s, 68 °C for 3 min.
3. Check the product by running 3 μL of the reaction in a 1% agarose gel.
4. Purify the PCR product using NucleoSpin Gel and PCR Clean-up Kit (or similar glass filter base purification system). If necessary, run the PCR product in a 1% agarose gel and excise the correct PCR product.
5. Digest ~1–2 μg of pMcoDel plasmid DNA with NotI-HF and 1×CutSmart buffer in 50 μL overnight. If the NEBuilder HiFi DNA Assembly system will be used, go to **step 6**. When T4 DNA ligase will be used, go to **step 9**.

6. Extract NotI-digested pMcoDel with NucleoSpin Gel and PCR Clean-up Kit and elute with 30 μL elution buffer.
7. Mix 1 μL of NotI-digested pMcoDel (**step 6**), 1 μL PCR-amplified targeting sequence (**step 4**) and 2 μL of 2× NEBuilder HiFi DNA Assembly Mix.
8. Incubate at 50 °C for 30 min and go to **step 14**.
9. Add 1 μL of FastAP, incubate at 37 °C for 1 h and then inactivate the enzyme for 15 min at 65 °C.
10. Extract DNA with a DNA Clean-up Kit and elute with 30 μL elution buffer.
11. Digest the PCR product (**step 4**) with NotI-HF as in **step 5**; follow by purification using a DNA Clean-up Kit.
12. Mix 1 μL of NotI-digested and dephosphorylated pMcoDel (**step 10**), 1 μL NotI-digested PCR product (**step 10**), and 2 μL of 2× ligation mix.
13. Incubate at 16 °C for a few hours to overnight.
14. Transform NEB5alpha competent cell (12. 5 μL) with a 0.5 μL reaction (either from **steps 8** or **13**) and inoculate on an LB-amp plate.
15. Check the resulting plasmid construct by colony PCR using McoDel_seqFW and McoDel_seqRV.
16. Purify the plasmid construct with the correct insert size by mini-prep, and check the sequence of the insert by Sanger sequencing using McoDel_seqFW.

3.3 Transformation for Inducing coDel

1. Mix 25 mL each 5×10^5/mL prestarved B2086 and CU428 in 10 mM Tris pH 7.5 (*see* **Note 10**).
2. Introduce 5 μg of pMcoDel-Targeting Sequence plasmid into 6×10^6 cells at 7 hpm using BioRad Biolistic Particle Gun system (*see* **Note 6**).
3. Release cells into 50 mL 10 mM Tris–HCl (pH 7.5) and incubate for ~12 h at 30 °C with slow (60 rpm) rotation.
4. Add 50 mL of 2×SPP and 500 μL of 100× AB/AM mix.
5. Incubate at 30 °C for an additional ~4 h with slow (90 rpm) rotation.
6. Add 100 μL of 100 mg/mL paromomycin (final conc. 100 μg/mL).
7. Make two diluted cultures: (a) 1.5 mL of the culture + 13.5 mL of 1× SPP + 135 μL 100× AB/AM mix + 13.5 μL of 100 mg/mL paromomycin; (b) 0.3 mL of the culture + 14.7 mL of 1× SPP + 100× AB/AM mix + 14.7 μL of 100 mg/mL paromomycin.
8. Aliquot cells into 96-well plates.

Original culture: 150 μL/well, 7× 96-well plates.

1/10 diluted culture (**7a**): 150 μL/well, 1× 96-well plates.

1/50 diluted culture (**7b**): 150 μL/well, 1× 96-well plates.

9. Place the plates in a moist chamber (plastic box with wet tissue papers) and incubate for 3–4 days at 30 °C (*see* **Note 14**).
10. Select 24 wells containing paromomycin resistance cells from 96-well plate(s) in which 2/3 or less of the wells had growing cells (*see* **Note 15**).

3.4 Detecting coDel by Direct PCR

1. Mix the following in a PCR tube and keep on PCR cooler (or on ice): 10 μL GoTaq Long PCR Master Mix (2×), 0.5 μL 10 μL FW primer, 0.5 μL 10 μL RV primer, 8.5 μL water.
2. Add 0.5 μL culture (Subheading 3.2, **step 8**) to the PCR tube above (Subheading 3.3, **step 1**).
3. Perform PCR under the following conditions: 94 °C for 2 min; 35 cycles of 94 °C for 20 s, 50–55 °C for 30 s, and 67 °C for 1 min per kb; then 67 °C for 5 min.
4. Analyze 3–5 μL of reaction by agarose gel electrophoresis (Fig. 2b as an example).

3.5 Establishing Gene Knockout Cell Lines from Cells with coDel

1. Select 6–8 cell lines in which deleted copies of the target locus are dominant over the wild-type copy in Subheading 3.4, **step 4** (*see* **Note 16**).
2. Prepare 48 drops of 1× SPP on a 10 cm culture dish (1 plate for 2 lines) using a DropMaker. Alternatively, drops (~50 μL each) can be made using a pipet.
3. Pick up 24 single cells each from the corresponding culture (Subheading 3.2, **step 10**) using a microcapillary pipette under a dissection microscope and separately place into the drops above.
4. Incubate for 3 days at 30 °C in a moist chamber.
5. Choose 6–8 drops containing well-growing cells per line and transfer them individually into 1 mL fresh 1× SPP in a 24-well plate.
6. Incubate for 2 days at 30 °C.
7. Perform direct PCR as in Subheading 3.4 to find clones showing only deleted copies of the target locus (Fig. 2c as an example) (*see* **Notes 17** and **18**).
8. (If sexually mature cells are needed for the experiment) Every 2 days, inoculate 5 μL of the selected clones into 1 mL fresh 1× SPP and incubate at 30 °C (or keep cells at room temperature for a week) until passage 10. Then, perform test mating to identify combinations of clones with complementary mating types.

9. Prepare frozen cell stocks using DMSO and Cryo 1 °C Freezing Container. Refer to [21] for detailed cell storage protocol.

3.6 Using coDel to Study DNA Elimination Mechanism

1. Design modified targeting sequences (Fig. 3a as examples) (*see* **Note 19**) and prepare them by PCR from the original target or by gene synthesis.
2. Introduce the modified target sequences into pMcoDel as in Subheading 3.2, **steps 4–16**.
3. Introduce the resulting constructs into mating Tetrahymena cells as described in Subheading 3.3.
4. Analyze the target locus by direct PCR as in Subheading 3.4.
5. Measure the deletion frequency by counting the number of cell lines showing PCR products shorter than that from the wild-type target locus (Fig. 3a) (*see* **Note 20**).

4 Notes

1. The wild-type *Tetrahymena thermophila* strains and pMcoDel plasmid are available from the *Tetrahymena* Stock Center <http://tetrahymena.vet.cornell.edu/>.
2. 1× and 2× SPP is prepared from 10× SPP stock, autoclaved for 20 min at 120 °C, and stored at 4 °C (for <10 months) or at RT (for <3 months). 10× SPP is prepared by dissolving 600 g of Bacto Proteose peptone, 120 g of D-(+)-Glucose, 60 g of Bacto yeast extract, and 0.9 g of EDTA iron (III) sodium salt into 3 L (final volume) of cell culture grade water. Bacto Proteose peptone should be added in several batches. It takes around 2 h to dissolve the materials. 10× SPP stock is stored at −20 °C.
3. Our previous study suggested that coDel can be induced if an ~1 kbp segment shares ~90% identity with the targeting sequence [13]. The results shown in Fig. 3 also support this view. We therefore suggest choosing a target sequence that does not share >90% identity with any other genomic locations in all possible 200 bp windows in the genome. Nonetheless, off-target deletion may be induced in an unpredictable manner even with a carefully chosen target. Therefore, it is important to perform a genetic rescue experiment to confirm whether an observed phenotype is caused by the deletion at the targeted locus but not by some off-target deletion.
4. These designs work well for assembly by the NEBuilder HiFi assembly system: FW = 5′-CTTTATTGTTATCATCTTATGACCGC-target-3′; RV = 5′-CTCATCAAGTTGTAATGCTAAAATGC-target-3′.

5. 5′-GCATGCGGCCGC-target-3′ for both primers works fine for ligation.
6. For preparation of materials and detailed protocols for biolistic transformation, refer to [22].
7. Because pMcoDel is based on the high-copy rDNA vector pD5H8 [23], a targeting construct can also be introduced into the new MAC by electroporation [24]. However, we have not tested whether coDel could be induced by electroporation.
8. Aliquots of 1.33 mL of DMSO from freshly opened bottles were placed into 15-mL tubes and stored at −20 °C. Add 8.67 mL 10 mM Tris (pH 7.5) immediately before preparation of frozen cell stocks.
9. The AT-rich genome of *Tetrahymena* sometimes prevents efficient delivery of synthetic double-stranded DNA corresponding to some *Tetrahymena* genome loci. If an alternative region cannot be chosen, multiple shorter synthetic single-stranded DNAs can be stitched by overlapping PCR to produce the desired DNA fragment.
10. For detailed *Tetrahymena* cell culture and mating conditions, *see* [21, 25].
11. Cells lysed in CHAOS can be stored at −20 °C for more than several months.
12. Solution will be opaque. If not, add ~30 μL more phenol extraction buffer.
13. Subheadings 3.1, **steps 13–21** are not necessary to amplify up to ~1 kb DNA from the MAC by standard Taq polymerases. These steps enhance amplification of longer DNA from the MAC or any size of the MIC DNA by Taq polymerase. These steps are often necessary to amplify DNA from any genomic region by high-fidelity polymerases such as Q5 DNA polymerase (NEB) or PrimeSTAR HS DNA Polymerase (Takara).
14. pMcoDel is derived from the rDNA vector pD5H8, which contains the MIC copy of the rDNA locus containing a point mutation that causes the derived rRNA to be insensitive to paromomycin [23]. Because mature, functional rDNA can be produced from the rDNA vector only when it undergoes programmed genome rearrangement in the developing new MAC, all paromomycin-resistant cells obtained by introduction of pMcoDel should be sexual progeny of B2086 and CU428 and have pMcoDel-derived mature rDNA in the (new) MAC.
15. It is preferable to avoid plates in which all wells have growing cells because, in such a plate, the cells in each well were likely established from multiple transformation events.

16. The efficiency of coDel differs among targets. If more than six cell lines with appropriate deletions at the targeted locus are not obtained, it is necessary to screen more transformants, target different regions of the targeted gene, and/or use a longer targeting sequence. However, in our experience, extending the targeting sequence longer than 1 kb does not obviously enhance coDel.
17. If the wild-type locus is not detected in the selected line before cloning in Subheading 3.3, **step 4**, examining a few clones is probably sufficient to obtain a clone in which all the copies of the target locus in the polyploid MAC are deleted.
18. Because coDel induces DNA elimination only in the MAC, genomic PCR may detect the intact wild-type (nondeleted) copy of the targeted locus in the MIC even when all the copies in the MAC have deletions. Therefore, complete absence of the wild-type copy of the targeted gene in the MAC in selected strains may be necessary to be confirmed at the product level (such as by western blot, RT-PCR, or northern blot).
19. A series of truncated sequences and sequences having base replacements in defined intervals.
20. When the 401 bp targeting sequence has base replacement mutations every 10 bp and scnRNAs use X nucleotides as seeds to initiate recognition of the targeted sequences, the total length of the targetable sequence by such scnRNAs (Υ) can be estimated as $\Upsilon = 401-(x - 1) - 40x$. From the truncation study above, the minimum length of the targeting sequence to induce some DNA elimination at this targeted locus was ~150 bp (Fig. 3b). Because the targetable sequence length (Υ) becomes ~150 bp when $X = 6$ (Fig. 3c), we can estimate that the scnRNAs use more than 6 nucleotides for their seed sequence. This estimation of the seed length also explains the result obtained by the 401 bp targeting sequence with mutations every 20 bp, which is estimated to reduce the actual targetable sequence length to 276 bp if the seed size is 6 nt (Fig. 3c). A 276 bp targeting sequence is predicted to induce coDel in 15.3% of transformants (Fig. 3b), which was close to the actual fraction of transformants with deletion by this targeting sequence (=21.7%, Fig. 3a). The seed of scnRNA in Twi1p-scnRNA complex during *Tetrahymena* DNA elimination is most likely 6–8 bases, as in other Argonaute-mediated systems, and some mismatch(s) in the seed can be tolerated if other regions of the scnRNA are also complementary to the target.

Acknowledgments

We would like to thank Tomoko Noto and Julie Saksouk for technical support and Katrina Woolcock for the initiation of the investigation of scnRNA–target interactions. Our research was supported by an advanced grant from the "'Investissements d'avenir" program Labex EpiGenMed (ANR-10-LABX-12-01) and an "Accueil de Chercheurs de Haut Niveau" grant (ANR-16-ACHN-0017) from the French National Research Agency and Emergence de projets (2019-E10) from Cancéropôle Grand Sud-Ouest to K.M.

References

1. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell 110:689–699
2. Fang W, Wang X, Bracht JR et al (2012) Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. Cell 151:1243–1255
3. Bouhouche K, Gout J-F, Kapusta A et al (2011) Functional specialization of Piwi proteins in paramecium tetraurelia from post-transcriptional gene silencing to genome remodelling. Nucleic Acids Res 39:4249–4264
4. Couvillion MT, Sachidanandam R, Collins K (2010) A growth-essential Tetrahymena Piwi protein carries tRNA fragment cargo. Genes Dev 24:2742–2747
5. Couvillion MT, Lee SR, Hogstad B et al (2009) Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of Tetrahymena thermophila. Genes Dev 23:2016–2032
6. Furrer DI, Swart EC, Kraft MF et al (2017) Two sets of Piwi proteins are involved in distinct sRNA pathways leading to elimination of germline-specific DNA. Cell Rep 20:505–520
7. Hamilton EP, Kapusta A, Huvos PE et al (2016) Structure of the germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. Elife 5:e19090. https://doi.org/10.7554/eLife.19090
8. Schoeberl UE, Kurth HM, Noto T, Mochizuki K (2012) Biased transcription and selective degradation of small RNAs shape the pattern of DNA elimination in Tetrahymena. Genes Dev 26:1729–1742
9. Noto T, Kurth HM, Kataoka K et al (2010) The Tetrahymena argonaute-binding protein Giw1p directs a mature argonaute-siRNA complex to the nucleus. Cell 140:692–703
10. Mochizuki K, Gorovsky MA (2005) A dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. Genes Dev 19:77–89
11. Malone CD, Anderson AM, Motl JA et al (2005) Germ line transcripts are processed by a dicer-like protein that is essential for developmentally programmed genome rearrangements of Tetrahymena thermophila. Mol Cell Biol 25:9151–9164
12. Noto T, Kataoka K, Suhren JH et al (2015) Small-RNA-mediated genome-wide trans-recognition network in Tetrahymena DNA elimination. Mol Cell 59:229–242
13. Hayashi A, Mochizuki K (2015) Targeted gene disruption by ectopic induction of DNA elimination in tetrahymena. Genetics 201:55–64
14. Tian M, Loidl J (2018) A chromatin-associated protein required for inducing and limiting meiotic DNA double-strand break formation. Nucleic Acids Res 46:11822–11834
15. Akematsu T, Fukuda Y, Garg J et al (2017) Post-meiotic DNA double-strand breaks occur in Tetrahymena, and require topoisomerase II and Spo11. Elife 6:e26176. https://doi.org/10.7554/eLife.26176
16. Urbanska P, Joachimiak E, Bazan R et al (2018) Ciliary proteins Fap43 and Fap44 interact with each other and are essential for proper cilia and flagella beating. Cell Mol Life Sci 75:4479–4493
17. Soh AWJ, van Dam TJP, Stemm-Wolf AJ et al (2020) Ciliary force-responsive striated fibers promote basal body connections and cortical interactions. J Cell Biol 219(1):e201904091. https://doi.org/10.1083/jcb.201904091
18. Bazan R, Schröfel A, Joachimiak E et al (2021) Ccdc113/Ccdc96 complex, a novel regulator of ciliary beating that connects radial spoke 3 to

dynein g and the nexin link. PLoS Genet 17: e1009388

19. Tian M, Agreiter C, Loidl J (2020) Spatial constraints on chromosomes are instrumental to meiotic pairing. J Cell Sci 133(22): jcs253724. https://doi.org/10.1242/jcs.253724
20. Noto T, Mochizuki K (2018) Small RNA-mediated trans-nuclear and trans-element Communications in Tetrahymena DNA elimination. Curr Biol 28:1938–1949.e5
21. Cassidy-Hanley DM (2012) Tetrahymena in the laboratory: strain resources, methods for culture, maintenance, and storage. Methods Cell Biol 109:237–276
22. Bruns PJ, Cassidy-Hanley D (2000) Biolistic transformation of macro- and micronuclei. Methods Cell Biol 62:501–512
23. Yao MC, Yao CH (1991) Transformation of Tetrahymena to cycloheximide resistance with a ribosomal protein gene through sequence replacement. Proc Natl Acad Sci U S A 88: 9493–9497
24. Gaertig J, Gorovsky MA (1992) Efficient mass transformation of Tetrahymena thermophila by electroporation of conjugants. Proc Natl Acad Sci U S A 89:9196–9200
25. Chalker DL (2012) Transformation and strain engineering of Tetrahymena. Methods Cell Biol 109:327–345
26. Noto T, Mochizuki K (2017) Whats, hows and whys of programmed DNA elimination in Tetrahymena. Open Biol 7:170172

Chapter 4

Planarian PIWI–piRNA Interaction Analysis Using Immunoprecipitation and piRNA Sequencing

Makoto Kashima, Atsumi Miyata, and Norito Shibata

Abstract

The freshwater planarian *Dugesia japonica* is a good in vivo model for studying the function of *piwi* genes in adult pluripotent stem cell (aPSC) due to their abundant aPSCs. Generally, PIWI family proteins encoded by *piwi* genes bind to small noncoding RNAs called piRNAs (PIWI–interacting piRNAs). The analysis of PIWI–piRNA complexes in the planarian is useful for revealing the functions of *piwi* genes in the aPSC system. In this chapter, we present an immunoprecipitation protocol for PIWI–piRNA complexes from whole planarians.

Key words PIWI, piRNA, Planarian, Neoblast, RNA-binding protein, Stem cell, Immunoprecipitation

1 Introduction

Asexually reproducing invertebrates, such as sponges, hydra, and planarians, can maintain adult pluripotent/totipotent stem cells (aPSCs) throughout their adulthood, while sexually propagating vertebrates can only maintain pluripotent stem cells at the very early stages of development [1–3]. The aPSCs of these invertebrates commonly express *piwi* genes, which are known to be guardians of germline genome integrity in animals [4, 5]. Although *piwi* genes are widely used as aPSC markers in invertebrates, the molecular functions of these stem cells remain ambiguous.

The freshwater planarian *Dugesia japonica* is a good in vivo model for studying the function of *piwi* genes in aPSC due to their abundant aPSCs, called neoblasts, and well-established experimental techniques [6–9]. Neoblasts comprise approximately 30% of total adult planarian cells and supply differentiated cells on demand for the maintenance of tissue homeostasis and asexual reproduction [10, 11]. Owing to the neoblasts, planarians can regenerate the whole body from even a tiny fragment [12, 13]. Neoblasts express

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_4,

three *piwi* genes (*DjpiwiA*, *DjpiwiB*, and *DjpiwiC*) [14, 15]. Knockdown of *DjpiwiB* or *DjpiwiC* causes severe defects in regeneration, tissue homeostasis, and maintenance of neoblasts [16, 17]. In particular, DjPiwiB plays multiple roles in both neoblasts and differentiated cells by regulating functional coding genes as well as transposable elements [17, 18]. Generally, PIWI family proteins encoded by *piwi* genes bind to small non-coding RNAs called piRNAs (PIWI–interacting piRNAs) [19]. PIWI–piRNA complexes repress their target genes transcriptionally or posttranscriptionally via piRNA base-pairing [19]. Thus, the analysis of PIWI–piRNA complexes in the planarian is useful for revealing the functions of *piwi* genes in the aPSC system. In a previous study, we established an immunoprecipitation protocol for PIWI–piRNA complexes from whole planarians [17]. Here, we present an updated version of the protocol using SMARTer technologies with improved reproducibility and sensitivity.

2 Materials

All buffers should be prepared using nuclease-free water. Nuclease-free disposable pipette tips and tubes should be used.

2.1 Maintenance of Planarian

1. Instant Ocean (Aquarium Systems).
2. Pure water (e.g., water purified with an Elix® Water Purification System (Merck)).
3. A freshwater planarian, *D. japonica*.
4. Chicken liver.

2.2 Preparation of Lysate

1. Medium salt buffer (MSB): 50 mM Tris–HCl (pH 7.5), 150 mM NaCl, and 0.05% NP-40 (*see* **Note 1**).
2. cOmplete™, EDTA-free Protease Inhibitor Cocktail (Roche).
3. Liquid nitrogen.
4. RNasin® Plus Ribonuclease Inhibitor (Promega).

2.3 Immunoprecipitation of Piwi–piRNA Complex

1. House-made anti-DjPiwiB rabbit-polyclonal antibody [17].
2. Dynabeads™ Protein G for Immunoprecipitation (Thermo Fisher).

2.3.1 Using Magnetic Bead

3. A magnetic stand for 1.5/2.0 mL tube (e.g., DynaMag SPIN (Thermo Fisher)).
4. 10 mg/mL Yeast tRNA (Thermo Fisher).

2.3.2 Using Sepharose Bead

1. House-made anti-DjPiwiB rabbit-polyclonal antibody [17].
2. Protein G Sepharose 4 Fast Flow (Cytiba).

2.4 Immunoprecipitation Quality Assessment (Optional)

1. Reagents and equipment required for SDS-PAGE.
2. Reagents for silver staining (e.g., Sil-Best Stain-Neo for Protein and Nucleic Acid/PAGE (Nacalai Tesque)).

2.5 Purification of Immunoprecipitated RNA

1. ISOGEN-LS (Nippongene).
2. Chloroform.
3. Isopropanol.
4. 80% EtOH.
5. Gene-Packman Coprecipitant (Nacalai Tesque).
6. 3 M Sodium acetate (pH 5.2).

2.6 Size Selection of piRNA

1. Prestained molecular weight marker: DynaMarker®, Prestain Marker for Small RNA plus (BioDynamics).
2. 2× gel loading buffer: 95% formamide, 18 mM EDTA, 0.025% SDS, 0.025% bromophenol blue, and 0.025% xylene cyanol.
3. 7 M urea 10% polyacrylamide gel (gel: 9 cm width, 7.5 cm height, and 1 mm thick; well: 5 mm width and 1 cm depth).
4. Tris-borate-EDTA buffer (TBE; 0.5×): 45 mM Tris-borate and 1 mM EDTA.
5. Razor blade.
6. RNA elution buffer (0.5 M): sodium acetate (pH 5.2), 0.1 mM EDTA, 0.1% SDS.
7. Phenol:Chloroform:Isoamyl Alcohol in 25:24:1 and mixed (pH 5.2).
8. Chloroform.
9. 3 M Sodium acetate (pH 5.2).
10. 100% Ethanol.
11. 70% Ethanol.
12. Gene-Packman Coprecipitant (Nacalai Tesque).

2.7 piRNA Quantity and Quality Assessment

1. Quantus™ Fluorometer (Promega).
2. QuantiFluor® RNA System (Promega).
3. Agilent 2100 Bioanalyzer (Agilent).
4. Agilent Small RNA Kit (Agilent).

2.8 piRNA-Seq Library Preparation

1. SMARTer® smRNA-Seq Kit for Illumina® (TaKaRa).
2. NucleoSpin Gel and PCR Clean-Up kit (TaKaRa) supplied with SMARTer .smRNA-Seq Kit for Illumina.

2.9 piRNA-Seq Library Quality Assessment

1. Agilent High Sensitivity DNA Kit (Agilent).

2.10 Sequencing

1. A HiSeq platform (Illumina).

3 Methods

All procedures are carried out at room temperature unless otherwise specified.

3.1 Maintenance of Planarian

1. Culture freshwater planarian *D. japonica* at 24 °C in pure water containing 0.05 g/L Instant Ocean. Feed the planarians with chicken liver once or twice a week. The detailed protocol is described in [9].
2. Select planarians that are approximately 5 mm in length and do not feed them, in order to limit contamination from chicken liver and reduce the endogenous protease and nuclease activities, for at least one week prior to the following experiments.

3.2 Preparation of Lysate

1. Add a tablet of cOmplete™, EDTA-free Protease Inhibitor Cocktail to 10 mL MSB buffer. Mix it by gently inverting the tube. Keep the mixture on ice.
2. Transfer 20 planarians into a 1.5-mL tube.
3. Spin down the tube to keep the planarians at the bottom and discard the culture medium.
4. Add 250 μL of the MSB buffer containing the protease inhibitor to the tube, and freeze it immediately using liquid nitrogen.
5. Thaw the sample on ice and homogenize it thoroughly by gently pipetting up and down using a 200 μL tip.
6. Freeze the sample with liquid nitrogen, thaw it on ice, and homogenize it thoroughly by gently pipetting up and down using a 200 μL tip.
7. Add 1 μL of RNasin Plus Ribonuclease Inhibitor and mix it by gently inverting the tube, avoid shaking.
8. Centrifuge at 18,000 × *g* for 10 min at 4 °C, and transfer the supernatant into a new tube and place it on ice.
9. Add 200 μL of the MSB buffer containing the protease inhibitor to the pellet, mix well, and freeze it using liquid nitrogen.
10. Thaw the sample on ice and homogenize it thoroughly by pipetting up and down using a 200 μL tip.
11. Centrifuge at 18,000 × *g* for 10 min at 4 °C, and transfer the supernatant into a new tube on ice.
12. Repeat **steps 9–11** once more.
13. Mix all the supernatants.

3.3 Immunoprecipitation of Piwi–piRNA Complex

Researcher can choose one of the following protocols (magnetic bead or Sepharose bead).

3.3.1 Using Magnetic Bead

1. Resuspend magnetic beads (Dynabeads) in the vial by vortexing.
2. Transfer 50 μL of Dynabeads into a 2.0-mL tube.
3. Place the tube on the magnet to separate the beads from the solution and discard the supernatant.
4. Add 1 mL of MSB buffer and 10 μg of anti DjPiwiB antibody to the tube.
5. Incubate the tube with rotation for an hour at 4 °C.
6. Add 1 μL of 10 mg/mL yeast tRNA to the tube for blocking and incubate it with rotation for an hour at 4 °C.
7. Place the tube on the magnet to separate the beads from the solution and discard the supernatant.
8. Resuspend the beads in 1 mL of MSB buffer by gently pipetting up and down.
9. Rotate the tube for 5 min at 4 °C.
10. Repeat **steps 7–9** twice.
11. Place the tube on the magnet to separate the beads from the solution and discard the supernatant.
12. Resuspend beads in 800 μL of MSB buffer containing 1 μL of RNasin Plus Ribonuclease Inhibitor by gently pipetting up and down.
13. Add 200 μL of the lysate to the tube.
14. Incubate the tube with rotation for an hour at 4 °C.
15. Place the tube on the magnet and discard the supernatant.
16. Resuspend the beads with 1 mL of MSB buffer by gently pipetting up and down.
17. Rotate the tube for 5 min at 4 °C.
18. Repeat **steps 15–17** three times.
19. Place the tube on the magnet and discard the supernatant.
20. Resuspend beads by adding 270 μL of nuclease-free water.

3.3.2 With Sepharose Bead

1. Centrifuge 300 μL of Protein G Sepahrose 4Fast Flow for 1 min at 5000 × *g*, and discard the supernatant.
2. Add 100 μL of MSB buffer and mix well by rotating for 5 min.
3. Centrifuge for 1 min at 5000 × *g*, and discard the supernatant.
4. Add 650 μL of MSB buffer and mix well.
5. Transfer aliquots of 100 μL into new tubes (for six samples).

6. Add 50 μL of the anti-DjPiwiB antibody to each aliquot, and bring up to a final volume of 1 mL using MSB buffer.
7. Incubate the mixtures at 4 °C for 3 h with rotation.
8. Centrifuge for 1 min at 5000 × *g* and discard the supernatant.
9. Add 1 mL of MSB buffer and mix well.
10. Centrifuge for 3 min at 5000 × *g*, and discard the supernatant.
11. Repeat **steps 9** and **10**, three times.
12. After mixing, add 200 μL of the lysate to the tube.
13. Rotate the mixtures at 4 °C for 1 h.
14. Centrifuge for 3 min at 5000 × *g*, and discard the supernatant.
15. Repeat **steps 9** and **10**, three times.
16. Centrifuge for 3 min at 5000 × *g*, and completely discard the supernatant.
17. Resuspend beads in 270 μL of nuclease-free water.

3.4 Immunoprecipitation Quality Assessment (Optional)

During the first trial of immunoprecipitating the Piwi–piRNA complex, validation of the immunoprecipitant is recommended. 20 μL of the sample should be sufficient for the validation. According to a general SDS-PAGE protocol (e.g., [20]), the immunoprecipitant is validated via electrophoresis and silver staining (Fig. 1).

3.5 Purification of Immunoprecipitated RNA

1. Mix 750 μL of ISOGEN-LS and 250 μL of the sample by vortexing. The solution is then incubated for 5 min at room temperature.
2. Centrifuge at 16,000 × *g* for 15 min at 4 °C and transfer the colorless upper aqueous phase (approximately 400 μL) into a new tube (*see* **Note 2**). Do not transfer the red organic phase or interphase.

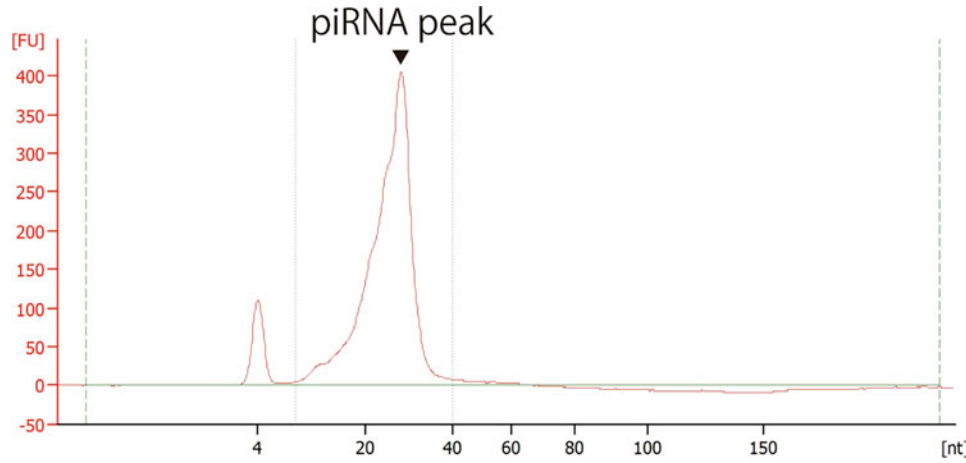


Fig. 1 A typical silver staining pattern of SDS-PAGE of the immunoprecipitant using anti-DjPiwiB antibody

3. Add 200 μL of chloroform to the aqueous phase and mix the sample by vortexing. The solution is then incubated for 5 min at room temperature.
4. Centrifuge at 16,000 × *g* for 10 min at 4 °C and transfer the upper aqueous phase (approximately 400 μL) into a new tube. Do not transfer the lower phase.
5. Add 400 μL of isopropanol, 1 μL of gene-packman, and 40 μL of 3 M sodium acetate (pH 5.2), and mix the sample by vortexing.
6. Centrifuge at 16,000 × *g* for 10 min at room temperature and discard the supernatant without disturbing the RNA pellet.
7. Wash the pellet using 1 mL of 80% EtOH.
8. Centrifuge at 16,000 × *g* for 5 min at room temperature and discard the supernatant without disturbing the RNA pellet.
9. Centrifuge to spin down the remaining liquid and discard the liquid completely.
10. Elute the RNA pellet using 10 μL of nuclease-free water.

3.6 Size Selection of piRNA

The immunoprecipitated RNA would contain RNAs longer than piRNAs, including yeast tRNA. To enrich the piRNAs, size selection is recommended.

1. Add an equal volume of the gel loading buffer to immunoprecipitated RNA. Heat-denature at 65 °C for 3 min and keep on ice.
2. After pre-run of 7 M urea 10% polyacrylamide gel at 150 V for 30 min in 0.5 × TBE buffer, load the samples and prestained molecular weight marker on the gel, and electrophorese at 150 V for 1 h (*see* **Note 3**).
3. Remove one side of the gel plate, and take out the other gel plate with the gel, and place it on a white paper so that the gel side is on top. Using the prestained molecular weight marker as an indicator, a razor blade is used to excise the gel between 20 and 40 bases of the lane to which the sample is applied (*see* **Note 4**). Transfer the excised gel into a 1.5-mL centrifuge tube (*see* **Note 5**).
4. The excised gel is kept in 1.4 mL RNA elution buffer at 37 °C overnight.
5. Spin down the eluate and aliquot it into 650 μL portions and transfer them to two new 1.5-mL centrifuge tubes (*see* **Note 6**).
6. Add 700 μL of phenol/chloroform (pH 5.2) to each tube and mix by vortexing. Centrifuge at 12,000 × *g* for 10 min at room temperature, and transfer 650 μL of the upper aqueous phase to a new 1.5-mL centrifuge tube (2 tubes in total).

7. Add 700 μL of phenol/chloroform (pH 5.2) to each tube and mix by vortexing. Centrifuge at 12,000 × *g* for 10 min at room temperature, and transfer 630 μL of the upper aqueous phase to a new 1.5-mL centrifuge tube (2 tubes in total).
8. Add 700 μL of chloroform to each tube and mix by vortexing. Centrifuge at 12,000 × *g* for 2 min at room temperature and collect 600 μL of the upper aqueous phase from two tubes (total 1200 μL). The samples are divided into 400 μL portions and transferred into three new 1.5-mL centrifuge tubes.
9. Add 1 mL of 100% ethanol, 2 μL of Gene-Packman coprecipitant, and 40 μL of 3 M sodium acetate (pH 5.2) and mix thoroughly by gentle inversion (3 tubes total) (*see* **Note 7**). Centrifuge at 15,000 × *g* for 60 min at room temperature and discard the supernatant without disturbing the RNA pellet.
10. Wash the RNA pellet using 500 μL of 70% EtOH.
11. Centrifuge at 15,000 × *g* for 15 min at room temperature and discard the supernatant without disturbing the RNA pellet.
12. Air-dry the RNA pellet.
13. Add a total of 400 μL of nuclease-free water to the three tubes and hold at 4 °C for 1 h to dissolve the RNA pellet. Collect samples of all the three tubes into one 1.5-mL centrifuge tube.
14. Add 1 mL of 100% ethanol and 40 μL of 3 M sodium acetate (pH 5.2), and mix thoroughly by gentle inversion. Centrifuge at 15,000 × *g* for 60 min at room temperature and discard the supernatant without disturbing the RNA pellet.
15. Wash the RNA pellet using 500 μL of 70% EtOH.
16. Centrifuge at 15,000 × *g* for 15 min at room temperature and discard the supernatant without disturbing the RNA pellet.
17. Repeat **steps 15** and **16**, once.
18. Air-dry the RNA pellet.
19. Add 7 μL of nuclease-free water and hold at 4 °C for 1 h to dissolve the RNA pellet. Keep the sample at −80 °C until use.

3.7 piRNA Quantity and Quality Assessment

1. Quantify 1 μL of the size-selected RNA using the Quanti-Fluor® RNA System and Quantus™ Fluorometer according to the manufacturer's instructions. The typical yield should be around 3 ng/μL (approximately 20 ng in total).
2. Analyze 1 μL of the size-selected RNA using Agilent Small RNA Kit and Agilent 2100 Bioanalyzer according to the manufacturer's instructions. The major peak should be approximately 30 nt (Fig. 2).

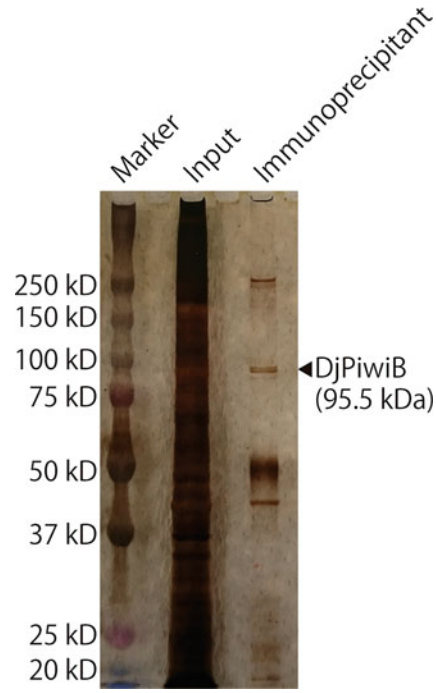


Fig. 2 A typical electrophoresed pattern of DjPiwiB–interacting piRNAs

3.8 Small RNA-Seq Library Preparation

The RNA solutions are kept on ice before setting up the reaction. Conduct all reactions using a thermal cycler.

1. Thaw all reagents of the SMARTer smRNA-Seq Kit on ice, except the reagents for amplification.
2. Mix 6 ng of the size-selected RNA solution with 0.25 μL of Poly (A) polymerase, 0.25 μL of RNase inhibitor, and 2.5 μL of smRNA Mix 1. Bring up the volume of the solution to 10 μL using nuclease-free water.
3. Mix the tube by tapping and spin it down.
4. Incubate at 16 °C for 5 min and immediately transfer the tube onto ice for more than 1 min (and not more than 5 min).
5. Add 1 μL of 3′ smRNAdT Primer to the tube, mix by tapping, and spin down.
6. Incubate at 72 °C for 3 min and transfer the tube onto ice for at least 2 min.
7. Add 6.5 μL of smRNA Mix2, 0.5 μL of RNase inhibitor, and 2 μL of PrimeScript RT to the mixture. The total sample volume is 20 μL.
8. Mix the tube by tapping and spin it down.

9. Place the tube in the thermal cycler which is preheated to 42 °C. Start the following program:

   42 °C for 60 min.

   70 °C for 10 min.

   4 °C hold

10. Mix 10 μL of cDNA, 12 μL of nuclease-free water, 25 μL of SeqAmp PCR Buffer (2×), 1 μL of SeqAmp DNA Polymerase, 1 μL of one each of Primer F (F1–F8), and 1 μL of one each of Primer R (R1–R12). The total sample volume is 50 μL.
11. Mix the tube by tapping and spin it down.
12. Place the tube in a preheated thermal cycler with a heated lid and run the following program:

    98 °C for 1 min.

    Followed by 13 cycles:

    98 °C for 10 s.

    60 °C for 5 s.

    68 °C for 10 s.

    4 °C hold.
13. Purify all the PCR products using the NucleoSpin Gel and PCR Clean-Up kit according to the manufacturer's instructions. Elute sample using 30 μL of NE buffer provided.

3.9 piRNA-Seq Library Quality Assessment

Analyze 1 μL of the piRNA-Seq library according to the manufacturer's instructions for the Agilent High Sensitivity DNA Kit (Fig. 3). The major peak detected should be approximately 185 bp (Fig. 3).

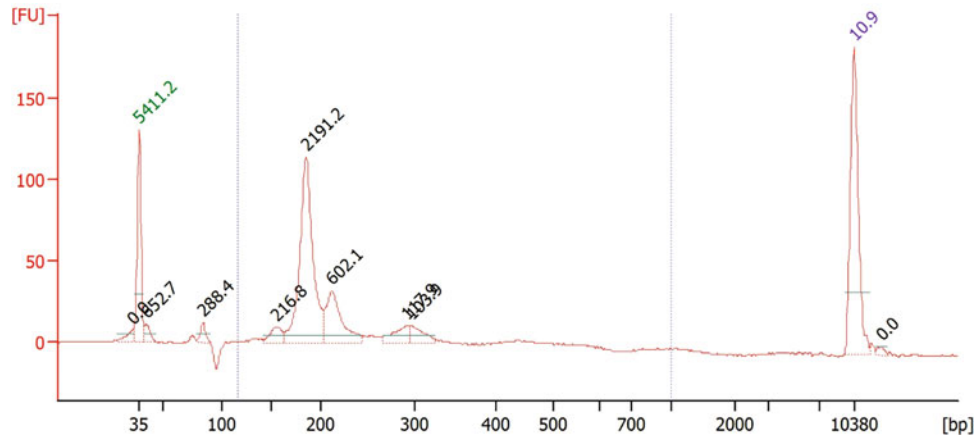


Fig. 3 A typical electrophoresed pattern of piRNA-Seq library

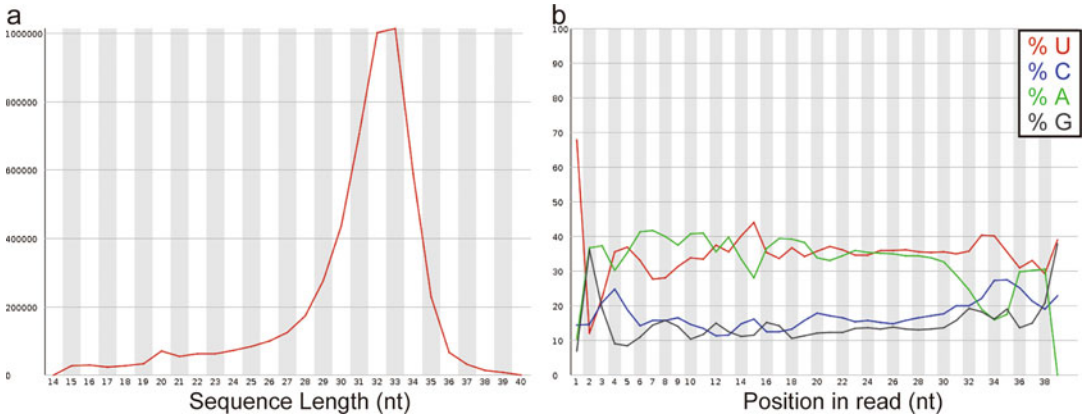


Fig. 4 Typical outputs of fastqc for DjPiwiB–interacting piRNAs. (**a**) A length distribution of the piRNAs. (**b**) A base content distribution of each position of the piRNAs

3.10 Sequencing

Sequence the piRNA library using an illumina sequencing platform according to the manufacturer's instructions. A 50 bp of single-end sequencing is of sufficient length for the library.

3.11 Bioinformatic Analysis of Planarian piRNA

1. Execute "cutadapt -m 15 -u 3 -a AAAAAAAAAA input.fastq > output.fastq" (*see* **Note 8**). This command retains only reads 15 nt or longer after trimming (-m 15), trims the first 3 nt of all reads, which are extra nts inserted due to the SMART template-switching mechanism (-u 3), and removes poly A added by Poly (A) polymerase at the 3′ end (6a AAAAAAAAAA) (*see* **Note 9**).
2. Run fastqc (*see* **Note 10**) to check the features of the piRNAs. Expected results are of 32 nt length in average, with strong uracil preference at the 5′ end (Fig. 4).

4 Notes

1. Since Tris can be degraded by diethyl pyrocarbonate (DEPC), RNase in the buffer and bottle cannot be deactivated using DEPC. If DEPC treatment is needed, treat the medium without Tris using DEPC, then autoclave it to degrade DEPC, and add nuclease-free commercial Tris–HCl buffer to the DEPC-treated medium.
2. To avoid contamination of the organic phase and the interphase, use a 200 μL tip, not a 1 mL tip.
3. The distance between 20 and 40 bases of prestained molecular weight marker lane will be about 1.5 cm.
4. The size of the excised gel will be approximately 7 mm by 1.6 cm ($W \times L = 1.12\ cm^2$).

5. Do not spin down to prevent the gel from breaking.
6. If the eluate is not enough for 1300 μL (650 μL × 2), RNA elution buffer can be used to top up the volume to 1300 μL.
7. Keep the sample at room temperature after adding the co-precipitant because the recovery rate decreases with cooling.
8. To install cutadapt, install conda and execute "pip install cutadapt."
9. By default, the cutadapt will accept an error rate of 10% in adapter detection. This parameter can be changed depending on the purpose.
10. To install fastqc, refer to https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Acknowledgments

This work was supported by JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (19K16149) for M. K.; Tomizawa Jun-ichi & Keiko Fund of Molecular Biology Society of Japan for Young Scientists to A. M. We would like to thank Editage (www.editage.com) for English language editing.

References

1. Funayama N (2018) The cellular and molecular bases of the sponge stem cell systems underlying reproduction, homeostasis and regeneration. Int J Dev Biol 62:513–525. https://doi.org/10.1387/ijdb.180016nf
2. Sánchez Alvarado A, Yamanaka S (2014) Rethinking differentiation: stem cells, regeneration, and plasticity. Cell 157:110–119. https://doi.org/10.1016/j.cell.2014.02.041
3. Agata K, Nakajima E, Funayama N et al (2006) Two different evolutionary origins of stem cell systems and their molecular basis. Semin Cell Dev Biol 17:503–509. https://doi.org/10.1016/j.semcdb.2006.05.004
4. Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12:246–258. https://doi.org/10.1038/nrm3089
5. van Wolfswinkel JC (2014) Piwi and potency: PIWI proteins in animal stem cells and regeneration. Integr Comp Biol 54:700–713. https://doi.org/10.1093/icb/icu084
6. Baguñà J, Slack JMW (1981) Planarian neoblasts. Nature 290:14–15. https://doi.org/10.1038/290014b0
7. Wolff EDF (1948) Sur la migration des cellules de regeneration chez les planaires. Rev Suisse Zool 55:218–227
8. Kashima M, Agata K, Shibata N (2020) What is the role of PIWI family proteins in adult pluripotent stem cells? Insights from asexually reproducing animals, planarians. Develop Growth Differ 62:407–422. https://doi.org/10.1111/dgd.12688
9. Rink JC (2018) Planarian regeneration Nethods and protocols. In: Methods in Molecular Biology. Springer, New York
10. Hayashi T, Asami M, Higuchi S et al (2006) Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. Develop Growth Differ 48:371–380. https://doi.org/10.1111/j.1440-169X.2006.00876.x
11. Orii H, Sakurai T, Watanabe K (2005) Distribution of the stem cells (neoblasts) in the planarian Dugesia japonica. Dev Genes Evol 215:143–157. https://doi.org/10.1007/s00427-004-0460-y
12. Agata K, Watanabe K (1999) Molecular and cellular aspects of planarian regeneration. Semin Cell Dev Biol 10:377–383. https://doi.org/10.1006/scdb.1999.0324

13. Morgan TH (1898) Experimental studies of the regeneration of Planaria maculata. In: Exp Stud Regen Planaria maculata, vol 7. Creative Media Partners, LLC, New York City, pp 364–397. https://doi.org/10.1007/BF02161491
14. Yoshida-Kashikawa M, Shibata N, Takechi K, Agata K (2007) DjCBC-1, a conserved DEAD box RNA helicase of the RCK/p54/Me31B family, is a component of RNA-protein complexes in planarian stem cells and neurons. Dev Dyn 236:3436–3450. https://doi.org/10.1002/dvdy.21375
15. Hayashi T, Shibata N, Okumura R et al (2010) Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. Develop Growth Differ 52:131–144. https://doi.org/10.1111/j.1440-169X.2009.01157.x
16. Kashima M, Kumagai N, Agata K, Shibata N (2016) Heterogeneity of chromatoid bodies in adult pluripotent stem cells of planarian Dugesia japonica. Develop Growth Differ 58: 225–237. https://doi.org/10.1111/dgd.12268
17. Shibata N, Kashima M, Ishiko T et al (2016) Inheritance of a nuclear PIWI from pluripotent stem cells by somatic descendants ensures differentiation by silencing transposons in planarian. Dev Cell 37:226–237. https://doi.org/10.1016/j.devcel.2016.04.009
18. Kashima M, Agata K, Shibata N (2018) Searching for non-transposable targets of planarian nuclear PIWI in pluripotent stem cells and differentiated cells. Develop Growth Differ 60: 260–277. https://doi.org/10.1111/dgd.12536
19. Iwasaki YW, Siomi MC, Siomi H (2015) PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem 84:405–433. https://doi.org/10.1146/annurev-biochem-060614-034258
20. Judd RC (1996) SDS-polyacrylamide gel electrophoresis of peptides. In: The protein protocols handbook, 2nd edn. Humana Press, New Jersey, pp 101–107

Chapter 5

Isolation and Processing of Bovine Oocytes for Small RNA Sequencing

Minjie Tan, Helena T. A. van Tol, and Elke F. Roovers

Abstract

In modern biomedical research, mice have been the mammalian model system of choice to investigate molecular pathways for potential future medical applications. Over the last years, it has become clear that female mice employ an exceptional piRNA pathway-independent mechanism to neutralize transposon activity in the ovary. In other model organisms studied to date, the piRNA pathway is indispensable for efficient targeting of transposable elements and fertility in both males and females. Moreover, recent studies have demonstrated that in other mammals, including humans, the piRNA pathway is highly active in the female germline as well, indicating that the situation in the mouse female germline is anomalous. For this reason, novel models to study piRNA pathways in female mammalian germlines are currently emerging, including *Bos taurus*. Here we describe a protocol for isolation and downstream processing of female bovine tissues in order to perform downstream applications including piRNA sequencing.

Key words Cumulus oocyte complexes (COCs), GV oocytes, piRNAs, β-Oxidation, PIWIL3, RNA-sequencing

1 Introduction

The piRNA pathway is a predominantly germline-specific RNA silencing pathway that protects the genome of germ cells through the targeted inactivation of transposable elements. Since the discovery of the piRNA pathway, it has become clear that in most species studied, including *Drosophila*, silkworm and zebrafish, the piRNA pathway plays an essential role in the function of both male and, particularly, female gonads [1–3]. Animals carrying mutations in piRNA pathway components typically show fertility defects [2, 4–6]. An interesting exception known thus far is the mouse, in which fertility defects are restricted to males, whereas in females, the piRNA pathway seems to be dispensable and absent during the majority of the life cycle [7–9]. Since biomedical research relies heavily on mice and rats as a model for the mammalian situation, it has been assumed that these results could be extrapolated to

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_5,

other mammals as well, including humans. However, a mammalian-specific PIWI paralog, PIWIL3, exists in most mammalian species but is absent in Muridae, comprising mice and rats. The important discovery that Muridae ovaries express a Dicer isoform accompanied by a unique population of transposon-targeting siRNAs, raised suspicions about the representability of Muridae for piRNA pathway function in females [10]. This led to the investigation of bovine, macaque and human ovary tissue, bovine oocytes, in vitro fertilized bovine embryos, and human fetal gonads, which confirmed that indeed, in many mammals, the female germline displays a remarkable variety of PIWI proteins and piRNA species [11]. A high resemblance was found to both transposon-targeting as well as pachytene piRNA species typically found in murine male gonads. In addition, high levels of PIWIL3 protein expression were detected together with a novel piRNA population, starting from mature oocytes up to early stages of in vitro fertilized embryos [11, 12]. These exciting results open up an entire new chapter in the piRNA field. Many fundamental questions remain unanswered including the overall requirement of the piRNA pathway in mammalian ovaries for fertility and the function of maternally inherited PIWIL3-piRNA complexes. This comes with many challenges including the establishment of new model organisms, genetic tools, and adjusted sample preparation. However, this development will be essential since piRNA biology in mice seems to be an exception, whereas many other mammalian models used thus far have severe limitations for follow-up studies for ethical and practical reasons. The establishment of CRISPR-Cas9-mediated genome editing has significantly lowered the threshold for the establishment of alternative research models that encode PIWIL3, among other PIWI proteins and piRNA species. Species including hamsters, guinea pigs, and rabbits are attractive alternatives for future piRNA research, with the first reports confirming piRNA pathway relevance in hamsters already available [13]. Still, adaptation and optimization of many existing protocols and techniques will be needed. Here, we describe how to isolate bovine oocytes from ovaries and the downstream processing required for sequencing of small RNA species including piRNAs.

2 Materials

2.1 Bovine Oocyte Collection

1. Fresh bovine ovarian tissue.
2. 0.9% NaCl.
3. Penicillin/streptomycin.
4. Vacuum suction system.
5. 18-gauge needle, winged infusion set (1.2 mm × 40 mm).

6. HEPES-buffered M199.
7. Recombinant human follicle-stimulating hormone.
8. Fetal Calf Serum.

2.2 RNA Extraction

1. TRIzol® Reagent.
2. Pellet pestle for Eppendorf tubes.
3. Chloroform.
4. Isopropanol.
5. Ethanol.
6. Nuclease-free water.
7. GlycoBlue™ coprecipitant (Ambion).

2.3 β-Oxidation

1. *mir*Vana™ miRNA Isolation Kit (Ambion).
2. 200 mM $NaIO_4$.
3. 5× Borate buffer: 148 mM borax, 148 mM boric acid, pH 8.6.
4. Glycerol.
5. 3 M Sodium Acetate, pH 5.5.
6. Isopropanol.
7. GlycoBlue™ coprecipitant.

3 Methods

3.1 Bovine Ovary Collection

1. Bovine ovaries are collected from a slaughterhouse and transported to the laboratory in a polystyrene box at 30 °C (*see* **Note 1**). Collection should take place as soon as possible, at most within 2 h after slaughter.
2. After rinsing with clean tap water at 30 °C, the ovaries are maintained in a beaker containing 0.9% NaCl supplemented with 0.1% penicillin/streptomycin (10,000 U/mL) in a 30 °C water bath until use. Continue with oocyte collection as soon as possible.

3.2 Oocyte Collection

1. Follicles that range from ø2 to 8 mm (*see* **Note 2**) are aspirated for the collection of cumulus oocytes complexes (COCs) by using a vacuum suction system (pressure at −0.6 hPa) connected to a 50-mL centrifuge tube, via a winged infusion set (Fig. 1a).
2. In each centrifuge tube, follicular fluid from 15 to 25 ovaries is collected (Fig. 1b and c). Let the follicular aspirates settle for at least 15 min. The COCs and fragments of membrana granulosa can then be found in the sediment.

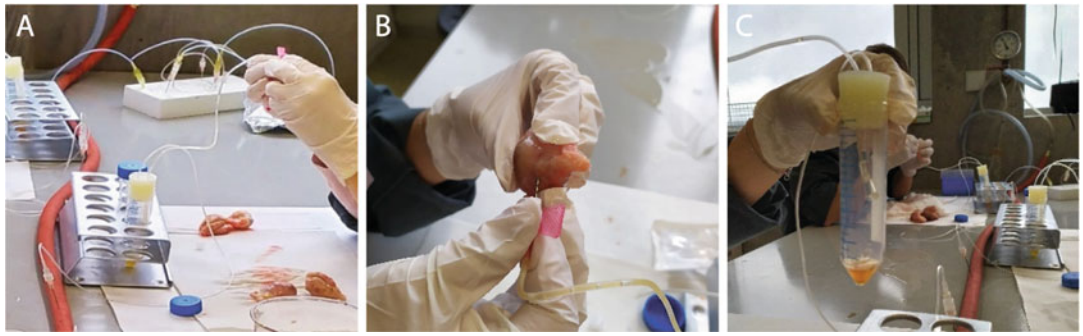


Fig. 1 Follicular fluid collection. (**A**) Vacuum suction system. (**B**) Follicular fluid collection from bovine ovary. (**C**) Follicular fluid in the tube

Fig. 2 Oocyte collection. (**A**) Oocytes searching system. (**B**) Follicular fluid searching under the microscope. (**C**) COCs in M199 medium

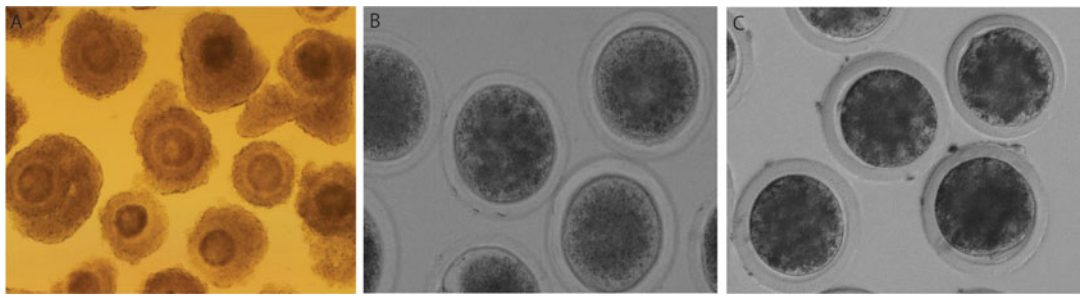


Fig. 3 COCs and oocytes. (**A**) Oocytes surrounded by cumulus cells. (**B**) Germinal Vesicle stage oocytes. (**C**) Metaphase II stage oocytes

3. Transfer the sediment to a 10-cm diameter petri dish (*see* **Note 3**) in order to collect the COCs by using a stereomicroscope with a 12× magnification (Fig. 2). This way, only the oocytes with a multilayered, compact cumulus complex (an intact "cumulus oophorus") are selected for the experiment (Fig. 3a).
4. Collect the COCs in a petri dish with follicular fluid.

5. Wash the collected COCs three times in HEPES-buffered M199 at room temperature and proceed with germinal vesicle (GV) stage processing (Subheading 3.3) or they can be matured until metaphase II (MII) stage (Subheading 3.4). Alternatively, the COCs can be used for IVF as described previously [14].

3.3 GV Stage Oocyte Collection

1. The cumulus cells surrounding the oocytes are removed by pipetting multiple times with a 1 mL Rainin pipette. After that, check for complete denudation by use of a microscope (Fig. 3b).
2. The oocytes are washed three times in PBS and around 1000 oocytes per sequencing sample are collected in 1.5 mL RNase-free tubes, snap frozen in liquid N_2 and stored at −80 °C until RNA isolation.

3.4 MII Stage Oocyte Collection

1. The isolated COCs from Subheading 3.2 are washed once in maturation medium, consisting of HEPES-buffered M199 supplemented with 0.02 U/mL follicle-stimulating hormone, 10% FCS, 100 U/mL penicillin, and 100 μg/mL streptomycin (*see* **Note 4**).
2. After washing, the COCs are transferred to a 4-well plate containing 500 μL maturation medium. Between 35 and 75 COCs can be cultured in each well. In order to avoid evaporation of the maturation medium, 1 mL of sterilized water is pipetted between the wells.
3. The COCs are placed in an incubator (humidified atmosphere, 39 °C with 5% CO_2) for 22–24 h to mature.
4. After maturation, the cumulus cells are removed by pipetting (the same way as in Subheading 3.3), and the successfully matured oocytes displaying the first polar body are collected (Fig. 3c). Per sequencing sample, around 1000 oocytes are pooled this way and washed in PBS and transferred to 1.5 mL RNase-free tubes and snap frozen in liquid N_2. Samples are stored at −80 °C until RNA isolation.

3.5 RNA Extraction

1. Add 350 μL TRIzol to the snap-frozen oocyte pellets. Mash the sample with a pellet pestle (while in TRIzol) until all oocytes/embryos are lysed (*see* **Note 5**). The TRIzol manufacturer's protocol is followed, adjusted for the 350 μL volume of TRIzol and some other minor changes as indicated below.
2. Following the lysis, 100 μL chloroform is added, and the tubes are vortexed vigorously.
3. Centrifuge the samples for 15 min at 12,000 × *g* at 4 °C.
4. Collect the upper (aqueous) phase in a new tube.

5. Add 1 μL GlycoBlue as a co-precipitant. This will increase the pellet mass and will make detection of the pellet easier because of its blue color, which is especially helpful when there is limited sample available.
6. Add 500 μL 100% isopropanol and vortex the sample.
7. Precipitate the RNA at −80 °C for 3 h to overnight.
8. Centrifuge for 30 min to 1 h at 16,000 × *g* at 4 °C and remove the supernatant.
9. Wash the pellet with 75% ethanol. Make sure the pellet detaches during the washing steps by vortexing, so the ethanol can access the salt contaminants.
10. Centrifuge 15 min at 16,000 × *g* at 4 °C.
11. Remove the ethanol and airdry the pellet at RT (*see* **Note 6**).
12. Take up the pellet in nuclease-free water (*see* **Note 7**) and proceed with β-oxidation (Subheading 3.6) or continue directly with library preparation.

3.6 β-Oxidation

Some piRNA classes are protected by a 2′*O*-methyl group at their 3′ ends. This modification will protect them from periodate treatment, also known as β-oxidation. Exposure of unprotected 3′ terminal nucleotides to $NaIO_4$ converts the ribose ring into a dialdehyde. Consequently, only 3′ ends of methylated (protected) piRNAs are available for ligation to adapters during preparation of small RNA libraries for deep-sequencing analysis. This way, certain small RNA classes can be distinguished or enriched for, such as PIWIL1/2/4-type piRNAs [13, 15]. The 26 nt long small RNA population, believed to be PIWIL3-type piRNAs in bovine oocytes [11], are not protected, just as miRNAs and siRNAs. They will therefore be depleted from downstream amplification following β-oxidation (*see* **Note 8**).

1. First, samples from Subheading 3.5 are enriched for small RNA species with the *mir*Vana kit, according to the manufacturer's protocol. Elute the RNA samples in 30 μL nuclease-free water.
2. Prepare 5× borate buffer (148 mM borax, 148 mM boric acid, pH 8.6) and 200 mM $NaIO_4$ (*see* **Note 9**).
3. For a 20 μL reaction, mix 4 μL 5× borate buffer, 2.5 μL 200 mM $NaIO_4$, and a maximum of 13.5 μL RNA sample (if a smaller volume is used, fill up the reaction to 20 μL with nuclease-free water) (*see* **Note 10**).
4. Incubate the mixture for 10 min at room temperature.
5. Add 2 μL glycerol and incubate for another minute at room temperature, in order to stop the reaction.
6. Precipitate the RNA: add 1/10th volume of 3 M NaAc (pH 5.5), 1 volume 100% isopropanol, and 1 μL glycoblue. Mix and precipitate 1 h to overnight at −80 °C.

7. Centrifuge for 30 min to 1 h at 16,000 × *g* at 4 °C and discard the supernatant.
8. Wash the pellet with 75% ethanol. Make sure the pellet detaches during the washing steps by vortexing.
9. Centrifuge again at 16,000 × *g* at 4 °C and discard the supernatant.
10. Repeat **steps 8** and **9** one to two times when the pellet looks large, irregular, and white (due to salt contaminants).
11. Once the pellet looks properly washed (you will typically see a small, slightly translucent pellet with a blue dot from the glycoblue), airdry the pellet at RT, and resuspend in nuclease-free water for downstream processing. However, particularly when the sample is used for library preparation, it is recommended to perform an additional purification step on a TBE-urea gel, especially when washing the pellet does not show much improvement (*see* **Note 11**).

4 Notes

1. The average temperature in the polystyrene box is not actively maintained and is instead kept at room temperature. Once the ovaries are removed and put in the box, the average temperature inside the box will increase to initially ~39 °C and gradually decreases to ~30 °C before it arrives in the lab (up to ~2 h). During the whole procedure, large changes in temperature (for instance by opening the box) are avoided.
2. The diameter of follicles can be measured with a ruler. The reason why follicles between 2 and 8 mm are selected is because oocytes present in follicles within this size range are mostly well developed.
3. The petri dish is marked with horizontal lines on the outside to efficiently search the area with follicular fluid.
4. The maturation medium has to be equilibrated in the incubator at 39 °C with 5% CO_2 for at least 30 min before use.
5. Lysis of your sample has to be adjusted to the type and size of the starting material. Some samples readily dissolve in TRIzol, whereas other samples are very sturdy. In Roovers et al. (2015), we also looked at human fetal ovaries from different trimesters. While first trimester ovaries were very tiny but extremely rigid, it was impossible to mash them with a pellet pestle, but they were too small to grind (which we also wanted to avoid in order to not lose too much material). In this particular case, we chose to sonicate the tissue while in TRIzol using a Diagenode Bioruptor. This was done 3× 30 s, with an interval of 30 s. Second and third trimester ovaries on the other hand are quite a bit

larger and too large to sonicate directly in TRIzol. In case of larger tissues like this, but also for instance adult ovary, it is preferred to grind the tissue under liquid N_2, collect the tissue powder in an Eppendorf tube and directly add the TRIzol.

6. If a large salt pellet is seen, which can be the result of using isopropanol for the precipitation, repeat the washing steps 1 or 2 more times with 75% ethanol. The salt contamination can interfere with downstream applications like library preparation.
7. If the sample is used for library preparation directly, it can be an option to *only* use the maximum volume that can be used in the library prep reaction, for instance, when really low amounts of RNA are expected (for instance, with limited oocytes available). It can then be preferable to use the entire sample for the library prep. For example, the samples used in Roovers et al. (2015) were resuspended in 6 μL nuclease-free water which could be used entirely as input for the NEBNext® Small RNA Library Prep Set for Illumina® [11, 16].
8. When it is unknown whether the sample contains any 2′*O*-methylated small RNAs at all, it can be considered to spike the sample with RNA that is known to contain methylated RNAs (derived from a different species or synthetic RNA). In Roovers et al. (2015), a fraction of macaque testis RNA was mixed with bovine GV oocytes prior to β-oxidation (1/10th concentration of the GV oocyte RNA). This way, a positive control was included, one that should definitely be detected in the oxidated sample, in case no protected piRNAs were present at all. PIWIL3-type piRNAs were depleted from these libraries, indicating they were not methylated. This was not due to RNA degradation during the β-oxidation procedure, but rather a specific feature of this piRNA species, since the (methylated) testis piRNAs in the same sample were still present in the sequenced libraries.
9. 5× borate buffer can be used for multiple experiments and stored at RT. It is recommended to check the pH each time before usage. The $NaIO_4$ solution should be prepared freshly.
10. For mock-treated samples, $NaIO_4$ can be preincubated with 1/10th volume of glycerol for an hour at RT, before adding the RNA. This way, ionic strength of the mock-treated samples remains comparable to the oxidated samples.
11. Running the RNA on a 15% TBE-urea gel removes final potential contamination from the β-oxidation step, which can interfere with the adapter ligation. In our hands, running samples from β-oxidation reactions on the 15% TBE-urea gel sometimes showed abnormal running patterns, possibly due to remaining contaminants. This indicates that we still experienced issues cleaning the RNA properly. Performing size selection of 15–35 nt followed by elution and precipitation of the RNA resulted in clean RNA for library preparation.

References

1. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. Cell 128:1089–1103
2. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, Plasterk RHA, Hannon GJ, Draper BW, Ketting RF (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. Cell 129:69–82
3. Kawaoka S, Hayashi N, Suzuki Y, Abe H, Sugano S, Tomari Y, Shimada T, Katsuma S (2009) The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. RNA 15(7): 1258–1264. https://doi.org/10.1261/rna.1452209
4. Huang HY, Houwing S, Kaaij LJT, Meppelink A, Redl S, Gauci S, Vos H, Draper BW, Moens CB, Burgering BM, Ladurner P, Krijgsveld J, Berezikov E, Ketting RF (2011) Tdrd1 acts as a molecular scaffold for Piwi proteins and piRNA targets in zebrafish. EMBO J 30:3298–3308
5. Cox DN, Chao A, Baker J, Chang L, Qiao D, Lin H (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev 12(23): 3715–3727. https://doi.org/10.1101/gad.12.23.3715
6. Kamminga LM, Luteijn MJ, den Broeder MJ, Redl S, Kaaij LJT, Roovers EF, Ladurner P, Berezikov E, Ketting RF (2010) Hen1 is required for oocyte development and piRNA stability in zebrafish. EMBO J 29:3688–3700
7. Deng W, Lin H (2002) Miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell 2: 819–830
8. Carmell MA, Girard A, van de Kant HJG, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ (2007) MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell 12: 503–514
9. Aravin AA, Van Der Heijden GW, Castaneda J, Vagin VV, Hannon GJ, Bortvin A (2009) Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. PLoS Genet 5: e1000764
10. Flemr M, Malik R, Franke V, Nejepinska J, Sedlacek R, Vlahoviček K, Svoboda P (2013) A retrotransposon-driven dicer isoform directs endogenous siRNA production in mouse oocytes. Cell 155:807–816
11. Roovers EF, Rosenkranz D, Mahdipour M, Han CT, He N, de Sousa Lopes SMC, van der Westerlaken LAJ, Zischler H, Butter F, Roelen BAJ, Ketting RF (2015) Piwi proteins and piRNAs in mammalian oocytes and early embryos. Cell Rep 10(12):2069–2082. https://doi.org/10.1016/j.celrep.2015.02.062
12. Tan M, van Tol HTA, Rosenkranz D, Roovers EF, Damen MJ, Stout TAE, Wu W, Roelen BAJ (2020) PIWIL3 forms a complex with TDRKH in mammalian oocytes. Cells 9(6): 1356. https://doi.org/10.3390/cells9061356
13. Ishino K, Hasuwa H, Yoshimura J, Iwasaki YW, Nishihara H, Seki NM, Hirano T, Tsuchiya M, Ishizaki H, Masuda H, Kuramoto T, Saito K, Sakakibara Y, Toyoda A, Itoh T, Siomi MC, Morishita S, Siomi H (2020) Hamster PIWI proteins bind to piRNAs with stage-specific size variations during oocyte maturation. bioRxiv. https://doi.org/10.1101/2020.12.01.407411
14. Van Tol HTA, Van Eerdenburg EFCM, Colenbrander B, Roelen BAJ (2008) Enhancement of bovine oocyte maturation by leptin is accompanied by an upregulation in mRNA expression of leptin receptor isoforms in cumulus cells. Mol Reprod Dev 75(4):578–587. https://doi.org/10.1002/mrd.20801
15. Simon B, Kirkpatrick JP, Eckhardt S, Reuter M, Rocha EA, Andrade-Navarro MA, Sehr P, Pillai RS, Carlomagno T (2011) Recognition of 2′-o-methylated 3′-end of piRNA by the PAZ domain of a Piwi protein. Structure 19(2): 172–180. https://doi.org/10.1016/j.str.2010.11.015
16. Rosenkranz D, Han C-T, Roovers EF, Zischler H, Ketting RF (2015) Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. Genomics Data 5:309–313

Chapter 6

3D Imaging and In Situ Hybridization for Uncovering the Functions of MicroRNA in Rice Anther

Koji Koizumi and Reina Komiya

Abstract

Small RNAs specifically expressed in reproductive tissues are key regulators of germline development in eukaryotes. Rice microRNA2118 (miR2118), which is enriched during reproduction in grasses, is a trigger to produce phased small interfering RNAs (phasiRNAs). These phasiRNAs demonstrate the temporal regulation with premeiotic phasiRNAs and meiotic phasiRNAs in anther development. Furthermore, the site-specific regulation via miR2118 and phasiRNAs is of importance in soma and germ development in anthers. Accordingly, histological imaging methods are essential tools for understanding spatiotemporal regulation during reproduction and elucidating the reproductive roles of miRNAs and phasiRNAs. We successfully developed a method to visualize the three-dimensional (3D) structure of entire rice anthers, which can also be used for distinguishing the internal structure of the anthers in other plants. Here, we describe the detailed methods of in situ hybridization for miR2118 localization and the visualization of the 3D structure of entire anthers of rice.

Key words MicroRNA, PhasiRNA, in situ hybridization, 3D imaging, Anther, Rice

1 Introduction

MicroRNA (miRNA) is a core component of the silencing system that significantly contributes to developmental regulation in many organisms. The functions of plant miRNAs can be divided into two main groups: (1) posttranscriptional repression and (2) induction of secondary small interfering RNA production [1, 2]. miR2118, the latter type of miRNA, acts as a trigger for the production of a large number of 21-nucleotide (nt) phasiRNAs in grasses [3]. During this process, 1300-2000 types of long non-coding RNAs with miR2118-recognition motifs are specifically expressed at the pre-meiotic stages in anther development. miR2118

Supplementary Information The online version contains supplementary material available at [https://doi.org/10.1007/978-1-0716-2380-0_6].

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_6,

recognizes these long non-coding RNAs, and its associated AGO cleaves them. This cleavage triggers the synthesis of double-stranded RNAs (dsRNAs) via RNA-dependent RNA polymerase 6. Next, Dicer-like 4 (DCL4) protein processes the dsRNAs into 21-nt phasiRNAs [4–6]. When proceeding to meiosis, 24-nt-long phasiRNAs are generated via microRNA2275 (miR2275) cleavage and DCL3b/DCL5 processing [7, 8]. Therefore, temporally regulated expression of pre-meiotic 21-nt-long phasiRNAs and meiotic 24-nt-long phasiRNAs could be required for anther development.

In addition to temporal regulation, spatial phasiRNA expression patterns are important for development. Anthers contain germ cells and somatic cell layers arranged in four types of anther wall layers. The development of somatic anther wall synchronizes with germ development. Therefore, defects in anther wall development cause pollen sterility [9, 10]. The miR2118 family members are enriched in the outer layers, called epidermis, in pre-meiotic anthers [7, 11, 12]. Furthermore, rice miR2118 regulates the somatic anther wall development via uracil (U)-rich phasiRNA production [12]. In contrast to the soma phasiRNAs, the germ cell-specific AGO, named MEIOSIS ARRESTED AT LEPTOTENE1 (MEL1), interacts with the first cytosine (C) of phasiRNAs [5]. The U-rich phasiRNAs in the miR2118-dependent soma are distinct from the C-phasiRNAs interacting with germ-specific MEL1, suggesting that the combination of site-specific phasiRNAs is involved in interactions between soma and germ development in anthers [12, 13]. Recently, it was reported that 21-nt germ-phasiRNAs cause the cleavage of target RNAs [14, 15]. However, the mechanism of action of phasiRNAs in the reproductive system, including soma regulation, is not yet available.

In animals, it has been reported that site-specific AGO-small RNA complexes, well-known as PIWI-piRNA complexes, are essential for germline development via transposable element silencing to preserve genome integrity [16, 17]. Thus, spatiotemporal histological analyses are required to understand reproductive development in both plants and animals. In this chapter, we describe in detail the methods of in situ hybridization for miR2118 localization using the cross-section of anthers at pre-meiotic stages. This method is applicable to other specific organs and for mRNA or miRNA localization. Moreover, to histologically clarify the reproductive development, a 3D imaging method using Lightsheet microscopy can serve as a practical tool to visualize the features from the cell level up to the whole structure of the anthers.

2 Materials

RNase-free water or distilled MilliQ water (DW) is used to prepare all solutions, and wearing gloves is recommended to prevent RNase contamination.

2.1 Reagents for In situ Hybridization

1. Formalin acetic acid alcohol (FAA) fixative: 1.85% formaldehyde, 5% acetic acid, and 30% ethanol.
2. Dehydration buffer I: 50% ethanol and 10% *t*-butanol.
3. Dehydration buffer II: 50% ethanol and 20% *t*-butanol.
4. Dehydration buffer III: 50% ethanol and 30% *t*-butanol.
5. Dehydration buffer IV: 40% ethanol and 50% *t*-butanol.
6. Dehydration buffer V: 25% ethanol, 75% *t*-butanol, and 1% neutral red.
7. Paraplast Plus (Merck/Sigma-Aldrich).
8. Toluidine blue solution: 0.1% (w/v) toluidine blue (Nacalai).
9. Lemosol (FUJIFILM Wako Pure Chemical Corporation).
10. Proteinase K Buffer: 100 mM Tris-HCl (pH 8.0) and 10 mM ethylenediaminetetraacetic acid (EDTA).
11. Proteinase K solution: Proteinase K recombinant PCR Grade (Sigma-Aldrich).
12. 10× HCMF: 80 g NaCl, 4 g KCl, 1.2 g $Na_2HPO_4{\cdot}(12H_2O)$, 24 g HEPES, 1.92 g NaOH, and 10 g glucose in 1 L distilled MilliQ water (DW). After sterilization, 10 mg phenol red (pH 7.4) is added.
13. Refixative solution: 4% PFA, 2.5 mL 1 M NaOH, and 2.5 mL 1 M HCl in 250 mL 1× HCMF.
14. Acetylation solution: 0.5 mL 12 N HCl, 3 mL triethanolamine, and 0.5 mL acetic acid in 200 mL DW.
15. Prehybridization solution: 50% (v/v) formamide, 1× Denhardt's solution, 2× SSC, 10 mM EDTA, 100 μg/mL yeast tRNA (Sigma-Aldrich), and 0.01% Tween 20.
16. Hybridization solution: 50% (v/v) formamide, 1× Denhardt's solution, 2× SSC, 10 mM EDTA, 100 μg/mL yeast tRNA, 0.01% Tween 20, and 5% dextran sulfate (Sigma-Aldrich; 20% dextran sulfate was stored at 4 °C).
17. 50× Denhardt's solution: 1% Ficoll (type 400), 1% polyvinylpyrrolidone, and 1% bovine serum albumin (filtered 50× Denhardt's solution can be stored at 20 °C).
18. Humidified solution: 2× SSC and 50% (v/v) formamide.
19. Wash buffer: 50% (v/v) formamide, 2× SSC, and 0.01% Tween 20.
20. RNase A buffer: 500 mM NaCl, 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, and 0.01% Tween 20.
21. RNase A stock: 50 mg/mL (Nippon Gene).
22. 2× SSC wash buffer: 2× SSC and 0.01% Tween 20.
23. 0.2× SSC wash buffer: 0.2× SSC and 0.01% Tween 20.

24. 0.2 N HCl: 3.33 mL 12 N HCl in 200 mL DW.
25. TBST: 1× Tris-buffered saline (TBS) and 0.01% Tween 20.
26. 1× Blocking reagent: dilution of 10× Blocking reagent (Roche; 10× Blocking reagent is prepared according to the manufacturer's instructions and stored at -20 °C) with TBST.
27. 10× Detection buffer 1: 1 M Tris-HCl (pH 7.5) and 1.5 M NaCl.
28. 10× Detection buffer 3: 0.5 M Tris-HCl (pH 9.5) and 0.5 M NaCl.
29. 1 M $MgCl_2 \cdot 6H_2O$.
30. miR2118 probe: osa-miR-2118f modified with LNA and digoxigenin at the 3′- and 5′-terminus, 5DiGN/TAGGAA-TGGGAGGCATCAGGAA/3DiGN (Exiqon, #618800-360).
31. Scramble-miR miRCURY LNA DETEC: the probe for the negative control (Qiagen; #339111 YD00699004-BCG; 5DiGN/GTGTAACACGTCTATACGCCCA/3DiGN).
32. Glycine (FUJIFILM Wako Pure Chemical Corporation).
33. Anti-digoxigenin-AP (Roche).
34. Alkaline phosphatase substrate solution: 50.7 μL NBT and 52.5 μL BCIP (Roche) in 15 mL detection buffer 3. The solution is prepared according to the manufacturer's instructions.
35. Diethylpyrocarbonate-H_2O: DEPC treated water (Nippon Gene).
36. Slide Glass: MAS-01 Adhesive Glass slide (Matsunami).

2.2 Reagents for 3D Imaging of Anthers

1. 5× PMEG stock buffer (pH 6.8): 250 mM PIPES (Dojindo Molecular Technologies), 25 mM EGTA, 25 mM $MgSO_4 \cdot 7H_2O$, and 20% glycerol (stock at 4 °C); 10 mL DMSO was added to 1000 mL 5× PMEG stock buffer before use (prepared at time of use).
2. PFA fixative (prepared at time of use): 4% PFA (Alfa Aesar) in 1× PMEG buffer.
3. SR2200 solution: 1% (v/v) SCRI Renaissance 2200 (Renaissance Chemicals) in 1× PBS.
4. RapiClear 1.52: RapiClear 1.52 (refractive index = 1.52) (SunJin Lab).
5. 1% agarose: 1% (w/v) low-melting agarose (Sigma-Aldrich) in 1× PBS.
6. Glass capillary: Glass capillary 10 μL (BRAND).
7. Pulling plunger: Pulling a plunger for 10 μL capillary (BRAND).

3 Methods

3.1 In Situ Hybridization Using the miR2118 Probe

In situ hybridization for miR2118 expression consists of the following six steps: (1) fixation, (2) embedding and sectioning, (3) proteinase K treatment, (4) hybridization, (5) washing, and (6) detection. We have optimized this protocol for imaging rice inflorescences; it may be possible to modify these basic steps for other plant tissues.

3.1.1 Fixation

1. Place the inflorescences in FAA fixative on ice and degas them at 0.1 MPa for 2 min.
2. Repeat the degassing three more times and incubate the inflorescences for 1 h at 4 °C.
3. Replace the FAA fixative and incubate the inflorescences overnight at 4 °C.
4. Dehydrate the samples in dehydration buffer I for 1 h at 4 °C with shaking.
5. Repeat the dehydration step using dehydration buffer II to dehydration buffer IV for 1 h each at 4 °C with shaking.
6. Dehydrate the samples using dehydration buffer V overnight at room temperature with shaking.
7. Remove dehydration buffer V, add *t*-butanol and incubate the mixture for 3-5 h at room temperature with shaking.
8. Replace *t*-butanol repeatedly and incubate overnight at room temperature.
9. Remove *t*-butanol, add the solution of 80% *t*-butanol and 20% chloroform, and incubate the sample for 1 h in a fume hood.

3.1.2 Embedding and Sectioning

1. Add Paraplast Plus to the sample and incubate the sample overnight at 60 °C (*see* **Note 1**).
2. Replace Paraplast Plus with new Paraplast Plus, and incubate the sample overnight at 60 °C.
3. Repeat **step 2** for a few days.
4. Embed the samples in Paraplast Plus.
5. Trim the wax block and mount it onto the microtome block.
6. Section the tissues to a thickness of 8 μm (*see* **Note 2**).
7. Make 8-cm serial sections (wax ribbons) and place them on a black paper (Fig. 1a).
8. Transfer a portion (1-cm) of the 8-cm wax ribbons to the top of water on a slide using a pair of forceps.
9. Stain the 1-cm wax ribbons using toluidine blue solution and wash the sections three times in DW.

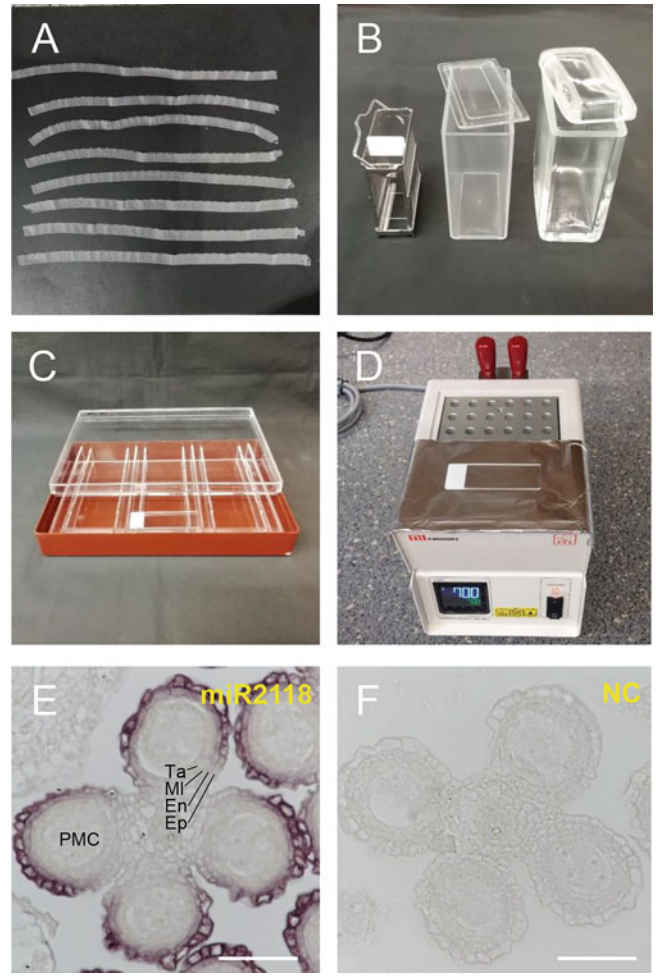


Fig. 1 Equipment used for in situ hybridization and images of miR2118 localization in rice anthers. (**a**) Paraffin-embedded tissue serial sections. The central parts of the anthers were selected using a part of the staining section. (**b**) Slide rack (left), plastic staining jar (center) for treatment and washing, and glass staining jar (right) for lemosol treatment. (**c**) Humidified box for probe hybridization. (**d**) Heat block incubator with aluminum foil for hybridization. (**e**, **f**) In situ hybridization of miR2118 using a miRCURY LNA miRNA detection probe and negative control (NC) using anthers (cross sectioning) of 2.0-2.5-mm inflorescences. The cross-sections of anthers show the somatic anther walls consisting of the epidermis (Ep), endothecium (En), middle layer (Ml), and tapetum layer (Ta) and the germ cells named pollen mother cells (PMC). The signal was observed in the epidermal and endothecium layers. Scale bar = 50 μm

10. Investigate the sample stage and inspect the section of interest under a stereo microscope.
11. Pick the 2-cm-long wax ribbons (sections of interest) in the rest of the 8-cm wax ribbons and transfer two or three 2-cm

ribbons to the top of water (diethylpyrocarbonate-H_2O) on a slide.

12. Remove the water using a PIPETMAN pipette and absorb the excess water using filter paper. Incubate the slide at 42 °C on a slide warmer overnight to dry (*see* **Note 3**).

3.1.3 Proteinase K Treatment of Samples

1. Place the slides in lemosol for 10 min and repeat this step with fresh lemosol (*see* **Note 4**; Fig. 1b).
2. Immerse the slides in 50% lemosol and 50% ethanol mixture for 5 min.
3. Immerse the slides in 99.5% ethanol for 1 min and repeat this step.
4. Hydrate the slides in 90% ethanol for 2 min.
5. Hydrate the slides in 70% ethanol for 2 min.
6. Hydrate the slides in 50% ethanol for 2 min.
7. Hydrate the slides in 30% ethanol for 2 min.
8. Wash the slides in DW for 2 min and repeat this step.
9. Treat the slides with 0.2 N HCl for 20 min.
10. Wash the slides in DW for 5 min.
11. Incubate the slides in prewarmed proteinase K buffer at 37 °C for 2 min.
12. Treat the slides with 0.5 μg/mL proteinase K in the proteinase K buffer at 37 °C for 15 min (*see* **Note 5**).
13. Transfer the slides to 1× PBS containing 0.2% (w/v) glycine for 10 min.
14. Wash the slides in 1× PBS for 5 min.
15. Refix the slides in refixative solution for 20 min.
16. Wash the slides in 1× PBS for 5 min and repeat the washing step.
17. Incubate the slides in acetylation solution for 15 min.
18. Wash the slides in DW for 5 min.
19. Dehydrate the slides in 30% ethanol for 2 min.
20. Dehydrate the slides in 50% ethanol for 2 min.
21. Dehydrate the slides in 70% ethanol for 2 min.
22. Dehydrate the slides in 90% ethanol for 2 min.
23. Dehydrate the slides in 99.5% ethanol for 5 min and repeat this step with 99.5% ethanol.
24. Dry the slides for 30 min in a vacuum desiccator (decreased to 0.07 MPa).

3.1.4 Hybridization

1. Place the humidified solution in a humidified box (Fig. 1c) prewarmed to 55 °C (*see* **Note 6**).
2. Add 100 μL of prewarmed prehybridization solution to each slide and incubate the slide at 55 °C in humidified boxes.
3. Dilute the miRNA probe with hybridization buffer (*see* **Note 7**).
4. Denature the probes at 80 °C for 10 min and place the tube on ice immediately after denaturation.
5. Remove the prehybridization solution from the slides.
6. Place the slides on a 70 °C heat block incubator (Fig. 1d) and apply the probe solution.
7. Place a coverslip on the slides (*see* **Note 8**).
8. Place the slides in the humid boxes, which were then sealed with adhesive tape.
9. Incubate the box overnight (for more than 16 h) at 55 °C.

3.1.5 Post-hybridization (Washing)

1. Remove the coverslip and place the slides in a slide rack (Fig. 1b).
2. Wash the slides in wash buffer at 55 °C for 30 min and repeat this washing step.
3. Transfer the slides to RNase A buffer and incubate them at 37 °C for 5 min.
4. Incubate the slides in RNase A buffer with 10 μg/mL RNase A at 37 °C for 30 min to 1 h (*see* **Note 9**).
5. Transfer the slides to RNase A buffer and incubate at 37 °C for 5 min.
6. Wash the slides in 2× SSC buffer at 55 °C for 30 min.
7. Wash the slides in 0.2× SSC buffer at 55 °C for 30 min.

3.1.6 Probe Detection

The following steps are performed at room temperature:

1. Transfer the slides to detection buffer 1 at room temperature for 5 min and repeat this step.
2. Add 500 μL of 1× Blocking reagent per slide and incubate the slides for 30 min.
3. Wash 1× Blocking Reagent with detection buffer 1.
4. Add 500 μL anti-digoxigenin-AP per slide and incubate the slides for 1 h.
5. Wash the slides in detection buffer 1, incubate it in detection buffer 1 for 10 min, and repeat this washing step twice. Agitate the slides every few minutes during washing.
6. Incubate the slides in detection buffer 3 with 50 mM $MgCl_2$.
7. Add 800 μL alkaline phosphatase substrate solution per slide.

8. Place the coverslip on top of the sample and transfer the slide to a humid box.
9. Incubate the humid box in darkness at 30 °C for 3 h to 2 days (*see* **Note 10**).
10. Remove the coverslip and wash the alkaline phosphatase substrate solution with Tris–EDTA buffer.
11. Mount the slides in 50% glycerol solution or an aqueous mounting medium, and seal the coverslip with nail polish.
12. Detect the signals under a light microscope such as Nikon Eclipse NiE (Fig. 1e and f).

3.2 Visualization of the 3D Structure of the Entire Anther

1. Place the inflorescences in PFA fixative on ice and degas at 0.09-0.1 MPa for 20 min.
2. Repeat degassing three more times.
3. Incubate the samples for 100 min at room temperature with shaking.
4. Wash the samples in 1× PMEG buffer for 20 min at room temperature with shaking.
5. Repeat the washing step five more times (*see* **Note 11**).
6. Cut the anthers from caryopsis in 1× PBS buffer under a stereo microscope.
7. Stain the samples with SR2200 solution (cell wall staining dye) for 2 h (*see* **Note 12**).
8. Wash the sample four times in 1× PBS.
9. Clear the sample using RapiClear 1.52 for 1 or 2 days (*see* **Note 13**).
10. Place the sample in a 35-mm dish on a cup warmer containing 1% (w/v) melted agarose (Fig. 2a).
11. Embed the anther in a glass capillary by pulling a plunger (Fig. 2b; *see* **Note 14**).
12. Capture the images using a light sheet illumination microscope (Lightsheet Z.1, Carl Zeiss) (Fig. 2c; *see* **Note 15**).

    Conditions: 20× (NA 1.0) Plan Apochromat lens for detection, 10× (NA 0.2) LSFM clearing lens for illumination, 405 and 561 nm laser lines for SR2200 and autofluorescence excitation, 420-470 nm (SR2200) and 585 nm long pass (autofluorescence) filter emission. Glass capillary and gel mounting, 1× PBS in the sample chamber, dual-side illumination (*see* **Note 16**).
13. Create images and animation using ZEN (Carl Zeiss) or Imaris 9 (Bitplane AG) software (Fig. 2d–i; Electronic Supplementary Movie 1).

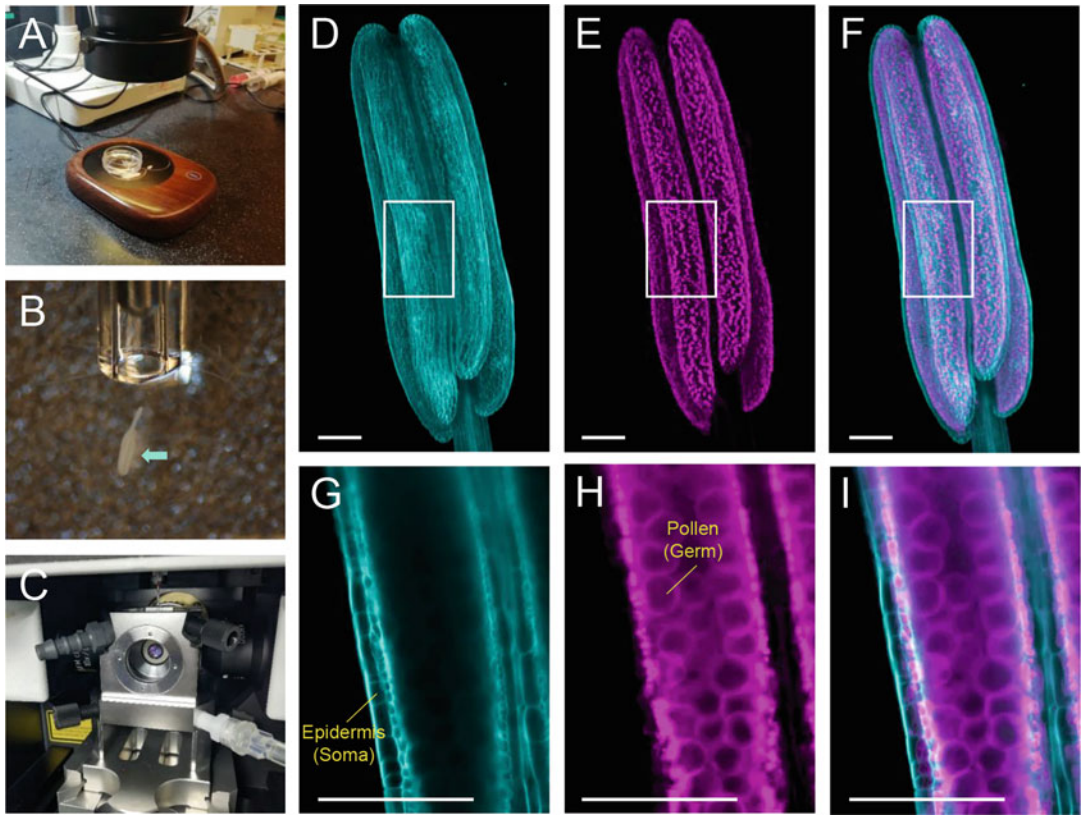


Fig. 2 Imaging of rice anthers using Lightsheet microscope. (**a**) A 35-mm dish on the cup warmer under a stereo microscope. (**b**) Rice anther in melted agarose and glass capillary (inner diameter: 0.7 mm). The anther was pulled into the capillary by pulling a plunger. The light blue arrow shows the anther. (**c**) Imaging of the anther embedded with 1% agarose gel in a glass capillary using the Lightsheet microscope. (**d**–**f**) 3D maximum intensity projection of the 1 mm rice anther. (**d**, **g**) Anthers stained with SR2200. Fluorescence was strongly observed in somatic anther walls, specifically in the epidermis. (**e**, **h**) Autofluorescence images of anthers. The pollens (germ cells) were detected (**h**). (**g**–**i**) Slice image of the anther before 3D reconstruction of the **d**–**f** images. Laser excitation/emission: 405 nm/420–470 nm (**d**), 561 nm/575 nm LP (**e**). (**f**, **i**) Merged image of **d** and **e**, and **g** and **h**, respectively. Scale bar = 100 μm

4 Notes

1. Paraplast is incubated and melted at 60 °C before use.
2. The angle of the blade is adjusted to 4-6°. The caryopsis should be cut vertically by adjusting the sample block folder in the microtome. The wax ribbon should be picked using a pair of forceps.
3. The dried slides can be stored in slide boxes at 4 °C for several days.

4. We used glass staining jars (Fig. 1b, right, AS ONE) and slide rack (Fig. 1b, left, AS ONE) for lemosol treatment of samples (Subheading 3.1.3, **steps 1** and **2**). We used plastic staining jars (Fig. 1b, center, AS ONE) and slide rack for proteinase K treatment of samples (Subheading 3.1.3, **steps 3-24**), post-hybridization Subheading 3.1.5, and probe detection Subheading 3.1.6, **step 1**, **5**, and **6**.
5. Proteinase K treatment is critical for successful in situ hybridization. This enzyme partially digests tissues to enable better probe penetration. It is recommended to change the concentration and incubation time if the image shows high background (insufficient digestion) or tissue damage (overdigestion).
6. A humidified box (Fig. 1c) was used for hybridization (Subheading 3.1.4) and probe detection (Subheading 3.1.6, **steps 2-4** and **7-9**).
7. An miRNA probe (25 nM) was used for the anther cross-section. A pretest of probe concentration is recommended when the expression of other miRNAs or genes is investigated for the first time.
8. A 24 mm × 60 mm coverslip is useful.
9. RNase A stock solution was added to RNase A buffer before incubation.
10. Signal for miR2118 was detected after 3 h.
11. Sample stock can be stored at 4 °C for 6 months.
12. Protocols for combined fluorescence dyes with a clearing solution were prepared as previously reported [18] (the conditions of double-staining methods has also been reported).
13. In the method reported by Ursache et al., the sample was cleared before staining, and the sample was stained with the dye dissolved in clearing solution.
14. Several anthers can be mounted in one capillary.
15. Herein, we provide a microscope setting for observing the sample in water. Microscope settings for clearing tissue in Carl Zeiss Lightsheet Z.1 or Lightsheet 7 may deliver better results.
16. Multiview acquisition and image fusion were performed using Carl Zeiss Z.1 and ZEN SP1 operation software. Image acquisition and fusion from two angles (180° in opposing directions) enable the acquisition of clear images from the back side.

Acknowledgments

This work was supported by JST PRESTO Program (Grant number JPMJPR17Q3, Japan), JST FORESTO Program (Grant Number JPMJFR204U, Japan), KAKENHI Programs (Grant numbers JP17H05608 and JP15H01476), the Naito Foundation, and the Okinawa Institute of Science and Technology Graduate University, Japan, to R.K. We thank Ms. Saori Araki and Ms. Hinako Tamotsu for rice growth and sampling.

References

1. Iwakawa HO, Tomari Y (2015) The functions of MicroRNAs: mRNA decay and translational repression. Trends Cell Biol 25(11):651–665
2. Komiya R (2017) Biogenesis of diverse plant phasiRNAs involves an miRNA-trigger and dicer-processing. J Plant Res 130(1):17–23
3. Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan GL, Walbot V, Sundaresan V, Vance V, Bowman LH (2009) Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. Genome Res 19(8):1429–1440
4. Song X, Li P, Zhai J, Zhou M, Ma L, Liu B, Jeong DH, Nakano M, Cao S, Liu C et al (2012) Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J 69(3): 462–474
5. Komiya R, Ohyanagi H, Niihama M, Watanabe T, Nakano M, Kurata N, Nonomura K (2014) Rice germline-specific Argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. Plant J 78(3): 385–397
6. Song X, Wang D, Ma L, Chen Z, Li P, Cui X, Liu C, Cao S, Chu C, Tao Y et al (2012) Rice RNA-dependent RNA polymerase 6 acts in small RNA biogenesis and spikelet development. Plant J 71(3):378–389
7. Zhai J, Zhang H, Arikit S, Huang K, Nan GL, Walbot V, Meyers BC (2015) Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc Natl Acad Sci U S A 112(10):3146–3151
8. Teng C, Zhang H, Hammond R, Huang K, Meyers BC, Walbot V (2020) *Dicer-like* 5 deficiency confers temperature-sensitive male sterility in maize. Nat Commun 11(1):2912
9. Nonomura K, Miyoshi K, Eiguchi M, Suzuki T, Miyao A, Hirochika H, Kurata N (2003) The *MSP1* gene is necessary to restrict the number of cells entering into male and female sporogenesis and to initiate anther wall formation in rice. Plant Cell 15(8):1728–1739
10. Fu Z, Yu J, Cheng X, Zong X, Xu J, Chen M, Li Z, Zhang D, Liang W (2014) The Rice basic helix-loop-helix transcription factor TDR INTERACTING PROTEIN2 is a central switch in early anther development. Plant Cell 26(4):1512–1524
11. Ta KN, Sabot F, Adam H, Vigouroux Y, De Mita S, Ghesquiere A, Do NV, Gantet P, Jouannic S (2016) miR2118-triggered phased siRNAs are differentially expressed during the panicle development of wild and domesticated African rice species. Rice 9(1):10
12. Araki S, Le NT, Koizumi K, Villar-Briones A, Nonomura KI, Endo M, Inoue H, Saze H, Komiya R (2020) miR2118-dependent U-rich phasiRNA production in rice anther wall development. Nat Commun 11(1):3115
13. Komiya R (2021) Spatiotemporal regulation and roles of reproductive phasiRNAs in plants. Genes Genet Syst 96(5):209–215
14. Jiang P, Lian B, Liu C, Fu Z, Shen Y, Cheng Z, Qi Y (2020) 21-nt phasiRNAs direct target mRNA cleavage in rice male germ cells. Nat Commun 11(1):5191
15. Zhang YC, Lei MQ, Zhou YF, Yang YW, Lian JP, Yu Y, Feng YZ, Zhou KR, He RR, He H et al (2020) Reproductive phasiRNAs regulate reprogramming of gene expression and meiotic progression in rice. Nat Commun 11(1):6031
16. Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136(4): 656–668
17. Sato K, Siomi MC (2018) Two distinct transcriptional controls triggered by nuclear PiwiRISCs in the *drosophila* piRNA pathway. Curr Opin Struct Biol 53:69–76
18. Ursache R, Andersen TG, Marhavý P, Geldner N (2018) A protocol for combining fluorescent proteins with histological stains for diverse cell wall components. Plant J 93(2):399–412

Part II

Methods to Study Roles of piRNAs in Classic Model Organisms

Chapter 7

Cloning, Sequencing, and Linkage Analysis of piRNAs

Rippei Hayashi

Abstract

Piwi-interacting RNAs (piRNAs) are 25- to 32-nucleotide-long small RNAs that silence transposable elements (transposons) in animal gonads. piRNAs have a large sequence diversity (over one million different sequences per organism) to target a variety of transposon sequences. This is achieved by flexible and distinct biogenesis pathways that are evolutionarily conserved. In this chapter, I describe a detailed method of purifying and cloning piRNAs from freshly dissected tissue samples, such as fruit fly ovaries, for the high-throughput sequencing. I also describe how to computationally process the sequencing data and interrogate the characteristic pattern of piRNA biogenesis, including ping-pong amplification and head-to-tail phasing.

Key words piRNA biogenesis, Small RNA library, Phasing, Ping-pong amplification

1 Introduction

There are three major classes of small RNAs in animals; small interfering RNA (siRNA), micro RNA (miRNA), and piRNA [1]. siRNAs and miRNAs are made by double-stranded RNA-dependent ribonuclease III enzymes Drosha and Dicer while piRNAs are made from single-stranded RNA [2]. The 5′ end of a piRNA is first generated by an endonuclease [3]. After the 5′ end of a piRNA precursor RNA is loaded onto an Argonaute protein, the 3′ end is cleaved by another endonuclease, which can be further trimmed by exonucleases [4].

There are two distinct piRNA biogenesis pathways called ping-pong and phasing (Fig. 1). The major difference of the two pathways is that the 5′ ends of ping-pong piRNA are defined by the slicing of piRNA-loaded Argonaute proteins while the endonuclease Zucchini/MitoPLD makes the 5′ end of phasing piRNAs. Because piRNA itself is used for the endonucleolytic cleavage, ping-pong amplifies a pair of piRNAs from opposing strands of double-stranded RNA, whose 5′ ends overlap by ten nucleotides [5, 6]. In contrast, phasing sequentially produces piRNAs from the same

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_7,

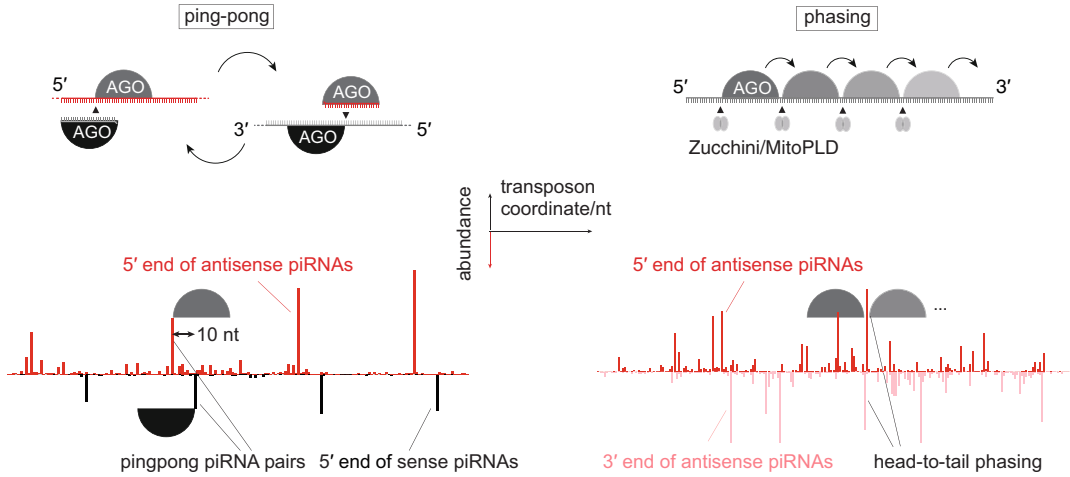


Fig. 1 Ping-pong and phasing piRNA biogenesis pathways. To the top left, "ping-pong" is initiated by a cleavage of a single-stranded RNA (colored in red) by a piRNA-carrying Argonaute protein (colored in black). The cleavage (indicated by arrowhead) makes the piRNA from the cleaved RNA, and the resulting piRNA-Argonaute complex in turn targets the opposite strand (colored in gray) to make another piRNA. The loop goes on to self-amplify the pair of piRNAs from opposite strands that overlap by ten nucleotides at 5′ ends. To the top right, the endonucleolytic cleavage by Zucchini/MitoPLD (depicted in gray ovals) creates "phasing" piRNAs. Zucchini/MitoPLD cleaves RNA sequentially from 5′ to 3′ in such a way that it leaves one piRNA at each step of the cleavage. As a result, piRNAs are lined up in a head-to-tail arrangement where the 3′ end of one piRNA is followed by the 5′ end of the next. To the bottom, 5′ and 3′ ends of fruit fly ovarian piRNAs mapping to *F-element* (left) and *gypsy* (right) transposon are shown, highlighting the ping-pong and phasing biogenesis, respectively

RNA strand, where a single cleavage of Zucchini/MitoPLD simultaneously determines the 3′ end of one piRNA and the 5′ end of the next [7, 8].

Because endo- and exonucleases involved in piRNA biogenesis do not depend on specific sequence motifs or RNA structures, the piRNA population is extremely diverse. High-throughput sequencing has been commonly used to identify piRNAs and reveal the mechanisms of piRNA production [9, 10]. In this chapter, I describe the method to clone piRNAs from fruit fly ovaries for Illumina deep sequencing, and the computational pipeline to examine the biogenesis pathways. The method can be applied to study piRNAs from any other tissues and organisms.

2 Materials

2.1 Preparation of Total RNA from Fly Ovaries

1. TriZol (Thermo Fisher) or alternatively, TRI Reagent (Sigma).
2. Acid-Phenol:Chloroform, pH 4.5 (Thermo Fisher).
3. 2 U/μl DNase I (RNase-free) (New England BioLabs).
4. 20 U/μl SUPERase IN RNase Inhibitor (Thermo Fisher).

5. 20 mg/ml Glycogen, molecular biology grade (Thermo Fisher).
6. Pellet pestles blue polypropylene (Sigma).

2.2 Small RNA Cloning

1. 2S ribosomal RNA (rRNA) capture oligo: /5Biotin-TEG/TA CAACCCTCAACCATATGTAGTCCAAGCA.
2. 3′ linker (DNA oligo, 1+26 nt): /5rApp/ NNNNAGATCG GAAGAGCACACGTCT/3ddC/.
3. 5′ linker (RNA oligo, 37 nt): ACACUCUUUCCCUACAC GACGCUCUUCCGAUCUNNNN.
4. 19 mer RNA spike (*see* **Note 1**): CGUACGCGGGUUUAAA CGA.
5. 35 mer RNA spike (*see* **Note 1**): CUCAUCUUGGUCG UACGCGGAAUAGUUUAAACUGU.
6. Solexa_RT_rev: GTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT.
7. Solexa_PCR_fw: AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCCGATCT.
8. TruSeq index primers (xxxxxx: reverse complement to the i7 6 mer barcode): CAAGCAGAAGACGGCATACGAGA TxxxxxxGTGACTGGAGTTCAGACGTGTGCTCTTCCGA TCT.
9. Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher).
10. 20× SSC buffer: 3 M sodium chloride, 300 mM sodium citrate, pH 7.0.
11. 6000 Ci/mmol [γ-32P] ATP (PerkinElmer).
12. 10 U/μl T4 polynucleotide kinase (New England BioLabs).
13. T4 RNA Ligase 2, truncated K227Q (New England BioLabs).
14. T4 RNA Ligase 1 (New England BioLabs).
15. 40% Acrylamide/Bis 29:1 (v/v) solution (BioRad).
16. Micro Bio-spin P-6, Tris buffer (BioRad).
17. Storage Phosphor Screen, Multipurpose Standard (Cytiva).
18. Centrifugal Filter Cellulose Acetate 2 ml 0.45 μm MS (MicroAnalytix).
19. Borax Anhydrous.
20. Boric Acid.
21. Sodium periodate.
22. AMPure XP (Beckman Coulter).
23. 200 U/μl SuperScript II Reverse Transcriptase (Thermo Fisher).
24. KAPA LongRange HotStart polymerase (Sigma).

25. EvaGreen Dye, 20× in Water (Biotium).
26. Reference Dye for Quantitative PCR (ROX), 100× (Sigma).
27. Agarose for molecular biology.
28. 1 kb Plus DNA Ladder.
29. 5 U/μl PmeI (New England BioLabs).
30. Agarose, low gelling temperature (Sigma).
31. Zymoclean gel DNA recovery kit (Integrated Science).
32. Qubit DNA high-sensitivity assay kit (Thermo Fisher).
33. Qubit Assay tubes (Thermo Fisher).

2.3 Equipment

1. NanoDrop One/One (Thermo Fisher).
2. Pellet pestles Cordless motor (Sigma).
3. DynaMag Magnet (Thermo Fisher).
4. Mini-PROTEAN gel casting stand, electrophoresis chamber (BioRad).
5. Mini-PROTEAN Short, flat and spacer plates (BioRad).
6. Mini-PROTEIN comb, 10-well, 1.0 mm (BioRad).
7. Typhoon phosphor imager, FLA 9000.
8. QuantStudio 12 k Flex Realtime PCR or equivalent for the quantitative PCR.
9. Standard thermal cycler for PCR.
10. Qubit 4 Fluorometer.

2.4 Software for the Computational Analyses

1. fastx_toolkit-0.0.14.
2. bowtie-1.2.3-linux-x86_64.
3. samtools/1.10.
4. bedtools/2.28.0.
5. weblogo/3.7.8.
6. R/4.0.2.
7. reshape2_1.4.4.

3 Methods

3.1 Preparation of Small RNA Libraries for Illumina Sequencing

The cloning of small RNA involves a sequential adapter ligation of size-selected pool of RNA, which was initially developed to clone miRNAs [11]. The procedure is summarized in Fig. 2. We use PAGE gel to purify small RNAs and remove unligated linker oligos because it is most reliable in our hands among other methods for the size-selection of RNA. We use a DNA-based 3′ linker oligo with modifications at both ends to achieve a directional ligation, which is

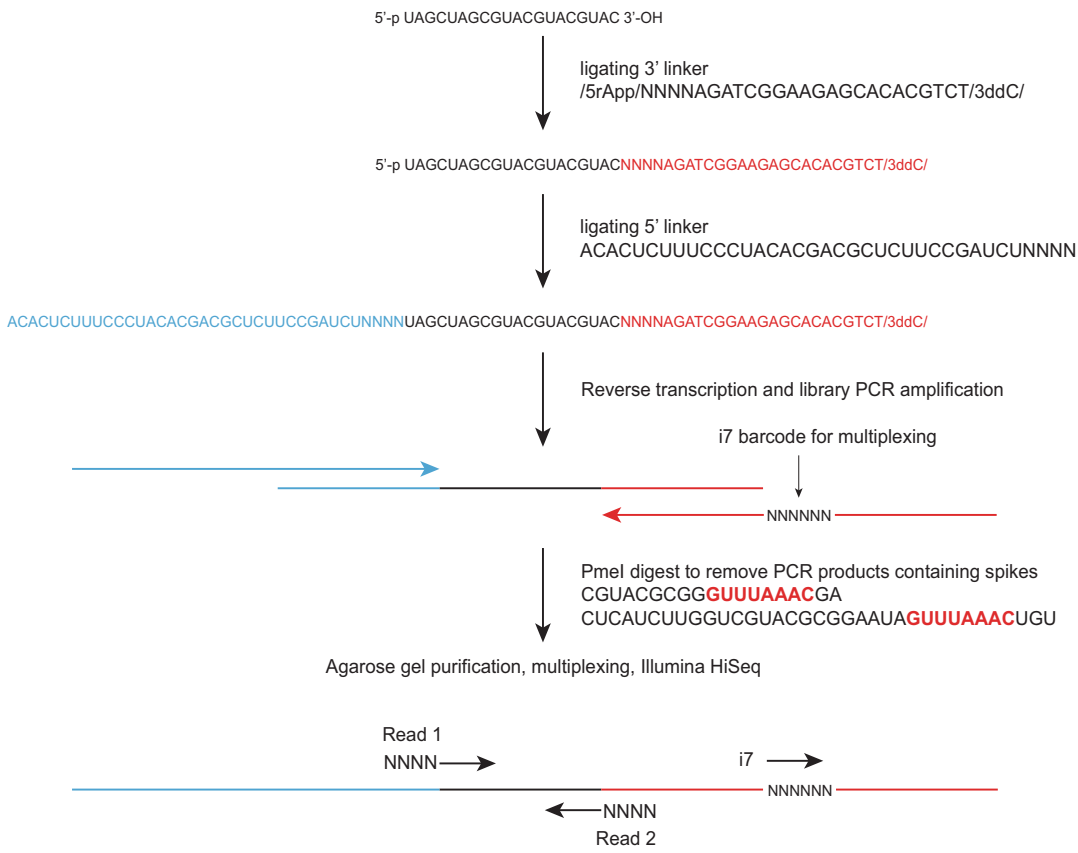


Fig. 2 The small RNA cloning workflow is summarized. The small RNA (a sequence is given for example) with 5′ monophosphate and 3′ hydroxyl groups is first ligated to a DNA oligo linker at the 3′ end (colored in red). This is followed by a ligation of an RNA oligo linker (colored in cyan) at the 5′ end. These 3′ and 5′ linkers carry four random nucleotides (NNNN) at the ends to allow a uniform ligation to the variety of sequences in small RNAs. The ligation products are subsequently reverse-transcribed and PCR-amplified to generate the library for Illumina sequencing. Reads from the paired end sequencing platforms are depicted in arrows at the bottom

followed by ligation of an RNA-based 5′ linker oligo. We use the restriction enzyme PmeI to selectively remove cDNAs that carry the spike RNAs after we amplify the complementary DNA by PCR using primers with Illumina adapter sequences.

3.1.1 Extracting Total RNA from Fly Ovaries

We extract total RNA from freshly dissected ovaries. It is important to keep female fruit flies healthy and give enough food and space to lay eggs to obtain ovaries that include all developmental stages.

1. Culture 3–10 days old female flies for one or two days in a plastic cage with apple agar plate with a brushful of yeast paste at the center (*see* **Note 2**).
2. Place flies on the CO_2 gas pad until they stop moving. Sort females and discard males.

3. Transfer females into a plastic petri dish and place it on ice (*see* **Note 3**).
4. Place ~50 μl of ice-cold phosphate-buffered saline (PBS) under the standard stereo microscope.
5. Transfer up to ten females and dissect ovaries in PBS using a pair of clean tweezers. Remove other tissues, such as gut, fat bodies and Malpighian tubules.
6. Transfer dissected ovaries directly into 500 μl TriZol in a 1.5-ml Eppendorf tube.
7. Repeat **steps 4–6** until a desired amount of ovaries is harvested (*see* **Note 4**).
8. Use plastic pellet pestles to homogenize ovaries in TriZol until there is no visible debris of ovaries (*see* **Note 5**).
9. Incubate ovaries on ice for 1 h to completely dissolve the tissue.
10. Add 300 μl Chloroform, vortex the tube for 10 s, centrifuge the tube at the maximum speed of the refrigerated bench-top centrifuge for 5 min (usually 13 krpm).
11. Take the aqueous phase (the top layer) to a new tube, then add 0.9× volume (add 270 μl of 2-propanol when the aqueous phase is 300 μl). Mix it thoroughly, incubate for 30 min on ice, and centrifuge the tube at 13 krpm for 25 min. A white pellet of RNA should be visible after centrifugation.
12. Remove the supernatant and add 300 μl of ice-cold 80% v/v EtOH, and centrifuge the tube at 13 krpm for 5 min (*see* **Note 6**).
13. Remove the supernatant using P200, and spin the tube briefly in the small bench-top centrifuge to bring the residual ethanol down to the bottom of the tube.
14. Use P10 or P2 pipette to carefully remove the residual ethanol on the pellet, and gently let the pellet air-dry (*see* **Note 7**).
15. Resuspend the pellet in 90 μl ddH_2O by pipetting.
16. Add 10 μl of 10x DNase I buffer and 1 μl of DNase I.
17. Incubate the tube at 37 °C for 1 h.
18. Add 100 μl of ddH_2O, 150 μl of Acid-Phenol:Chloroform, vortex the tube for 10 s, centrifuge the tube at 13 krpm for 5 min.
19. Take the aqueous phase to a new tube, add 20 μl of 3 M sodium acetate, pH 5.2, 1 μl of 20 mg/ml Glycogen, and 180 μl of 2-propanol. Mix it thoroughly, incubate at −20 °C for 1 h, and centrifuge the tube at 13 krpm for 25 min.
20. Remove the supernatant and add 300 μl of ice-cold 80% v/v EtOH, and centrifuge the tube at 13 krpm for 5 min.

21. Remove the supernatant using P200, and spin the tube briefly in the small bench-top centrifuge to bring the residual ethanol down to the bottom of the tube.
22. Use P10 or P2 pipette to carefully remove the residual ethanol on the pellet, and gently let the pellet air-dry (*see* **Note 8**).
23. Dissolve the pellet in 30 μl of ddH_2O and measure the concentration of RNA using NanoDrop.

3.1.2 2S Ribosomal RNA Depletion

2S ribosomal RNA is a 30-nucleotide-long fly-specific structural component of the ribosome. It is abundant in all tissues, much more so than piRNAs in ovaries. Since its size overlaps with piRNAs, it is important to selectively remove 2S ribosomal RNA before cloning the small RNA pool. Additionally, or alternatively, the complementary DNA oligo with 3′ C3 spacer can be used during the 5′ linker ligation to specifically block the ligation of 2S rRNA [12].

1. Take 50 μl slurry of Dynabeads Streptavidin C1 in a 1.5-ml Eppendorf tube.
2. Capture the beads on the magnetic stand, rinse once with 500 μl of 0.5× SSC (1:40 dilution of 20× SSC in ddH_2O), and resuspend the beads in 100 μl of 0.5× SSC.
3. Add 50 pmol of biotinylated 2S ribosomal RNA capture oligo DNA to the beads, and incubate on ice for 30 min.
4. Capture and wash the beads twice with 500 μl 0.5× SSC, and resuspend the beads in 100 μl of 0.5× SSC.
5. Incubate the beads with the oligo at 65 °C for 5 min. In parallel, prepare 5 μg of total RNA in ddH_2O and incubate it in 80 °C for 5 min.
6. Mix RNA and the beads, and incubate it in 50 °C for 1 h.
7. Capture the beads and take the unbound fraction.
8. Add 10 μl of 3 M sodium Acetate, pH 5.2, 1 μl of 20 mg/ml Glycogen, and 100 μl of 2-propanol. Mix it thoroughly, incubate at −20 °C for 1 h, and centrifuge the tube at 13 krpm for 25 min.
9. Remove the supernatant and add 300 μl of ice-cold 80% v/v EtOH, and centrifuge the tube at 13 krpm for 5 min.
10. Remove the supernatant using P200, and spin the tube briefly in the small bench-top centrifuge to bring the residual ethanol down to the bottom of the tube.
11. Use P10 or P2 pipette to carefully remove the residual ethanol on the pellet, and gently let the pellet air-dry. Ready to load samples on the UREA-PAGE (*see* **Note 9**).

3.1.3 Radiolabel 19 mer and 35 mer Spike RNA

We use radiolabeled RNA of known sizes to guide the gel excision to size-select small RNA pool.

1. Set up a phosphorylation reaction: 10 pmol 19 mer and 35 mer spike RNA in separate tubes, 2 μl of fresh 6000 Ci/mmol [γ-^{32}P] ATP, 0.5 μl of 10 U/μl T4 Polynucleotide Kinase.
2. Incubate the reaction for 30 min at 37 °C.
3. Add 40 μl of TE buffer (10 mM Tris–HCl pH 8.0, 1 mM EDTA) and apply the solution to Micro Bio-spin P-6. Gel filtration typically yields ~60 μl (*see* **Note 10**).
4. Count the radioactivity by a Geiger counter. Usually, 1 μl of labeled spike solution gives 300–800 counts per second. Labeled spikes can be stored at −20 °C for 2 months.

3.1.4 Size Selection

1. Cast 11% w/v UREA-PAGE, 1 × TBE on Mini-PROTEAN short plate with 1.0 mm spacers.
2. Resuspend 2S ribosomal RNA-depleted RNA from Subheading 3.1.3 in ddH_2O, add 0.5 μl each of radiolabeled 19 mer and 35 mer spikes (estimated to be 50 fmol each), and 6 μl of 2× UREA-TBE loading dye to make the solution 1×.
3. Heat the RNA samples at 95 °C for 3 min, and quickly transfer the tubes on ice.
4. Load RNA samples on the gel and run it at 100 C.V. until the bromophenol blue dye runs about three quarters of the gel (*see* **Note 11**).
5. Open the glass plate and place two pieces of radiolabeled triangle filter papers onto the gel to mark the position of the gel, then wrap the gel by cling film (Fig. 3a).
6. Insert the glass plate into a phosphor imager cassette, place a phosphor screen atop, and wait for 30 min.
7. Image the phosphor screen by the phosphor imager, and print out the image.
8. Place the printed image underneath the glass plate, merge the triangle markers to accurately determine the positions of the 19 mer and 35 mer spikes.
9. Excise the gel between 19 mer and 35 mer while it is covered by the cling film (Fig. 3b).
10. Remove the cling film, cut the gel into a few pieces, and transfer them into 750 μl of Elution buffer (0.5 M sodium chloride and 0.02% w/v SDS) (*see* **Note 12**).
11. Rotate the gel in the elution buffer for overnight at 4 °C.
12. Pass the solution through the centrifugal filter, save the flow-through into a fresh Eppendorf tube, add 1 μl of 20 mg/ml Glycogen and 680 μl of 2-propanol. Mix it thoroughly, incubate for 2 h in −20 °C, and centrifuge the tube at 13 krpm for 25 min (*see* **Note 13**).

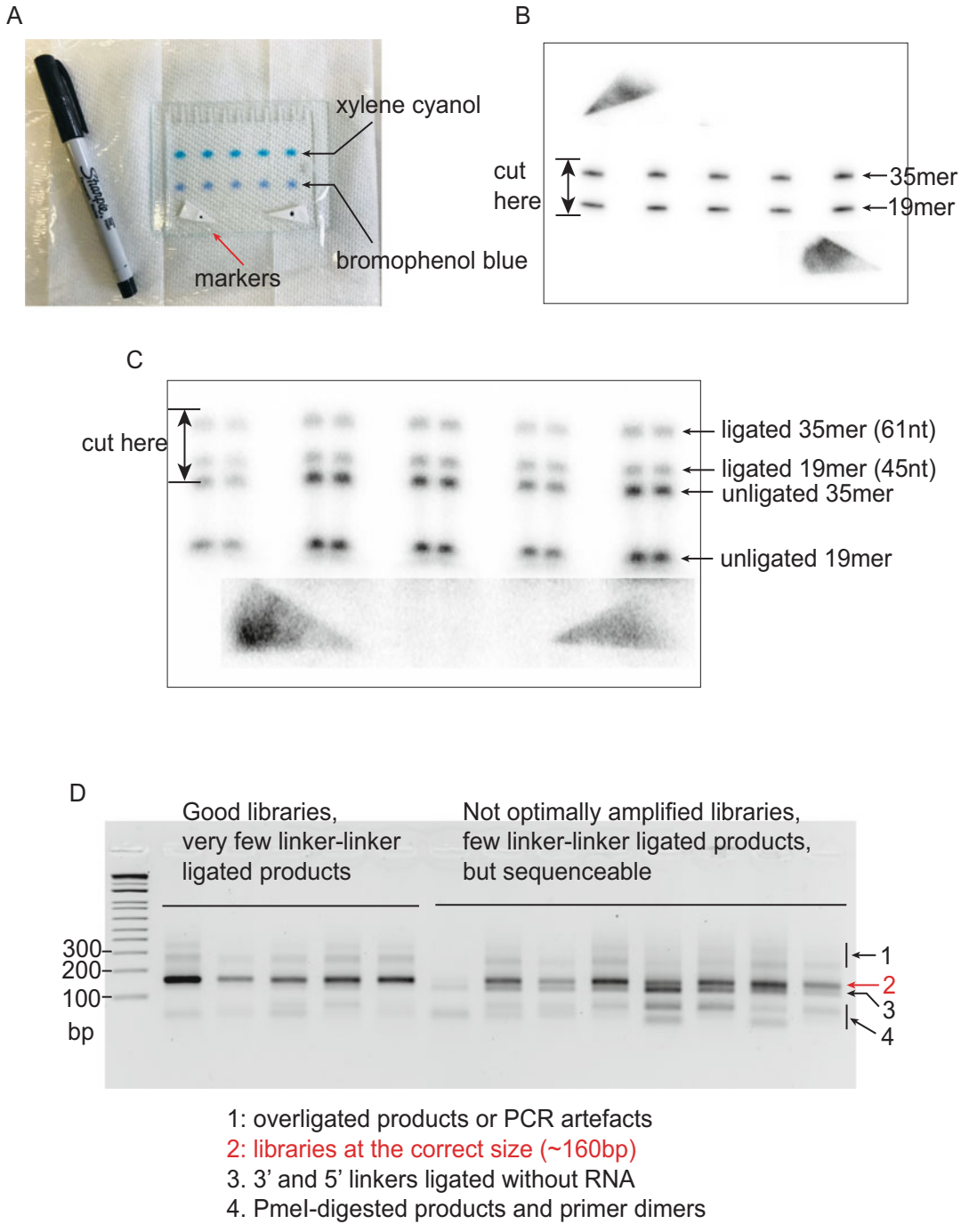


Fig. 3 Gel extraction of ligated and PCR-amplified small RNA. (**a**) Small RNAs are separated from large RNAs by UREA-PAGE using a Mini-PROTEAN gel. Samples are separated by empty lanes to minimize cross-contamination. (**b**) The size selection is aided by two radiolabeled spike-in RNA oligos of 19 and 35 nucleotides in length. Excise the gel that well includes both spikes. The radiolabeled triangular filter papers are put on the gel to mark the position. (**c**) The products of 3′ linker ligation are run on a UREA-PAGE. Run the gel long enough to separate the ligated products from unligated products. (**d**) PCR-amplified libraries are run on a low-melting Agarose gel to separate from other unwanted fragments produced by PCR

13. Remove the supernatant and add 150 μl of ice-cold 80% v/v EtOH, and centrifuge the tube at 13 krpm for 5 min.
14. Remove the supernatant using P200, and spin the tube briefly in the small bench-top centrifuge to bring the residual ethanol down to the bottom of the tube.
15. Use P10 or P2 pipette to carefully remove the residual ethanol on the pellet, and gently let the pellet air-dry.
16. Resuspend the pellet in 6 μl of ddH_2O by pipetting, and move to the 3′ linker ligation.

3.1.5 Oxidation of SMALL RNA Pool (Optional)

Several types of small RNAs including piRNAs are modified at the 3′ end by 2′-*O*-methylation, which can be used to selectively clone piRNAs. Periodate oxidation of unmodified 2′ and 3′ hydroxyl groups at the 3′ end of RNA makes the groups inert for the ligation of the adapter oligos while 2′-*O*-methylation renders the 3′ end of RNA resistant to oxidation [13]. A selective enrichment of piRNAs in the cloning step helps determine whether an RNA of the size of a piRNA found in the sequencing library is a bona fide piRNA or an RNA of different origins, such as intermediate products of RNA degradation. We do the oxidation reaction after the size selection, then repeat the size selection step before 3′ linker ligation.

1. Make 5× Borate buffer. Dissolve Borax anhydrous and Boric acid into ddH_2O to prepare 300 mM solutions. Mix approximately six and five portions of Borax anhydrous and Boric acid, respectively, to make 5× Borate buffer at pH 8.6. Keep it at room temperature.
2. Dissolve a small amount of Sodium periodate into ddH_2O to make a 10.3 mg/ml solution. Prepare a fresh solution before oxidation.
3. Mix 1.6 μl of 5× Borate buffer to RNA in a total of 7 μl, add 1 μl of 10.3 mg/ml Sodium periodate solution, and mix well by pipetting.
4. Incubate at room temperature for 40 min.
5. Add 8 μl of 2× UREA-TBE loading dye to the reaction.
6. Heat the RNA samples at 95 °C for 3 min, and quickly transfer the tubes on ice.
7. Run a 11% w/v UREA-PAGE as described above, and extract RNA from the gel.
8. Resuspend the pellet in 6 μl of ddH_2O, including about 50 fmols each of 19 mer and 35 mer spikes, and move to the 3′ linker ligation (*see* **Note 14**).

3.1.6 3′ Linker Ligation

1. Mix the size-selected RNA in 6 μl with 0.2 μl of 100 μM 3′ linker, denature RNA by heating at 65 °C for 5 min, and chill the tube on ice.
2. Make a mastermix of 1 μl of 10× T4 Ligase buffer, 2.5 μl of 50% PEG8000, 0.5 μl of T4 RNA Ligase 2, truncated K227Q, and 0.2 μl of SUPERase IN RNase Inhibitor per sample.
3. Add 4.2 μl of the reaction mastermix to the heat-denatured RNA + 3′ linker, incubate at 16 °C for overnight.
4. Run the reaction on a 11% w/v UREA-PAGE in the same way as in the size selection (Fig. 3c, *see* **Note 15**).
5. Excise the gel between the ligated products of spikes (45 nt and 61 nt in size, respectively) (*see* **Note 16**).
6. Follow the same protocol as in the size selection to elute and precipitate RNA, resuspend the pellet in 5 μl of ddH_2O, and move to the 5′ linker ligation.

3.1.7 5′ Linker Ligation

1. Mix the size-selected RNA with 0.2 μl of 100 μM 5′ linker, denature RNA by heating at 70 °C for 5 min, and chill the tube on ice.
2. Make a mastermix of 1 μl of 10× T4 Ligase buffer, 2.5 μl of 50% PEG8000, 1 μl of 10 mM ATP, 0.5 μl of T4 RNA Ligase 1, and 0.2 μl of SUPERase IN RNase Inhibitor.
3. Add 5.2 μl of the reaction mastermix to the heat-denatured RNA + 5′ linker, incubate at 16 °C for overnight.
4. Add 30 μl of ddH_2O and 80 μl of AMPure XP Beads, mix it thoroughly, incubate it for 5 min at room temperature, and capture the beads by the magnetic stand (*see* **Note 17**).
5. Discard the unbound fraction, wash the beads twice with 150 μl of 80% EtOH and air-dry the beads.
6. Resuspend the beads with 13 μl of ddH_2O, incubate it for 2 min at room temperature, and capture the beads by the magnetic stand.
7. Collect the eluted RNA into a fresh PCR tube, and move to the reverse transcription.

3.1.8 Reverse Transcription

1. Add 0.2 μl of 100 μM Solexa_RT_rev primer to the RNA sample after 5′ linker ligation.
2. Heat denature RNA at 65 °C for 5 min, then on ice.
3. Make a mastermix of 4 μl of 5× SuperScript II reaction buffer, 2 μl of 0.1 M DTT, 1 μl of 10 mM dNTPs, 0.3 μl of SuperScript II, and 0.2 μl of SUPERase IN RNase Inhibitor per sample.
4. Add 7.5 μl of the reaction mastermix to the heat-denatured RNA, incubate at 25 °C for 5 min, 42 °C for 50 min, and 70 °C for 15 min. Store the complementary DNA (cDNA) in −20 °C.

3.1.9 Amplify the Library Using KAPA DNA Polymerase

1. Prepare the following reaction for qPCR: 2 μl of cDNA, 4 μl of 5× KAPA reaction buffer, 2 μl of 25 mM $MgCl_2$, 1 μl of 10 mM dNTPs, 5 pmols each of Solexa_PCR_fw and TruSeq index primers, 0.1 μl of 2.5 U/μl KAPA DNA polymerase, 1 μl of 20× EvaGreen, and 0.2 μl of 100× ROX in 20 μl.
2. Split the reaction into two wells of a 384-well plate, and run a quantitative PCR with the following temperature cycle: 98 °C for 2 min, followed by 30 cycles of 98 °C for 12 s, 65 °C for 30 s, and 72 °C for 30 s.
3. Identify the cycle at which the amplified libraries reach a half to three quarters of the plateau.
4. Run the amplified libraries on a 1% w/v Agarose gel to check the size of the libraries (successfully cloned libraries are sized at around 160 bp).
5. Re-run the PCR in 60 μl using 6 μl of cDNA (20 μl each in three PCR tubes to make the reaction thermodynamically efficient) for the number of cycles determined in the **step 3**.
6. Precipitate the PCR products by 2-propanol as described above, resuspend the pellet in a reaction for the PmeI restriction enzyme digest. Use 0.5 μl of 5 U/μl PmeI in a 10 μl reaction, and incubate it at 37 °C for 2 h (*see* **Note 18**).
7. Prepare a 2% w/v low-melting Agarose gel (*see* **Note 19**).
8. Load the PmeI-digested libraries along with the 1 kb PLUS DNA ladder.
9. Excise the bands around 160 bp and purify DNA using Zymoclean spin columns. Elute the libraries in 20 μl ddH_2O, and measure the DNA concentration by Qubit DNA High Sensitivity kit. It is important to run the gel long enough to separate the libraries from other unwanted products (Fig. 3d).
10. Combine samples together for a multiplexed high-throughput sequencing as required. ~30 million reads per library usually give an enough complexity for the downstream analyses.

3.2 Computational Analysis of Small RNA Sequencing Libraries

High-throughput sequencing is usually done through Illumina platforms. For example, the standard paired end sequencing on the HiSeq platform produces Read 1 (R1) reads that read from the 5′ end to the 3′ end of the following structure where the sequencing starts from four Ns at the 5′ end of the sense strand of small RNA (*see* also Fig. 2).

5′ AATGATACGGCGACCACCGAGATCTACAC[TCTTTCCC TCACGACGCTCTTCCGATCT]--- NNNN---[small RNA sense] --- NNNN ---[AGATCGGAAGAGCACACGTCTGAACT CCAGTCAC] ---[i7 6mer barcode] --- ATCTCGTATGCCGTCT TCTGCTTG 3′.

We use bash scripts to process the sequencing data and map processed reads against genomic and transposon reference sequences. We use R scripts for manipulating tables, linkage analyses, and generating graphs. Codes are shaded in gray in the text.

3.2.1 Filter Poor-Quality Reads and Trim Adapter Sequencing

We first trim sequencing reads to the size that covers the small RNA and the adapter. Suppose we store the raw fastq file in the directory `${raw_data}` and place the processed data in the directory `${analysis}/${lib}` where `${lib}` is the name of individual sequenced libraries. Fastx-toolkit from Hannon Lab (http://hannonlab.cshl.edu/fastx_toolkit/commandline.html) offers a variety of tools for processing the Illumina sequencing libraries.

As the first step, if the sequencing is performed at the Paired-End 150 bp mode, we only keep the first 90 bases of R1 reads by `fastx_trimmer -l 90`. Reads of poor sequencing quality are discarded by applying the filter `fastq_quality_filter -q 33 -p 40` that takes reads in which more than 40% of bases have PHRED quality scores of greater than 33. This filter usually removes 1–2% of poor-quality reads.

The fastq format is then converted to the fasta format by `fastq_to_fasta -Q33` (*see* **Note 20**), before trimming the adapter sequence by `fastx_clipper -c -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. -l 18` option removes reads that become shorter than 18 bases after the adapter trimming.

Reads of 18–40 bases in size are taken, and the reads from the same sequences are collapsed to one entry per sequence while retaining the number of reads per sequence after "@", which is operated by the array function of awk, `READS[substr($NF,5,length($NF)-8)]++` (*see* **Note 21**).

```
### make collapsed fasta file
fastx_trimmer -l 90 -i <(zcat ${raw_data}/${lib}_R1.fastq.gz) |\
fastq_quality_filter -q 33 -p 40 |\
fastq_to_fasta -Q33 |\
fastx_clipper -c -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -l 18 |\
fasta_formatter - -t |\
awk '{if(length($NF)>25 && length($NF)<49) READS[substr
($NF,5,length($NF)-8)]++
} END {for(var in READS) print ">"var"@"READS[var]"\n"var
}' > ${analysis}/${lib}/${lib}_collapsed.Q33P40.fa
```

3.2.2 Filter Reads that Come from Infrastructural RNAs as Well as MicroRNAs

In the next step, we filter reads that are derived from ribosomal, transfer, small nuclear, and small nucleolar RNA. Fragments of these infrastructural RNAs can be loaded to the PIWI clade Argonaute proteins [14]. However, they confound the downstream analysis of piRNAs from mRNA and transposons especially when they are located nearby in the genome. We include microRNAs in the filter for the same reason.

First, we build the bowtie index of the reference sequences of the infrastructural RNA, which can be downloaded from the FlyBase (http://flybase.org) or Ensembl (https://ensembl.org). We recommend using the same assembly and the annotation throughout the same study. Reads are filtered in the following command of bowtie allowing up to one mismatch: `bowtie -f -v 1 -a`. MicroRNA reads can be used for the normalization of the sequencing depth in a later step.

```
### map reads to infrastructural RNA and keep unmapped reads
bowtie -f -v 1 -a -S ${misc_index} ${analysis}/${lib}/${lib}
_collapsed.Q33P40.fa --un
${analysis}/${lib}/${lib}_collapsed.Q33P40.misc-unmapped.fa |\
samtools view -bS - | \
bamToBed -i - > ${analysis}/${lib}/${lib}_collapsed.Q33P40.
misc-mapped.bed
```

3.2.3 Map the Reads Against the Genome and Take the Uniquely Mapping Reads

piRNAs are abundantly produced from RNA transcribed from distinct genomic loci called piRNA clusters or piRNA source loci, which usually consist of repeat elements and transposon sequences [5]. piRNA clusters include both intact copies and degenerated fragments of transposons. piRNAs are also derived from mRNAs and intergenic regions. Notable examples include 3′ UTR-derived genic piRNAs in fruit flies and pachytene piRNAs in mice [15, 16].

To map piRNA reads to these sequences, we separately map sequencing reads to the genomic sequence and the transposon reference sequences.

When mapping reads to the genome, we recommend only taking the reads that map to one specific locus of the genome, which can be done by `-m 1` option in bowtie.

```
### map reads to the genome, and report uniquely-mapped reads
bowtie -f -v 1 -m 1 -S ${genome_index_genome} ${analysis}/
${lib}/${lib}_collapsed.Q33P40.misc-unmapped.fa |\
samtools view -bS - | bamToBed -i - > ${analysis}/${lib}/
${lib}_misc-unmapped_genome-unique-mappers.bed
```

3.2.4 Map the Reads Against Transposon Reference Sequences

Transposon reference sequences can be downloaded from the RepBase (https://www-girinst-org/repbase/) or are available from Senti et al. (2015) for a refined reference of transposons found in *Drosophila melanogaster* [17]. When mapping reads to the transposon sequences, we recommend allowing up to three mismatches (`-v 3` option in bowtie) to cover the frequent nucleotide variations in transposon copies. We usually use `--all --best --strata` option to take piRNAs that map to multiple transposons while only taking the best mapping results. This is necessary to have a seamless mapping of reads to the reference, which is important for the linkage analyses.

```
### map reads to transposon sequences
bowtie -f -v 3 --all --best --strata -S ${TE_index}
${analysis}/${lib}/${lib}_collapsed.Q33P40.misc-unmapped.fa |\
samtools view -b -S - | bamToBed -i - > ${analysis}/${lib}/
${lib}_TE_3MM_mappers.bed
```

3.2.5 Evaluating the Complexity of the Sequencing Libraries (Optional)

PCR amplification can introduce a bias in the sequenced pool of small RNAs. Four random nucleotides at the ends of 3′ and 5′ linkers can be used to evaluate the complexity of sequencing libraries. Reads of the same sequence that are flanked by the same four nucleotides are likely the products of PCR duplication. We use genome-unique mappers to evaluate the degree of PCR duplication.

First, we process the fastq file in the same way as described above except for retaining the read id instead of collapsing reads by sequence.

```
### process reads while keeping the read ids
fastx_trimmer -l 90 -i <(zcat ${raw_data}/${lib}_combined_R1.
fastq.gz) |\
fastq_quality_filter -q 33 -p 40 |\
fastq_to_fasta -Q33 |\
fastx_clipper -c -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -l
18 -i - |\
fasta_formatter - -t |\
awk '{if(length($NF)>25 && length($NF)<49) print
">"$1"\n"substr($NF,5,length($NF)-8)
}' > ${analysis}/${lib}/${lib}.Q33P40.fa
```

Then filter reads that map to the infrastructural RNAs:

```
bowtie -f -v 1 -a -S ${misc_index} ${analysis}/${lib}/${lib}.
Q33P40.4Ns.fa --un ${analysis}/${lib}/${lib}.Q33P40.misc-un-
mapped.fa
```

We fetch the read ids that uniquely map to the genome.

```
bowtie -f -v 1 -m 1 -S ${genome_index} ${analysis}/${lib}/
${lib}.Q33P40.misc-unmapped.fa |\
samtools view -bS - | bamToBed -i - |\
awk '{print $4}' > ${analysis}/${lib}/${lib}.Q33P40.misc-un-
mapped.genome-unique-mappers.ids
```

We use seqtk to intersect the original fastq file with ids of the genome-unique mappers. We obtain uniquely identifiable molecules before PCR after applying the same filters as above but without trimming the four nucleotides at both ends of small RNA sequence.

```
seqtk subseq <(zcat ${raw_data}/${lib}_R1.fastq.gz) ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-unique-mappers.ids |\
fastx_trimmer -l 90 |\
fastq_quality_filter -q 33 -p 40 |\
fastq_to_fasta -Q33 |\
fastx_clipper -c -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -l 18 |\
grep -v ">" | sort | uniq
```

The degree of library complexity is calculated by the number of uniquely identifiable molecules divided by the number of genome-unique mappers. Small RNA libraries usually amplify after 10–16 cycles of PCR at the last step of library preparation. The library complexity is typically between 75% and 82% when about 30 million reads are sequenced from fruit fly ovary samples. The complexity starts to drop after 18 cycles of PCR.

3.2.6 Linkage Analysis of Ping-Pong and Phasing piRNA Biogenesis

piRNAs are generated by two distinct mechanisms called ping-pong and phasing (Fig. 1). The major difference of the two pathways is that ping-pong is mediated by the slicing activity of piRNA-loaded Argonaute proteins while the phasing is mediated by the endonuclease Zucchini/MitoPLD. Both nucleases produce the 5′ end of piRNAs. Ping-pong produces a pair of piRNAs from opposing strands of double-stranded RNA (usually sense (plus strand) and antisense (minus strand) transposon RNAs), whose 5′ ends overlap by ten nucleotides. Phasing sequentially produces piRNAs from the same RNA strand, where the production of 3′ end of one piRNA precedes the 5′ end of the next piRNA.

The purpose of the linkage analysis is to measure the proportion of these different piRNA biogenesis pathways that have occurred in vivo to result in the observed piRNA pool. We start by detecting the ping-pong biogenesis in transposon-mapping reads.

Prepare the "End Counts" Table for Transposon-Mapping Reads

The first step is to generate a table that counts the number of 3′ and 5′ ends of the read at every coordinate of the transposon reference sequences. To make a seamless table where every coordinate gets a number, we make a bed file of pseudo piRNAs that map to every position. Take *F-element* as an example.

```
TE="F-element"
length="4708"
for ((i=0; i<=${length}-1; i++)); do
j=$((${i}+1))
printf "${TE} ${i} ${j} AAAAAAAAAAAAAAAAAAAAAAA@1 . +""\n""${TE} ${i} ${j} AAAAAAAAAAAAAAAAAAAAAAA@1 . -""\n" | tr ' ' '\t' >> ${analysis}/${lib}/End.plots.TEs/${TE}_pseudo.piRNAs.bed
done
```

Combine the bed file of transposon mapping reads and the pseudo piRNAs while only retaining reads that are longer than 22 nucleotides to exclude siRNAs from the analysis.

```
cat ${analysis}/${lib}/${lib}_TE_3MM_mappers.bed ${analysis}/
${lib}/End.plots.TEs/${TE}_pseudo.piRNAs.bed |\
awk -v TE=${TE} '{split($4,a,"@"); if($6=="+" && length(a[1])>
22 && $1==TE) {PLUS5[$2]+=a[2]; PLUS3[$3-1]+=a[2]
} else if($6=="-" && length(a[1])>22 && $1==TE) {MINUS5[$3-1]
+=a[2]; MINUS3[$2]+=a[2]
}} END {for(var in PLUS5) print var,PLUS5[var]-1,"plu-
s_5end""\n"var,PLUS3[var]-1,"plus_3end""\n"var,MINUS5[var]-
1,"minus_5end""\n"var,MINUS3[var]-1,"minus_3end"
}' > ${analysis}/${lib}/End.plots.TEs/${lib}_${TE}_3MM_map-
pers.counts
### the resulting table looks as follows:
<coordinate> <number of reads> <end>
1140 111 plus_5end
1140 10 plus_3end
1140 0 minus_5end
1140 25 minus_3end
...
```

Calculate the Linkage Using R

We import the table into R and transform it to a table that has the counts of "plus_5end," "plus_3end," "minus_5end," and "minus_3end" at each column using `dcast` function:

```
library(reshape2)
table=read.table("3MM_mappers.counts", header=F)
table_d <- dcast(table,V1~V3, value.var="V2")
```

The resulting table looks as in Fig. 4a. To measure the linkage between the 5' ends of plus and minus reads, we first take the counts of "plus_5end" at each coordinate and take the counts of "minus_5end" from a window of 20 nucleotides' offset starting at its position (Fig. 4a). We divide the counts of "minus_5end" at each nucleotide position by the total counts from the 20 nucleotides' window to obtain the frequency of "minus_5end" at each position within the window. We calculate the product of the frequency at each position and the count of "plus_5end." We do this for the entire length of transposon and sum the products per position in the 20 nucleotides' window, which yields the chance by which the 5' end of minus-strand-mapping piRNAs are found at given offsets to the 5' end of plus-strand-mapping piRNAs.

To achieve this calculation, we first set the length, start, and end positions of the window as variables in R:

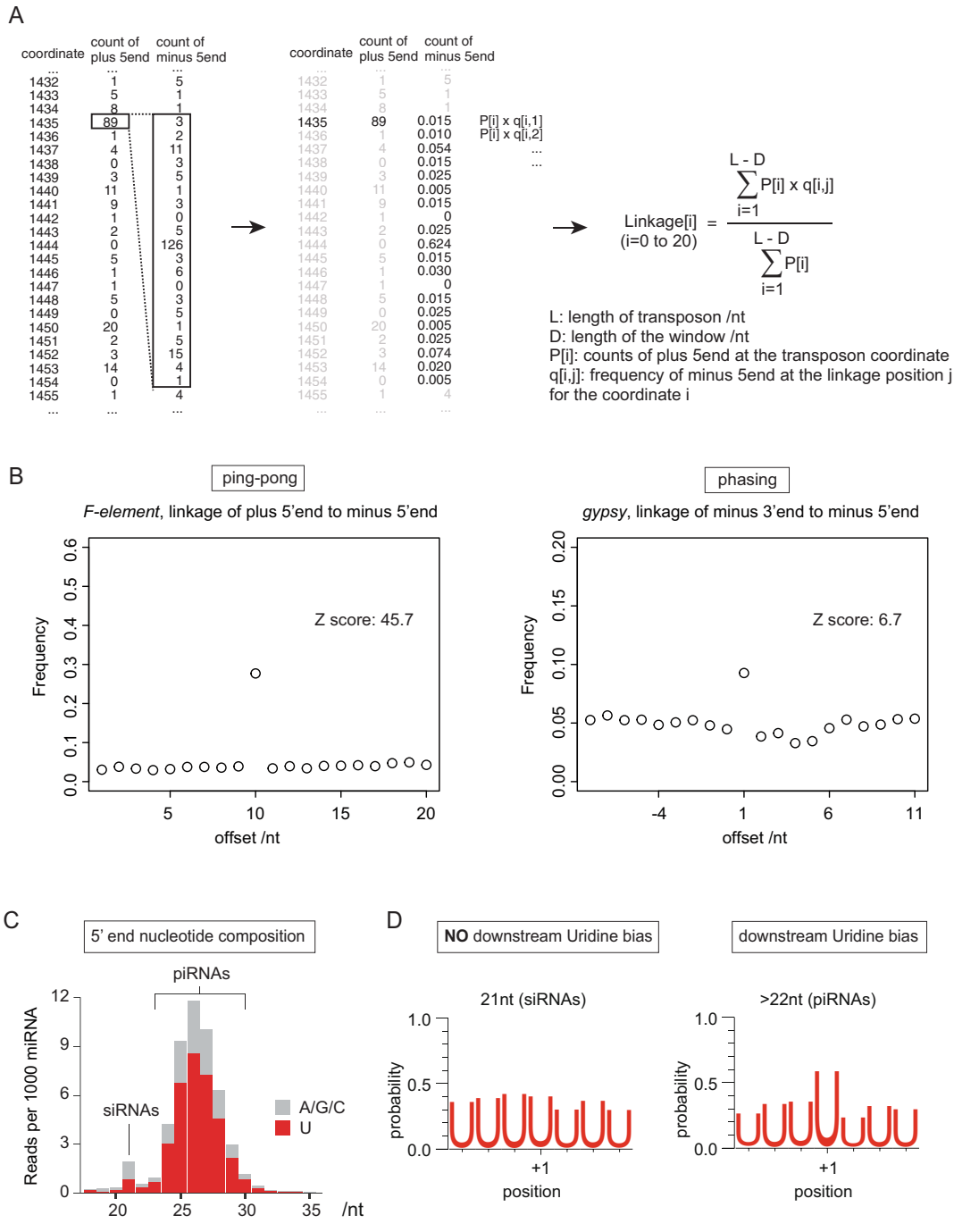


Fig. 4 The linkage analysis of ping-pong and phasing piRNAs. (**a**) The analysis takes a table of read counts of piRNA 3′ and 5′ ends mapping at each nucleotide position of transposon. For the ping-pong linkage analysis, the frequencies of the 5′ ends of minus-strand piRNAs in a 20 nucleotides' window are calculated for a given 5′ end of plus-strand piRNA. The frequencies of minus-strand piRNAs at each position in the 20 nucleotides' window are weighted by the abundance of the matching plus-strand piRNA to yield the overall frequencies. (**b**) Shown are examples of frequency plots for ping-pong and phasing piRNAs from *F-element* and *gypsy* transposons, respectively. Z-scores are calculated as the statistical significance of frequencies at offset nucleotides of 10 and +1, respectively. (**c**) Shown is the abundance of fruit fly ovarian small RNAs of 18–35

```
### parameters for the plus 5end to minus 5end pingpong
LENGTH=length(table_d[,1])
START=0
END=19
POSITION=9
table_a=table_d$plus_5end
table_b=table_d$minus_5end

### make variables numeric
length=as.numeric(LENGTH)
start=as.numeric(START)
end=as.numeric(END)
link=as.numeric(POSITION)
start.position_B=1-start
end.position_B=length-end
link.position=link-start+1
span=end-start+1
```

Then we measure the linkage as follows:

```
### define a data.frame and a vector
offset <- data.frame()
sum_offset = NULL

### core of linkage calculation
for (i in 1:span) {
 for (j in start.position_B:end.position_B) {
 offset[j,i] = table_a[j]*table_b[j+start+i-1]/sum(table_b[(j
+start):(j+end)])
 }
}

offset[is.na(offset)] <- 0

for (i in 1:span) {
 sum_offset[i] = sum(offset[,i])/sum(offset)
}
```

The linkage values for the ping-pong differ when the calculation is done reciprocally. The above analysis fixes the position of "plus_5end." Alternatively, the ping-pong linkage can be measured by calculating the read counts of the plus 5′ ends per each fixed

Fig. 4 (continued) nucleotides in length mapping to mRNA 3′ UTRs. 5′ end nucleotides are biased to Uridine for piRNAs, but not for siRNAs. (**d**) Weblogo plots for the frequencies of Uridine around the 3′ end of 3′ UTR-derived siRNAs and piRNAs. One nucleotide downstream of piRNA 3′ ends (+1 position) is biased to Uridine, whereas there is no nucleotide bias around the 3′ end of siRNAs

minus 5′ end. We only need to change the parameters at the initial setting below and go through the whole calculation to obtain the linkage value.

```
### parameters for the minus 5end to plus 5end pingpong
START=-19
END=0
POSITION=-9
table_a=table_d$minus_5end
table_b=table_d$plus_5end
```

The linkage of the plus 5′ ends to the minus 5′ ends and vice versa should yield similar values because the abundance of ping-pong piRNA pairs usually correlate.

The linkage of the phasing piRNA biogenesis can also be calculated in a similar way by simply changing the initial parameters as follows:

```
### parameters for the plus 3end to plus 5end phasing
START=-9
END=10
POSITION=1
table_a=table_d$plus_3end
table_b=table_d$plus_5end

### parameters for the minus 3end to minus 5end phasing
START=-10
END=9
POSITION=-1
table_a=table_d$minus_3end
table_b=table_d$minus_5end
```

Evaluate the Statistical Significance of the Linkage Analysis

Z scores are commonly used to evaluate the statistical significance of the linkage. Z score is defined by the magnitude of deviation in frequency at the linkage position from the mean frequency. We first calculate the mean and the standard deviation of frequencies at all nucleotide offsets excluding the linkage position. We then divide the difference between frequencies of the linkage position and the mean by the standard deviation to obtain the Z score (*see* **Note 22**).

Z scores can be calculated in R as follows:

```
sum_offset2=sum_offset[c(1:(link.position-1),(link.position
+1):span)]
m=mean(sum_offset2)
s=sd(sum_offset2)
z=(sum_offset[link.position]-m)/s
```

Figure 4b shows examples of the frequency plots for ping-pong and phasing of piRNAs mapping to *F-element* and *gypsy*, respectively. The signature of ping-pong biogenesis is more readily observed in the piRNA pool from fruit fly ovaries while the phasing is less clear. This is evident by frequencies at the linkage-offset position and the Z scores. This is partly explained by the further processing of piRNA 3′ ends after the cleavage by Zucchini/MitoPLD [18]. Hence, the 5′ end of the next piRNA is not always one nucleotide downstream of the 3′ end of upstream piRNA. Furthermore, we observe that a number of phasing biogenesis appear to overlap to each other more so than ping-pong piRNA pairs do.

3.2.7 Measure the Downstream Uridine Bias for Phased piRNAs

Uridine is preferentially found at the 5′ end position of piRNAs. This is because the 5′ end Uridine residue is preferred by PIWI proteins and also because the endonuclease Zucchini/MitoPLD prefers to cut a nucleotide before Uridine [7, 19]. Since the phasing predominantly involves Zucchini-mediated cleavage, the immediate downstream nucleotide of the 3′ end of a phased piRNA is also biased to Uridine. This so-called downstream Uridine bias can be used to assess the per-transcript and the transcriptome-wide prevalence of phasing biogenesis [7]. This is particularly useful when phased piRNAs are not densely populated in certain regions and the serially produced piRNAs are not detected due to low abundance.

Obtain Reads that Are Derived from mRNA 3′ UTRs and Measure the 5′ End Nucleotide Composition

As a first step, we extract small RNA reads that map to the mRNA 3′ UTRs using `bedtools intersect`.

```
### intersect genome unique mappers to annotated 3UTR regions,
keep reads that overlap at least half of their length
bedtools intersect -wa -s -f 0.5 \
-a ${analysis}/${lib}/${lib}_misc-unmapped_genome-unique-map-
pers.bed \
-b ${3UTR_bed} > ${analysis}/${lib}/${lib}.Q33P40.misc-un-
mapped.genome-unique-mappers.3UTR.bed
```

The awk command below counts the number of reads by size and the identity of nucleotide at the 5′ end:

```
### measure the 5'end nucleotide composition for each size of
small RNA reads
cat ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-un-
ique-mappers.3UTR.bed |\
awk '{ split($4,a,"@"); FIRST[$3-$2" "substr(a[1],1,1)]+=a[2]
} END { for(var in FIRST) print var,FIRST[var]
}' > ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-
```

```
unique-mappers.3UTR.1st.10th.nucleotide.length.table
### the resulting table looks as follows:
<size> <nucleotide> <counts>
22 T 3581
25 A 1381
32 T 1091
...
```

The resulting table can be used to plot the nucleotide composition at the 5′ end of reads in the histogram of size distribution. As shown in Fig. 4c, the piRNA population peaking at the length around 24–26 nucleotides mostly have Uridine at the 5′ end while siRNAs whose dominant size is 21 nucleotides do not show the bias. Note that mRNAs that overlap at the 3′ UTR form double-stranded RNA and become substrates of Dicer in flies [20–22].

Measure the Downstream Uridine Bias

Using the information of the genomic coordinates, one can retrieve the sequence around the 3′ end of reads using `bedtools getfasta`. The code below allows to extract the seven nucleotides' window centered at the immediate downstream position of the read. We do this for reads of 21 nucleotides in length and reads that are longer than 22 nucleotides, which mostly consist of siRNAs and piRNAs, respectively.

```
### extract the sequence around the 3'end of small RNA reads
### only considering reads of 21 nucleotides in length (siRNAs)
awk '{if($3-$2==21) print}' ${analysis}/${lib}/${lib}.Q33P40.
misc-unmapped.genome-unique-mappers.3UTR.bed |\
awk '{if($6=="+") print $1,$3-3,$3+4,$4,$5,$6; else print $1,
$2-4,$2+3,$4,$5,$6}' | tr ' ' '\t' |\
bedtools getfasta -s -name -tab -fi ${fasta-file-for-the-
genome} -bed - |\
tr '@()' ' ' | awk '{for(i=1;i<=$2;i++) print ">"$NF"\n"$NF}'
|\
tr 'T' 'U' > ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.
genome-unique-mappers.3UTR.downstream.21nt.fa
### the resulting fasta file looks as follows:
>UAGCUAC
UAGCUAC
>GUCUAAC
GUCUAAC
...
```

```
### extract the sequence around the 3'end of small RNA reads
### only considering reads longer than 22 nucleotides (piRNAs)
awk '{if($3-$2>22) print}' ${analysis}/${lib}/${lib}.Q33P40.
misc-unmapped.genome-unique-mappers.3UTR.bed |\
awk '{if($6=="+") print $1,$3-3,$3+4,$4,$5,$6; else print $1,
$2-4,$2+3,$4,$5,$6}' | tr ' ' '\t' |\
bedtools getfasta -s -name -tab -fi ${fasta-file-for-the-
genome} -bed - |\
tr '@()' ' ' | awk '{for(i=1;i<=$2;i++) print ">"$NF"\n"$NF}'
|\
tr 'T' 'U' > ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.
genome-unique-mappers.3UTR.downstream.gt22nt.fa
```

The resulting fasta file can be used to measure the nucleotide composition at every base position using `weblogo`.

```
### weblogo for 21mers
cat ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-un-
ique-mappers.3UTR.downstream.21nt.fa |\
weblogo -U probability -A rna -F pdf -n 50 -c classic \
-o ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-un-
ique-mappers.3UTR.downstream.7mers.21nt.logo_prob.pdf

### weblogo for reads longer than 22nt
cat ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-un-
ique-mappers.3UTR.downstream.gt22nt.fa |\
weblogo -U probability -A rna -F pdf -n 50 -c classic \
-o ${analysis}/${lib}/${lib}.Q33P40.misc-unmapped.genome-un-
ique-mappers.3UTR.downstream.7mers.gt22nt.logo_prob.pdf
```

The logos that are generated from weblogo are shown in Fig. 4d. As previously shown, the immediate downstream nucleotide of the 3′ end of piRNAs (longer than 22 nucleotides) is biased to Uridine while siRNAs (21 nucleotides) do not show the bias [7].

4 Notes

1. Purify fresh 3′ and 5′ linker on a 12% w/v UREA-PAGE, measure concentration and dilute to 100 μM. Freeze aliquots at −80 °C.
2. Do not put too many flies in the cage, and change the plate at least twice a day to make sure that flies can find space to lay eggs. Otherwise, ovaries will be occupied by matured eggs and skew the representation of egg chambers from different developmental stages.

3. This step is necessary when the dissection takes more than 30 min. We experience that flies suffer if they are under CO_2 for a long time and may impact RNA in ovaries in an unpredicted manner.
4. We typically obtain 30 μg of total RNA from ~30 μl bed volume of ovaries. Having more ovaries may result in an incomplete dissolution of ovaries in TriZol. Prepare another tube with TriZol if more RNA is needed. Keep ovaries in Trizol on ice.
5. One can use the portable grinding device or squish ovaries by hand. When using the grinding device, make sure that the volume of TriZol is less than 500 μl to avoid splashing the liquid out of the tube. Egg cuticles usually do not dissolve in TriZol, and remain visible after the grinding. But, they do not contribute to the small RNA pool.
6. We do not disturb the pellet when adding EtOH. Put the tube back to the centrifuge right after EtOH is added in the tube.
7. Do not over-dry the pellet. When there is a significant contamination of genomic DNA, over-dried pellet becomes difficult to dissolve in water.
8. It is not so critical for not over-drying the pellet after the DNase digest because RNA readily dissolves in water.
9. We prepare two solutions containing 0% and 20% w/v of Acrylamide, respectively, in 7 M UREA, 1 × TBE in ddH_2O.- Solutions are stable at room temperature for at least 3 months. 6.5 ml is enough to fill a Mini-PROTEAN gel. To 6.5 ml of Acrylamide solution, we add 65 μl of 10% w/v Ammonium persulfate (APS), and 3.5 μl of *N*,*N*,*N*′,*N*-′-Tetramethylethylenediamine (TEMED). It is important to remove the comb once the gel polymerizes, and rinse inside the well with water. If this step is omitted, the residual Acrylamide polymerizes inside the well and disturbs the sample loading. It is also important not to add too much TEMED for the same reason.
10. Follow manufacturer's instruction.
11. Make at least one lane empty between samples to avoid cross-contaminations.
12. Keep the gel pieces moist to allow an efficient elution. Prepare the elution buffer in Eppendorf tube in advance so that the gel can be transferred as soon as they are cut into pieces.
13. Any centrifugal columns that do not bind RNA can be used.
14. 19 mer and 35 mer spike RNAs carry unmodified 2′ hydroxyl group. Therefore, it is necessary to add fresh RNAs after oxidation to guide the gel excision after the 3′ linker ligation. The

second PAGE is necessary after the oxidation and before the 3′ linker ligation in order to completely remove periodate.

15. We recommend using 15-well comb instead of 10-well and load samples in two lanes per sample to avoid spill-over of samples from the well.
16. It is crucial to remove the unligated 3′ linker (26 mer). Keep in mind that 3′ linker smears out of its size and can contribute a large quantity into the final library. Therefore, it is important to cut tightly at or above, but not below, the unligated 35 mer spike.
17. One can run the UREA-PAGE after the 5′ linker ligation to remove the unligated 5′ linkers. However, we found this process unnecessary, and it only reduces the recovery of RNA. One can do standard isopropanol precipitation instead of using AMPure beads to remove the salts and PEG from the reaction before the reverse transcription reaction. The 5' linkers would remain in the reverse transcription reaction, but would not be amplified as long as they are not ligated to the 3' linkers.
18. PmeI digests cDNAs derived from the 19 mer and 35 mer spikes.
19. Use a hot plate stirrer to gradually melt the low-melting Agarose to achieve consistency. This takes about 1.5 h. The low-melting Agarose is delicate. Carefully remove the comb off the gel and make sure that there are no holes in the wells by adding 1 μl of loading dye before loading the samples. Do not overheat the gel during the electrophoresis. A lower voltage is recommended.
20. FASTX-toolkit assumes quality scores with ASCII offset 64, but the data is Sanger encoded (offset 33), so this needs to be specified explicitly by adding the -Q33 flag.
21. Four Ns at both ends of the small RNA reads are removed by substr function in awk.
22. Z scores change when the window size changes. We take the window of 20 bases, which can be shortened or lengthened depending on the purpose of the analysis.

Acknowledgments

I acknowledge members of the Julius Brennecke lab, especially Dominik Handler, for technical inputs in the library preparation and the computational analyses. This work is supported by the Australian Research Council.

References

1. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10:94–108. https://doi.org/10.1038/nrg2504
2. Sato K, Siomi MC (2020) The piRNA pathway in drosophila ovarian germ and somatic cells. Proc Jpn Acad Ser B Phys Biol Sci 96:32–42. https://doi.org/10.2183/pjab.96.003
3. Vourekas A et al (2012) Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat Struct Mol Biol 19:773–781. https://doi.org/10.1038/nsmb.2347
4. Hayashi R et al (2016) Genetic and mechanistic diversity of piRNA 3′-end formation. Nature 539:588–592. https://doi.org/10.1038/nature20162
5. Brennecke J et al (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 128:1089–1103. https://doi.org/10.1016/j.cell.2007.01.043
6. Gunawardane LS et al (2007) A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. Science 315:1587–1590. https://doi.org/10.1126/science.1140494
7. Mohn F, Handler D, Brennecke J (2015) Noncoding RNA. piRNA-guided slicing specifies transcripts for zucchini-dependent, phased piRNA biogenesis. Science 348:812–817. https://doi.org/10.1126/science.aaa1039
8. Han BW, Wang W, Li C, Weng Z, Zamore PD (2015) Noncoding RNA. piRNA-guided transposon cleavage initiates zucchini-dependent, phased piRNA production. Science 348:817–821. https://doi.org/10.1126/science.aaa1264
9. Ishizu H et al (2015) Somatic primary piRNA biogenesis driven by cis-acting RNA elements and trans-acting Yb. Cell Rep 12:429–440. https://doi.org/10.1016/j.celrep.2015.06.035
10. Özata DM et al (2020) Evolutionarily conserved pachytene piRNA loci are highly divergent among modern humans. Nat Ecol Evol 4:156–168. https://doi.org/10.1038/s41559-019-1065-1
11. Hafner M et al (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods 44:3–12. https://doi.org/10.1016/j.ymeth.2007.09.009
12. Wickersheim ML, Blumenstiel JP (2013) Terminator oligo blocking efficiently eliminates rRNA from *Drosophila* small RNA sequencing libraries. BioTechniques 55:269–272. https://doi.org/10.2144/000114102
13. Akbergenov R et al (2006) Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res 34:462–471. https://doi.org/10.1093/nar/gkj447
14. Honda S et al (2017) The biogenesis pathway of tRNA-derived piRNAs in Bombyx germ cells. Nucleic Acids Res 45:9108–9120. https://doi.org/10.1093/nar/gkx537
15. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316:744–747. https://doi.org/10.1126/science.1142612
16. Saito K et al (2009) A regulatory circuit for piwi by the large Maf gene traffic jam in *Drosophila*. Nature 461:1296–1299. https://doi.org/10.1038/nature08501
17. Senti KA, Jurczak D, Sachidanandam R, Brennecke J (2015) piRNA-guided slicing of transposon transcripts enforces their transcriptional silencing via specifying the nuclear piRNA repertoire. Genes Dev 29:1747–1762. https://doi.org/10.1101/gad.267252.115
18. Izumi N et al (2016) Identification and functional analysis of the pre-piRNA 3′ trimmer in silkworms. Cell 164:962–973. https://doi.org/10.1016/j.cell.2016.01.008
19. Matsumoto N et al (2016) Crystal structure of silkworm PIWI-clade Argonaute Siwi bound to piRNA. Cell 167:484–497.e489. https://doi.org/10.1016/j.cell.2016.09.002
20. Kawamura Y et al (2008) Drosophila endogenous small RNAs bind to Argonaute 2 in

somatic cells. Nature 453:793–797. https://doi.org/10.1038/nature06938

21. Czech B et al (2008) An endogenous small interfering RNA pathway in *Drosophila*. Nature 453:798–802. https://doi.org/10.1038/nature07007

22. Ghildiyal M et al (2008) Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. Science 320:1077–1081. https://doi.org/10.1126/science.1157396

Chapter 8

Drosophila Genetic Resources for Elucidating piRNA Pathway

Kuniaki Saito

Abstract

Emerging evidence indicates that PIWI proteins, in collaboration with PIWI-interacting RNAs (piRNAs), play a critical role in gonadal development and retrotransposon silencing in metazoans. Numerous studies have characterized the mechanism of retrotransposon silencing and identified dozens of factors involved in the piRNA pathways. *Drosophila* is an attractive model organism for piRNA studies due to its great availability of genetic tools and the low cost of maintenance. Here, I introduce *Drosophila* genetic resources and techniques valuable for studying piRNA pathway genes via their impact on retrotransposon silencing.

Key words *Drosophila*, piRNAs, *Drosophila* resource, Stock center

1 Introduction

PIWI-interacting RNAs (piRNAs) are a class of single-stranded small RNAs mainly expressed from retrotransposon-enriched genomic loci. piRNAs are associated with PIWI family proteins and guide their complex to silence the expression of retrotransposon mRNAs [1–7]. In the *Drosophila* germline, the RDC complex composed of Rhino, Deadlock, and Cutoff defines noncanonical dual-strand piRNA clusters and ensures long piRNA precursor transcription by preventing splicing and RNA pol-II termination [8, 9]. The resulting piRNA precursors are then exported to the cytoplasm and further processed into mature piRNAs [10–12]. Processed mature piRNAs are loaded onto PIWI family proteins. Furthermore, piRNAs are amplified in a feed-forward loop, termed the secondary processing pathway, by Aubergine and Argonaute3 [13, 14]. In Drosophila gonadal somatic cells, such as follicle cells, piRNAs repress retrotransposons by transcriptional gene silencing, which is mediated by nuclear localized Piwi and piRNAs [15, 16]. Piwi recognizes nascent transcripts complementary to piRNAs and leads to deposition of a repressive chromatin mark,

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_8,

histone H3 lysine 9 trimethylation, onto the target chromatin loci in a collaboration with factors such as histone methyltransferase Eggless [17–26].

Drosophila has been used as a key model organism for identifying factors in piRNA pathways and elucidating molecular mechanisms. There are many reasons why *Drosophila* is so significant for studying piRNA pathways. First, the *Drosophila* genome database is well-established and carefully maintained [27]. Second, *Drosophila* has a small size, short generation time and is easy to cultivate in the laboratory. Finally, powerful genetic tools in *Drosophila* have been established and shared with the research community. By long-lasting efforts of researchers, *Drosophila* strains are collected and maintained by several institutes and centers around the world as living stocks, since cryopreservation and recovery of *Drosophila* are quite difficult. The Bloomington Drosophila Research Center (BDSC), KYOTO stock center (KYOTO), National Institute of Genetics (NIG), and Vienna Drosophila RNAi Center (VDRC) collect and distribute a large number of *Drosophila* strains. Using these genetic resources, researchers can evaluate gene functions and mechanisms in a short period of time without sustained maintenance of the strains.

RNA interference (RNAi) is a powerful tool for elucidating gene function in *Drosophila* [28]. This is achieved by the binary GAL4/UAS system in which GAL4 protein is driven by a tissue-specific promoter, and the expressed GAL4 binds the upstream activation sequences (UAS) and activates a gene downstream of UAS [29]. GAL4 and UAS strains can be maintained as a separate stock. Thus, mating of these strains results in tissue- and/or stage-specific expression of a protein-coding gene or double-stranded (dsRNA) under UAS sequence (Fig. 1). Indeed, screens for factors involved in retrotransposon silencing in germlines or ovarian follicle cells have been widely conducted [30, 31]. Retrotransposon derepression becomes an indication of involvement of genes in the

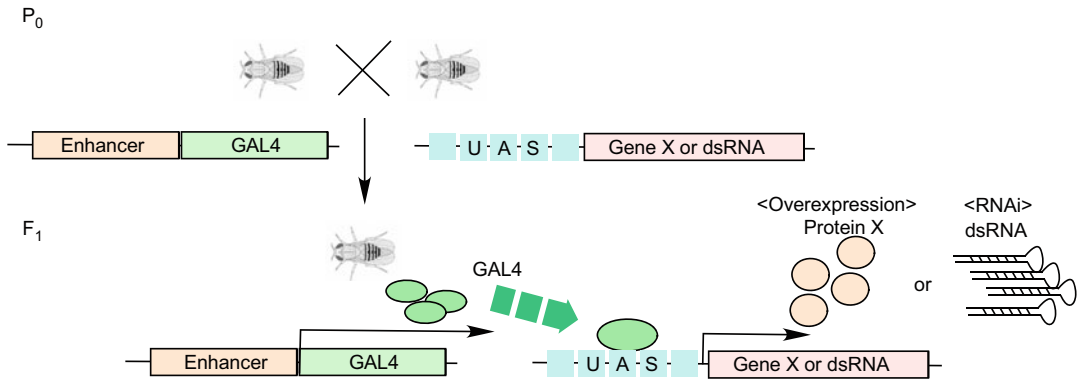


Fig. 1 Schematic drawing of the binary GAL4/UAS expression system

piRNA pathways. More recently, tethering piRNA pathway components on the reporter system in *Drosophila* enabled us to elucidate the mechanisms and establish a genetic hierarchy of piRNA pathway genes in vivo [17, 19]. This reporter system is also available from VDRC. Here, I introduce the techniques related to piRNA researches and advantages of utilization of *Drosophila* genetic resources.

2 Materials

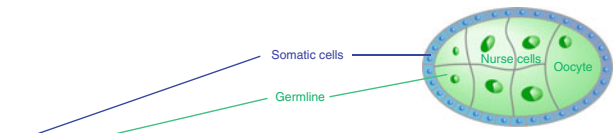
2.1 Drosophila Strains

In germline cells, *nos-GAL4 (BDSC, KYOTO)* and *MTD-GAL4 (BDSC)* are often used for driving GAL4 expression. In ovarian somatic cells, *tj-GAL4* (BDSC) is generally used (Fig. 2). Currently, numerous *Drosophila* strains adapted to express dsRNAs designed against a target gene downstream of UAS are generated and maintained, which covers more than 90% of *Drosophila* genes (Fig. 2). Such strains can be quickly obtained from resource centers (*see* **Note 1**).

Drosophila strains expressing tethering system for the piRNA pathway are distributed from VDRC center and the related plasmid vectors are available from Addgene (Fig. 3).

2.2 Dissection of Drosophila Gonad

1. Tweezers.
2. 1× PBS (without Mg^{2+}/Ca^{2+}), ice-cold.
3. Petri dish 60 mm.
4. Microcentrifuge tubes.



	Strains	Distributing Stock Center		Note
Germline	nos-Gal4	BDSC(#4937)	KYOTO(#107955)	GAL4 is expressed not only in adult GSCs but also differentiating germline cysts
	MTD-Gal4	BDSC(#31777)		3 GAL4 insertions robustly express GAL4 in the germline, unhealthy might be due to multiple VP16 sequences
Gonadal somatic cell	tj-Gal4	KYOTO(#104055)		moderately expressed in cap, escort, follicle stem and follicle cells
	bab1-Gal4	BDSC(#6802)		moderately expressed in terminal filament, cap, escort, stark and polar cells

UAS-RNAi line	KD library	Distributing Stock Center			Note
Long hairpins	VDRC KK and GD	VDRC			Combination with UAS-Dcr2 is required for efficient knockdown in germline
Short hairpins	TRiP Valium 20-22	BDSC	VDRC	NIG	Efficient in germline and weak in somatic cells; UAS-Dcr2 is not required; Lower off-target effects

Improved UAS vector	Distributing Stock Center	Note
pUASz1.0, -MIR, -1.1	DGRC	Well functions in both germline and somatic cells. Researchers don't need to use different constructs (UASt or UASp) for somatic and germline cells

Fig. 2 Genetic resources for RNAi-based analysis of piRNA pathways in *Drosophila* germline and gonadal somatic cells. UAS-TRiP (vallium 20–22) is miRNA-based knockdown system which enabled us to efficiently knockdown in germline and somatic cells without overexpressing Dcr2. VDRC also distributes shRNA-based RNAi lines against piRNA-related genes. UASz is recently developed by Allan Spradling lab [32]. UAST-based constructs function effectively in somatic cells but not in germline. By eliminating piRNA-targeted sequences in UAST vector, UASz-based overexpression shows significantly efficient expression in both germline and somatic cells [32]

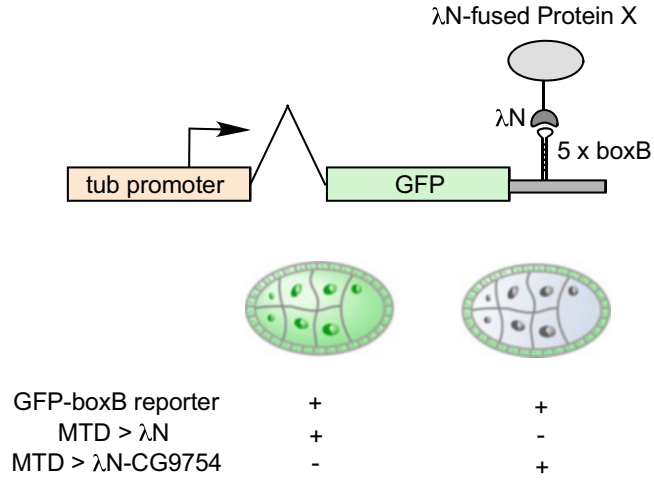


Fig. 3 Tethering reporter system for piRNA pathway elucidation. GFP sensor flies for tethering experiment is distributed from VDRC (#313408). GFP sensor expression is significantly repressed in germline cells when λN-fused CG9754/Silencio is expressed by crossing with MTD-GAL4 strain [17]. This system enabled us to elucidate the mechanisms and establish a genetic hierarchy of piRNA pathway genes in vivo [17, 19]. Plasmid vector for preparing λN-fused protein can be obtained from Addgene (#128011)

5. Stereo Microscope (Leica M80, etc.).
6. CO_2 gas.
7. Diethyl Ether.
8. 50-mL conical tube.
9. Micropipette and tip.

2.3 Measurement of TE mRNAs

1. Pellet pestle.
2. ISOGEN (Nippon Gene).
3. SuperScript IV reverse transcriptase (Invitrogen).
4. Oligo dT primer.
5. Gene-specific primers.
6. Quantitative PCR instrument.

3 Methods

3.1 Drosophila Culture and Mating

1. Anesthetize flies with CO_2 gas and collect virgin females less than 12 h after eclosion.
2. Anesthetize flies with CO_2 gas and collect males.
3. Mix virgin females and males in a same tube and culture them for 4 days at room temperature to facilitate egg laying.
4. Discard the parental flies and continue to culture.

3.2 Dissection of Drosophila *Gonad*

1. Anesthetize flies (*see* **Note 2**) with CO_2 gas and carefully sort them according to the balancer phenotype.
2. Transfer the sleeping flies into a 50-mL tube.
3. Anesthetize flies deeply with diethyl ether in a 50-mL tube and transfer them into 60-mm Petri dish filled with ice-cold 1× PBS.
4. Dissect ovary or testis under a stereo microscope using tweezers.
5. Transfer ovary or testis into a microcentrifuge tube filled with 1× PBS.
6. Continue dissection until the volume of ovaries and testis becomes 50 and 10 μL, respectively.
7. Remove the PBS overlaying the dissected tissues carefully with a pipette tip.
8. Store the isolated ovary and testis in a deep freezer or proceed directly to the next steps.

3.3 Preparation of Total RNAs and RT-qPCR

1. Add 50 μL of ISOGEN and homogenize flies vigorously using a pestle.
2. Add 950 μL of ISOGEN and mix by vortexing for 10 s.
3. Add 200 μL of chloroform and vortex for 15 s.
4. Store for 3 min at room temperature.
5. Centrifuge 12,000 × *g* for 5 min at 4 °C.
6. Collect transparent aqueous phase and add 0.5 mL isopropanol.
7. Precipitate total RNAs and rinse it with 70% Ethanol.
8. Dissolve total RNAs with H_2O (*see* **Note 3**).
9. Treat DNase reaction with 2 μg of total RNAs in 10 μL of reaction volume at 37 °C for 30 min to remove genomic DNA contamination.
10. Heat the DNase-treated samples at 75 °C for 10 min to inactivate DNase.
11. Perform a reverse transcription reaction with SuperScript III according to the manufacturer's instruction using the oligo-dT primer.
12. Perform the quantitative RT-PCR (RT-qPCR) analysis of retrotransposon mRNAs. Relative amount of retrotransposon mRNAs compared to that of a transcript not regulated by piRNAs (e.g., RP49) is calculated.

4 Notes

1. When *Drosophila melanogaster* is exposed to severe heat shock, most animals die. Thus, avoiding shipping during season is recommended to ensure the strains are viable upon arrival.
2. The ovaries of 2- to 4-day-old females should have egg chambers corresponding to various developmental stages.
3. Attempting to extract RNA from an excess amount of ovary tissue decreases the quality of total RNAs. We usually use less than 50 μL of tissue samples for 1 mL of ISOGEN solution.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 18H02379, 21H00236 and 22H02669.

References

1. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10(2):94–108
2. Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136(4): 656–668
3. Khurana JS, Theurkauf W (2010) piRNAs, transposon silencing, and drosophila germline development. J Cell Biol 191(5):905–913
4. Saito K, Siomi MC (2010) Small RNA-mediated quiescence of transposable elements in animals. Dev Cell 19(5):687–697
5. Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12(4): 246–258
6. Iwasaki YW, Siomi MC, Siomi H (2015) PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem 84:405–433
7. Mohn F, Sienski G, Handler D, Brennecke J (2014) The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in drosophila. Cell 157(6):1364–1379
8. Zhang Z, Wang J, Schultz N, Zhang F, Parhad SS, Tu S, Vreven T, Zamore PD, Weng Z, Theurkauf WE (2014) The HP1 homolog rhino anchors a nuclear complex that suppresses piRNA precursor splicing. Cell 157(6):1353–1363
9. ElMaghraby MF, Andersen PR, Pühringer F, Hohmann U, Meixner K, Lendl T, Tirian L, Brennecke J (2019) A heterochromatin-specific RNA export pathway facilitates piRNA production. Cell 178(4):964–979
10. Kneuss E, Munafò M, Eastwood EL, Deumer US, Preall JB, Hannon GJ, Czech B (2019) Specialization of the drosophila nuclear export family protein Nxf3 for piRNA precursor export. Genes Dev 33(17–18):1208–1220
11. Nishimasu H, Ishizu H, Saito K, Fukuhara S, Kamatani MK, Bonnefond L, Matsumoto N, Nishizawa T, Nakanaga K, Aoki J, Ishitani R, Siomi H, Siomi MC, Nureki O (2012) Structure and function of zucchini endoribonuclease in piRNA biogenesis. Nature 491(7423): 284–287
12. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. Cell 128(6):1089–1103
13. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC (2007) A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in drosophila. Science 315(5818): 1587–1590
14. Sienski G, Donertas D, Brennecke J (2012) Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 151: 964–980
15. Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC (2010) Roles for the Yb body components

Armitage and Yb in primary piRNA biogenesis in drosophila. Genes Dev 24:2493–2498

16. Sienski G, Batki J, Senti KA, Donertas D, Tirian L, Meixner K, Brennecke J (2015) Silencio/CG9754 connects the Piwi-piRNA complex to the cellular heterochromatin machinery. Genes Dev 29:2258–2271
17. Ohtani H, Iwasaki YW, Shibuya A, Siomi H, Siomi MC, Saito K (2013) DmGTSF1 is necessary for Piwi-piRISC-mediated transcriptional transposon silencing in the *drosophila* ovary. Genes Dev 27:1656–1661
18. Yu Y, Gu J, Jin Y, Luo Y, Preall JB, Ma J, Czech B, Hannon GJ (2015) Panoramix enforces piRNA-dependent cotranscriptional silencing. Science 350:339–342
19. Osumi K, Sato K, Murano K, Siomi H, Siomi MC (2019) Essential roles of windei and nuclear monoubiquitination of eggless/SETDB1 in transposon silencing. EMBO Rep 20:e48296
20. Ninova M, Chen YA, Godneeva B, Rogers AK, Luo Y, Fejes Tóth K, Aravin AA (2020) Su(var)2-10 and the SUMO pathway link piRNA-guided target recognition to chromatin silencing. Mol Cell 77:556–570
21. Murano K, Iwasaki YW, Ishizu H, Mashiko A, Shibuya A, Kondo S, Adachi S, Suzuki S, Saito K, Natsume T, Siomi MC, Siomi H (2019) Nuclear RNA export factor variant initiates piRNA-guided co-transcriptional silencing. EMBO J 38:e102870
22. Batki J, Schnabl J, Wang J, Handler D, Andreev VI, Stieger CE, Novatchkova M, Lampersberger L, Kauneckaite K, Xie W, Mechtler K, Patel DJ, Brennecke J (2019) The nascent RNA binding complex SFiNX licenses piRNA-guided heterochromatin formation. Nat Struct Mol Biol 26:720–731
23. Fabry MH, Ciabrelli F, Munafò M, Eastwood EL, Kneuss E, Falciatori I, Falconio FA, Hannon GJ, Czech B (2019) piRNA-guided co-transcriptional silencing coopts nuclear export factors. eLife 8:e47999
24. Zhao K, Cheng S, Miao N, Xu P, Lu X, Zhang Y, Wang M, Ouyang X, Yuan X, Liu W, Lu X, Zhou P, Gu J, Zhang Y, Qiu D, Jin Z, Su C, Peng C, Wang JH, Dong MQ, Wan Y, Ma J, Cheng H, Huang Y, Yu Y (2019) A pandas complex adapted for piRNA-guided transcriptional silencing and heterochromatin formation. Nat Cell Biol 21:1261–1272
25. Eastwood EL, Jara KA, Bronelov S, Munafo M, Frantzis V, Kneuss E, Barbar EJ, Czech B, Hannon GJ (2021) Dimerisation of the PICTS complex via LC8/Cut-up drives co-transcriptional transposon silencing in Drosophila. eLife 10:e65557
26. Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, Tabone CJ, Thurmond J, FlyBase Consortium (2021) FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res 49(D1):D899–D907
27. Mohr S, Bakal C, Perrimon N (2010) Genomic screening with RNAi: results and challenges. Annu Rev Biochem 79:37–64
28. Elliott DA, Brand AH (2008) The GAL4 system : a versatile system for the expression of genes. Methods Mol Biol 420:79–95
29. Handler D, Meixner K, Pizka M, Lauss K, Schmied C, Gruber FS, Brennecke J (2013) The genetic makeup of the *drosophila* piRNA pathway. Mol Cell 50:762–777
30. Czech B, Preall JB, McGinn J, Hannon GJ (2013) A transcriptome-wide RNAi screen in the drosophila ovary reveals factors of the germline piRNA pathway. Mol Cell 50:749–761
31. Muerdter F, Guzzardo PM, Gillis J, Luo Y, Yu Y, Chen C, Fekete R, Hannon GJ (2013) A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in drosophila. Mol Cell 50:736–748
32. DeLuca SZ, Spradling AC (2018) Efficient expression of genes in the drosophila germline using a UAS promoter free of interference by Hsp70 piRNAs. Genetics 209(2):381–387

Chapter 9

Generation of Stable *Drosophila* Ovarian Somatic Cell Lines Using the *piggyBac* System

Chikara Takeuchi, Kensaku Murano, Mitsuru Ishikawa, Hideyuki Okano, and Yuka W. Iwasaki

Abstract

Transposable elements (TEs) constitute a large proportion of the genome in multiple organisms. Therefore, anti-transposable element machineries are essential to maintain genomic integrity. PIWI-interacting RNAs (piRNAs) are a major force to repress TEs in *Drosophila* ovaries. Ovarian somatic cells (OSC), in which nuclear piRNA regulation is functional, have been used for research on piRNA pathway as a cell culture system to elucidate the molecular mechanisms underlying the piRNA pathway. Analysis of piRNA pathway using a reporter system to monitor the gene regulation or overexpression of specific genes would be a powerful approach. Here, we present the technical protocol to establish stable cell lines using the *piggyBac* system, adopted for OSCs. This easy, consistent, and timesaving protocol may accelerate research on the piRNA pathway.

Key words Ovarian somatic cell, piRNA pathway, PIWI, Stable line, *piggyBac*

1 Introduction

To maintain genome integrity and ensure the stable inheritance of genomic information, organisms require mechanisms to repress the expression and mobility of transposable elements (TEs) [1]. PIWI-interacting RNAs (piRNAs) are a class of small RNAs associated with PIWI family proteins whose well-established function is to silence transposable elements (TEs) [1, 2]. The PIWI-piRNA pathway is essential for repressing expression of TEs by both posttranscriptional and transcriptional gene silencing mechanisms. In a number of organisms, dysfunction of this pathway leads to impaired germline development and infertility. These observations suggest that the PIWI-piRNA pathway acts as a "guardian of the genome" [3].

PIWI-piRNA-mediated transcriptional and posttranscriptional silencing pathways function in *Drosophila* germline cells. Due to the

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_9,

close association of piRNA biogenesis and TE silencing mechanisms, the precise genetic and mechanistic dissection of transcriptional silencing is challenging. OSCs, a stable cell line derived from *Drosophila* ovarian somatic cells, express only the nuclear Piwi; therefore, OSCs are a powerful tool to analyze transcriptional silencing by the Piwi-piRNA pathway [4, 5]. This cell line is capable of both knockdown and overexpression experiments [6], making it a good model system to analyze the molecular mechanism of nuclear Piwi-piRNA regulation.

By tagging appropriate epitopes or fluorescent proteins, modified OSCs can be used for the purification of protein complex, chromatin immunoprecipitation (ChIP)/crosslinking-immunoprecipitation (CLIP) assay, reporter assay, or live imaging. For example, we have previously characterized a Piwi-piRNA factor Nxf2, by using myc-tagged Nxf2-expressing OSC line [7]. This was especially useful since we were unable to generate Nxf2 antibody for immunoprecipitation experiment. With appropriate controls, various biochemical experiments were performed using tagged proteins. Indeed, we could perform CLIP-seq experiment of myc-tagged Nxf2. In this sense, generation of stable cell lines accelerated Piwi-piRNA research. Meanwhile, it has been difficult to obtain stable lines from OSCs by using traditional vector transfection (e.g., a plasmid expressing the protein of interest as well as a drug resistance cassette) followed by drug screening. Therefore, we utilized the *piggyBac* system for the first time in OSCs to easily obtain stable lines.

PiggyBac (originally called IFP2) is one of the most active DNA transposons, which was found in *Trichoplusia ni* (Cabbage looper of moth) [8, 9]. *PiggyBac* transposase (PBase) recognizes inverted terminal repeats (ITRs) and cleaves at a TTAA motif to form a hairpin structure at both ends with footprint-free removal [10]. The excised fragment can then be integrated again into another TTAA motif in the host genome by PBase (Fig. 1a). By harnessing this mechanism, *piggyBac* transposon was engineered to introduce transgenes [9], and this system is used frequently for mammalian or insect culture cells [11]. In addition, hyperactive PBase (hyPBase), a more active form of PBase engineered by random mutagenesis was reported [12]. The hyPBase improves efficiency of successful stable culture cell line generation.

In this chapter, we provide an easy, consistent, and time-saving (average 2 weeks) protocol to establish a stable cell line by using *piggyBac* (hyPBase) system in OSC (Fig. 1b).

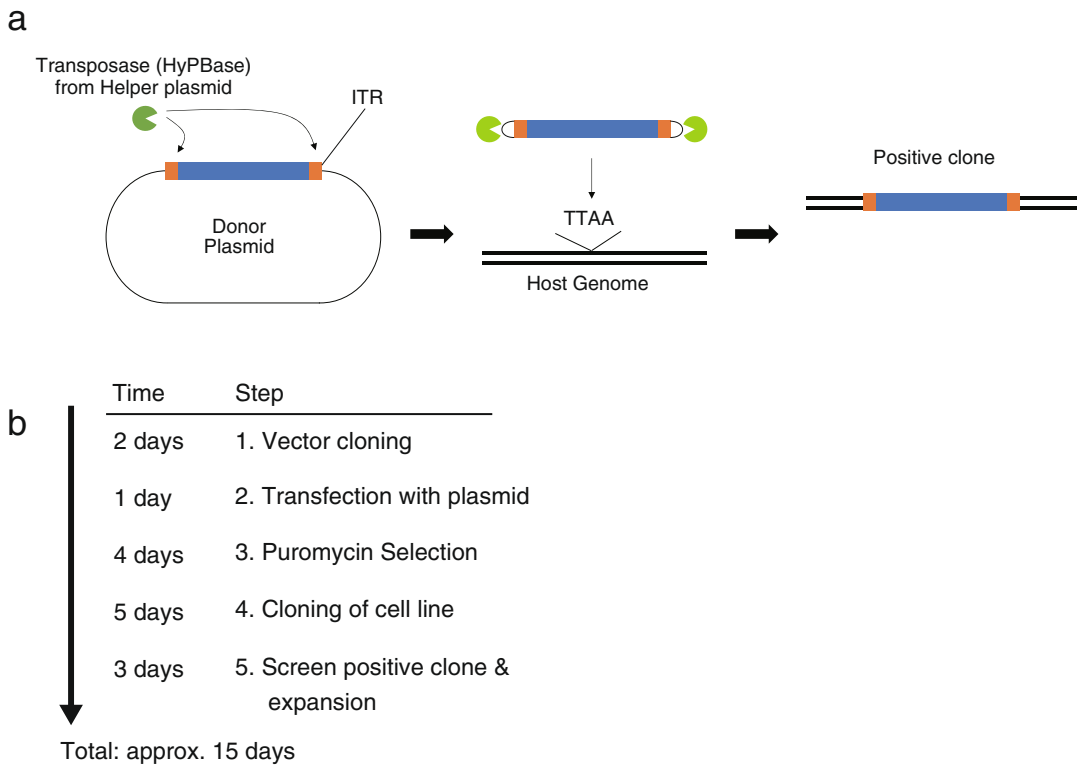


Fig. 1 Overview of stable OSC line generation by using *piggyBac* system. (**a**) Donor plasmid contains the transgene sequence flanked by two inverted terminal repeat (ITR) sequences. *PiggyBac* transposase (PBase) expressed by helper plasmid cleaves DNA sequence by recognizing ITR sequence and forms hairpin structure at both ends of excised fragment. Excised sequence is inserted into host genome randomly at TTAA motif region. (**b**) Flow of generation of stable OSC line using *piggyBac* system. The gene of interest is cloned into a donor plasmid (2 days), and the donor and helper plasmids are transfected into OSC (1 day). Cells without transgene are eliminated using puromycin selection (4 days). Monoclonal cells are isolated with diluted culture on a 10 cm plate (5 days). Positive clones are confirmed by appropriate method (3 days). Collectively, total time to obtain the monoclonal stable OSC is approximately 15 days

2 Materials

2.1 Preparation of PiggyBac Plasmids

1. QIAquick PCR Purification Kit.
2. EcoRI-HF restriction enzyme.
3. XhoI restriction enzyme.
4. NEBuilder HiFi DNA Assembly Master Mix.
5. QIAprep Spin Miniprep kit.
6. NucleoBond Xtra Midi.
7. DNase-free TE buffer.
8. Competent *E. coli* strain (DH5α or MACH1).
9. Helper plasmid (pHsp70-Myc-HyPBase; *see* **Notes 1** and **2**).

10. Donor plasmid (pPB-AcF-Tjen-PuroR; *see* **Note 3**).
11. Primers for amplifying gene of interest (*see* **Note 4**).

2.2 OSC Cell Culture

1. Ovarian Somatic Cells (OSCs) (Drosophila Genomic Resource Center, Stock Number: 288).
2. Shields & Sang M3 medium (Caisson).
3. Insulin.
4. L-Glutathione reduced.
5. Fetal bovine serum (*see* **Note 5**).
6. Fly Extract (*see* **Note 6**).
7. 0.05% Trypsin–EDTA, phenol red.
8. D-PBS(−) without Ca and Mg.
9. 100-mm tissue culture dishes.
10. 0.2-μm filters.
11. OSC culture medium: Thaw Fly Extract in a 37 °C water bath; centrifuge at max speed for 5 min if debris was visible upon thawing. Sterilize Fly Extract by passing through a 0.2-μm filter. To prepare 450 ml of complete OSC culture medium, supplement 351 ml of M3 medium with 10% FBS (45 ml), 4.5 ml glutathione (60 mg/ml stock solution), 4.5 ml insulin (1 mg/ml stock solution), and 10% sterilized Fly Extract (45 ml). Mix well and filter the complete OSC culture medium by passing through a 0.2-μm filter. OSC culture medium can be stored at 4 °C.

2.3 Transfection and Single Colony Isolation

1. Xfect (Clontech).
2. Puromycin Dihydrochloride Ready Made Solution (SIGMA).
3. Rubber bulb dropper.
4. Sterile P200 pipette tip.
5. Cell Banker 1 (Takara).

3 Methods

3.1 PiggyBac Donor Plasmid Design and Cloning Gene of Interest

Donor plasmid contains two transcription units flanked by 5′ and 3′ *piggyBac* ITR (Fig. 2a, *see* **Note 7**). First transcriptional unit is for expression of the gene of interest, whereas the second unit is for the gene conferring drug resistance. Described below is the example for digesting the *piggyBac* donor plasmid with EcoRI and XhoI restriction enzymes, followed by ligation of the amplified fragment using NEBuilder HiFi DNA Assembly Master Mix.

Hepler plasmid expresses hyPBase using Hsp70 promoter and does not need to be modified for each experiment (Fig. 2b).

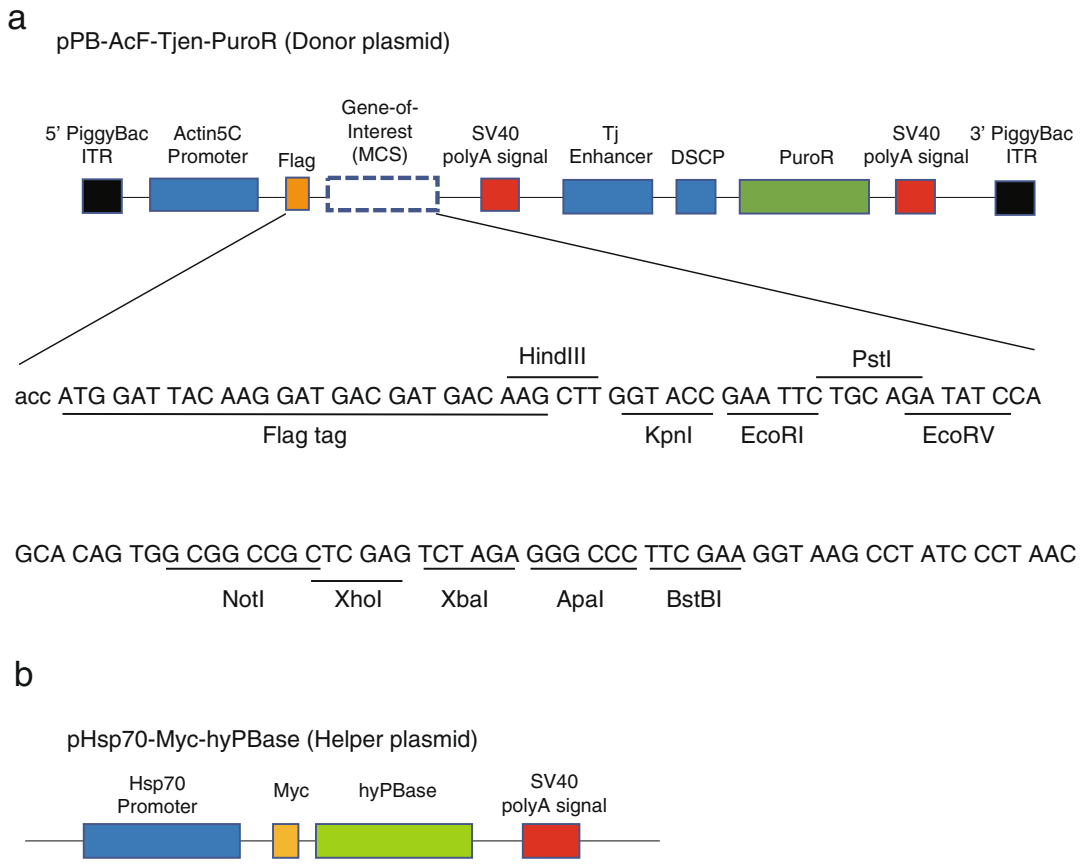


Fig. 2 Construction of *piggyBac* donor and helper plasmid. (**a**) Transgene sequence in donor plasmid (pPB-AcF-Tjen-PuroR) has two transcription units. The first unit is for expression of the gene of interest and contains actin promoter, coding sequence, and SV40 polyA signal. The second unit is for the expression of drug resistance gene and contains Tj enhancer, *Drosophila* synthetic core promoter, puromycin n-acetyltransferase, and SV40 polyA signal. The sequence and restriction site of multiple cloning site is indicated below. ITR: inverted terminal repeat, MCS: multiple cloning site, DSCP: *Drosophila* synthetic core promoter, PuroR: puromycin resistance gene. (**b**) Structure of Helper plasmid (pHsp70-Myc-hyPBase). Myc-tagged hyPBase is transcribed by hsp70 promoter

1. Digest *piggyBac* plasmid using EcoRI and XhoI restriction enzymes with standard protocol and purify the digested plasmid with QIAquick Spin according to the manufacturer's instructions.
2. Amplify fragment by standard PCR protocol (*see* **Note 8**), confirm amplified PCR product by agarose gel electrophoresis, and purify the PCR product with QIAquick Spin.
3. Mix the following reagents.

HiFi DNA Assembly Master Mix	5 μl
Vector fragment (50 ng/μl)	1 μl

(continued)

Insert fragment	x μl (=0.04 pmol, *see* **Note 9**)
Milli-Q	$(4 - x)$ μl
Total	**10 μl**

4. Incubate samples in a thermocycler at 50 °C for 15 min.
5. Transform competent cells of *E. coli* with 2 μl of the assembled product. Culture transformed *E. coli* on ampicillin plate (*see* **Note 10**).
6. Pickup colonies from ampicillin plates, isolate plasmids using plasmid isolation kit (e.g., QIAprep Spin Miniprep kit), and confirm the sequence of insertion by Sanger sequencing.
7. Perform large-scale culturing of *E. coli* transformed with donor and helper plasmids, and isolate plasmids using plasmid isolation kit (e.g., NucleoBond Xtra Midi). Quantify the plasmid concentration using a Nanodrop spectrophotometer, and adjust concentration to 1 μg/μl using TE buffer.

3.2 OSC Cell Culture

1. OSCs are cultured at 26 °C, without CO_2. Observe OSCs and split them upon 80–100% confluency. Upon passage, wash the cells with PBS two times, and treat them with 0.05% trypsin/EDTA for 1 min at 37 °C, followed by the addition of OSC culture medium to inactivate trypsin. Split OSCs every ~2 days at a ratio of 1:10.

3.3 Transfection of Plasmids to OSC

1. Seed approximately 1.0×10^6 OSC to six-well plate and culture (*see* **Note 11**).
2. On the next day, confirm that confluency of OSC is approximately 60%.
3. Mix the following regents in 1.5 ml tubes.

<Tube 1>	
Donor plasmid (1 μg/μl)	4 μl (4 μg)
Helper plasmid (1 μg/μl)	2 μl (2 μg)
Xfect reaction regent	94 μl

<Tube 2>	
Xfect polymer	1.2 μl
Xfect reaction buffer	98.8 μl

4. Mix each tube well by vortexing for 10 s at high speed.
5. Add tube 2 solution to tube 1, to a final volume of 200 μl.
6. Mix the tube well by vortexing for 10 s at high speed.

7. Incubate for 10 min at room temperature to allow nanoparticle complexes to form.
8. Remove OSC culture medium by aspirating, wash the cells two times with 2 ml PBS, and replace the medium with 1 ml of M3 medium, to remove serum from the cell culture (*see* **Note 12**).
9. Add the entire 200 µl of nanoparticle complex solution in tube 1 dropwise into the cell culture medium. Rock the plate gently back and forth to mix.
10. Incubate OSC at 26 °C for 4 h (*see* **Note 13**).
11. Remove nanoparticle-containing medium from cells by aspirating, replace with 2 ml of fresh complete OSC culture medium. Incubate OSC at 26 °C for 48 h.

3.4 Puromycin Selection and Obtaining Monoclonal OSC Line

As described previously [13], limited dilution cannot be used to obtain monoclonal cell line since OSC may need a certain number of neighboring cells to proliferate. Due to this reason, OSCs are passed at very low concentration, and colonies are picked up to obtain monoclonal clones.

Here, we describe an example of cloning Flag-mCherry-Lamin into the gene of interest region of donor plasmid and isolating monoclonal OSC clone. By following this protocol, Flag-mCherry-Lamin expressing OSCs are obtained as shown in Fig. 3.

1. Two days later after transfection, passage all of the cells into 2 ml of OSC culture medium containing 4 µg/ml puromycin into another six-well plate (*see* **Note 14**).
2. Culture OSC 48 h at 26 °C.
3. Seed approximately 1.0×10^6 OSC into 2 ml of OSC culture medium containing 4 µg/ml puromycin into a six-well plate.
4. Culture OSC for 48 h at 26 °C.
5. Seed approximately 1×10^5 OSC into 10 ml OSC culture medium containing 4 µg/ml puromycin in a 10-cm dish.
6. Culture OSC for at least 5 days.
7. Set dropper Rubber Bulb to sterilized P200 tip (Fig. 4a) carefully without touching the edge of the tip.
8. Fill a 24-well plate with 500 µl OSC culture medium per well.
9. Under the microscope with 100× magnification, scrape a colony with a tip and aspirate the scraped OSC by releasing the bulb as illustrated (Fig. 4b).
10. Add the scraped OSC colony to the well of 24-well plate.
11. Culture these OSC for 3–4 days. Replace medium with a fresh one every second day.

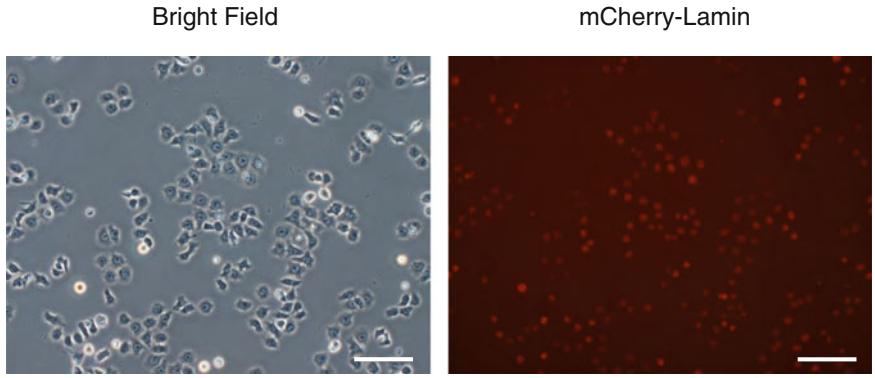


Fig. 3 Example of stable transgene expressing OSC. Fluorescent microscopic image of flag-mCherry-Lamin-expressing OSC. Images were captured under the conditions of brightfield and TRITC filter at 200× magnification. White bars represent 100 μm

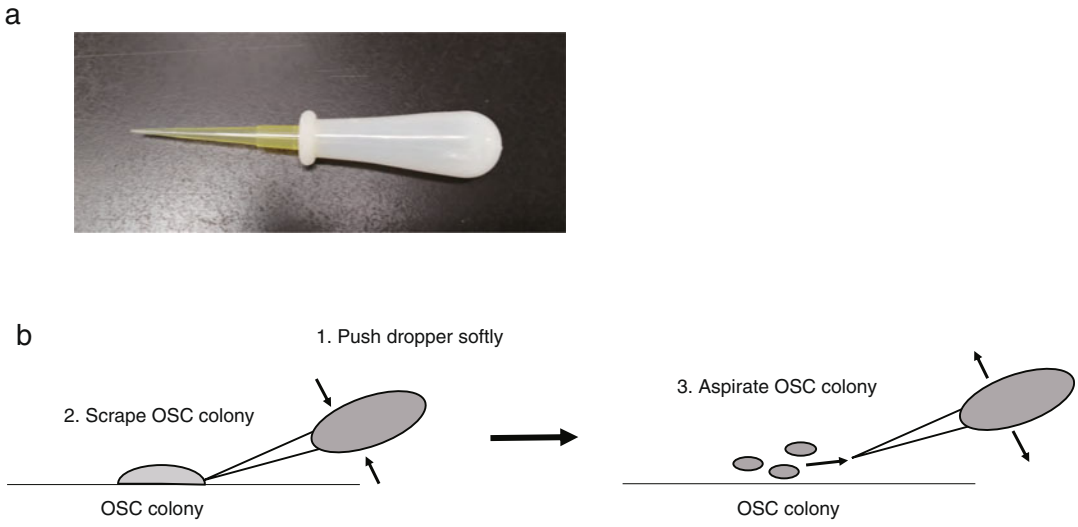


Fig. 4 Procedure to pick up single OSC colony. (**a**) Assembly of yellow tip and dropper rubber bulb. (**b**) Schematic illustration of picking up single OSC colony using a dropper rubber bulb

3.5 Screening for Positive Clones and Cryopreservation of the Cells

1. Dissociate OSC with Trypsin–EDTA and use half the culture for appropriate screening for positive clones and culture the rest of the cells.
2. If the positive cells express fluorescent protein, the screening can be performed using microscopic observation. Otherwise, Western blot screening may be applied.
3. Expand positive clones, according to the result of the screening (*see* **Note 15**). OSC can be stored in liquid nitrogen by using Cellbanker1 cryopreservation medium, according to the manufacturer's instructions.

4 Notes

1. Sequences of plasmids can be downloaded from github (https://github.com/Chikara-Takeuchi/2021_Piggybac_method), and these plasmids can be distributed upon request. For construction of donor plasmid, OSC_Reporter_UAS_traffic jam_BoxB_d2eGFP_t2a_Blast (Addgene plasmid # 128010; http://n2t.net/addgene:128010; RRID: Addgene_128,010) [14] and PB-P(tetO)-hDLX2-pA-floxPG-Kneo-pA [15] were used. For construction of helper plasmid, pCMV-HyPBase-PGK-Puro [15] was used.
2. Hsp70 promoter enhanced the efficiency of stable cell line establishment compared to Actin5C or Ubiquitin promoter. This is because excessive expression of transposase causes degradation of donor plasmid, and therefore weaker expression of transposase by Hsp70 promoter without heat shock is suitable.
3. In addition to puromycin, blasticidin is acceptable for selection. If blasticidin is used for selection, replace puromycin-resistant gene (puromycin n-acetyltransferase) within donor plasmid to blasticidin-S deaminase. Also, culture the cells using 50 μg/ml blasticidin-containing OSC culture medium.
4. For cloning using NEBuilder HiFi DNA Assembly Master Mix, primer to amplify the gene of interest should be designed as indicated below.

   Primer F: CGATGACAAGCTTGGTACCGAATTC*XXXX*.

   Primer R: CTTCGAAGGGCCCTCTAGACTCGAG*XXXX*.

   XXXX is 18–25 bp sequences which correspond to the beginning and the end of the coding sequence (CDS) of the gene of interest. See manufacturers' protocol (https://international.neb.com/products/e2621-nebuilder-hifi-dna-assembly-master-mix) for further information on primer design.
5. Batch check of fetal bovine serum is necessary. Doubling time of OCSs differ among FBS batches.
6. Fly extract is prepared as described previously [6, 13]. In short, 50 g of frozen fly was smashed by mixer with 150 ml M3 medium + 10% FBS, followed by centrifugation at 4,800 × *g* for 10 min at 4 °C. Next, the supernatant was packed in Hybribag (Cosmobio S-1001) and was inactivated by incubation at 60 °C for 5 min. Centrifuge twice at 10,000 × *g* for 15 min at 4 °C. Finally, the supernatant was filtered using a 1.2-μm filter (Sartorius 16555K). This filtered supernatant can be used as a fly extract and can be stored at −20 °C until use.

7. In addition to actin promoter, Tet-On 3G tetracycline-inducible gene expression system (Clontech) can be used in OSC. When using Tet-On 3G system, replace actin promoter of donor plasmid to TRE3G promoter and insert rtTA:P2A sequence between *Drosophila* synthetic core promoter and PuroR gene. For detailed information about Tet-3G system, *see* https://www.takarabio.com/products/inducible-systems/tet-systems-tet-on-3g.
8. Any high-fidelity PCR enzyme, such as KOD One PCR Master Mix (Toyobo KMM-101), can be used. Perform PCR reaction according to the manufacturer's instructions.
9. DNA molar concentration can be calculated using the following equation: (pmol DNA) = (μg DNA) × 10^6/660/length (bp). For example, 50 ng of 2000 bp DNA fragment is equivalent to approximately 0.04 pmol.
10. Both donor and helper plasmids harbor gene for ampicillin resistance.
11. Electroporation can also be used for plasmid transfection into OSC. In this case, use Lonza Cell Line Nucleofector Kit V (Lonza VCA-1003) and transfect with the T-029 program. Since transfected cells undergo puromycin selection, transfection efficiency does not affect downstream analysis. Efficiency is not very different between Xfect method and electroporation.
12. Although it is not recommended to remove serum from the culture medium in the manufacturer's protocol, removal increases the efficiency of plasmid transfection in case of OSC.
13. Xfect is slightly toxic to OSC, especially for the cells at the center of culture dish. However, no strong effect for the following procedures was observed.
14. Wild-type OSC can be eliminated with 0.125 μg/ml puromycin in OSC culture medium for 60% confluency. However, selection efficiency can be improved at a higher concentration of puromycin (~4 μg/ml), in case negative cells seem to survive.
15. If the "clones" are still polyclonal at this step, the monoclonal screening can be performed again starting from Subheading 3.4.

Acknowledgments

We thank Dr. Haruhiko Siomi for the critical reading of the manuscript. OSC_Reporter_UAS_traffic jam_BoxB_d2eGFP_t2a_Blast, used for construction of donor plasmid, was a kind gift from Dr. Julius Brennecke (Addgene plasmid # 128010; http://n2t.net/

addgene:128010; RRID:Addgene_128010). YWI is supported by funding from JSPS KAKENHI Grant Numbers 22H02547, 21H00259 and 18H02421, JST PRESTO Grant Number JPMJPR20E2. KM is supported by funding from JSPS KAKENHI Grant Number 20H03439.

References

1. Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD (2019) PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet 20(2):89–108. https://doi.org/10.1038/s41576-018-0073-3
2. Iwasaki YW, Siomi MC, Siomi H (2015) PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem 84:405–433. https://doi.org/10.1146/annurev-biochem-060614-034258
3. Senti KA, Brennecke J (2010) The piRNA pathway: a fly's perspective on the guardian of the genome. Trends Genet 26(12):499–509. https://doi.org/10.1016/j.tig.2010.08.007
4. Niki Y, Yamaguchi T, Mahowald AP (2006) Establishment of stable cell lines of Drosophila germ-line stem cells. Proc Natl Acad Sci U S A 103(44):16325–16330. https://doi.org/10.1073/pnas.0607435103
5. Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, Kotani H, Asai K, Siomi H, Siomi MC (2009) A regulatory circuit for piwi by the large Maf gene traffic jam in drosophila. Nature 461(7268):1296–1299. https://doi.org/10.1038/nature08501
6. Saito K (2014) RNAi and overexpression of genes in ovarian somatic cells. Methods Mol Biol 1093:25–33. https://doi.org/10.1007/978-1-62703-694-8_3
7. Murano K, Iwasaki YW, Ishizu H, Mashiko A, Shibuya A, Kondo S, Adachi S, Suzuki S, Saito K, Natsume T, Siomi MC, Siomi H (2019) Nuclear RNA export factor variant initiates piRNA-guided co-transcriptional silencing. EMBO J 38(17):e102870. https://doi.org/10.15252/embj.2019102870
8. Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, Fraser MJ (1989) Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology 172(1):156–169. https://doi.org/10.1016/0042-6822(89)90117-7
9. Fraser MJ, Cary L, Boonvisudhi K, Wang HG (1995) Assay for movement of lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. Virology 211(2):397–407. https://doi.org/10.1006/viro.1995.1422
10. Chen K, Birkinshaw RW, Gurzau AD, Wanigasuriya I, Wang R, Iminitoff M, Sandow JJ, Young SN, Hennessy PJ, Willson TA, Heckmann DA, Webb AI, Blewitt ME, Czabotar PE, Murphy JM (2020) Crystal structure of the hinge domain of Smchd1 reveals its dimerization mode and nucleic acid-binding residues. Sci Signal 13(636):eaaz5599. https://doi.org/10.1126/scisignal.aaz5599
11. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T (2005) Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell 122(3):473–483. https://doi.org/10.1016/j.cell.2005.07.013
12. Yusa K, Zhou L, Li MA, Bradley A, Craig NL (2011) A hyperactive piggyBac transposase for mammalian applications. Proc Natl Acad Sci U S A 108(4):1531–1536. https://doi.org/10.1073/pnas.1008322108
13. Ishizu H, Sumiyoshi T, Siomi MC (2017) Use of the CRISPR-Cas9 system for genome editing in cultured drosophila ovarian somatic cells. Methods 126:186–192. https://doi.org/10.1016/j.ymeth.2017.05.021
14. Batki J, Schnabl J, Wang J, Handler D, Andreev VI, Stieger CE, Novatchkova M, Lampersberger L, Kauneckaite K, Xie W, Mechtler K, Patel DJ, Brennecke J (2019) The nascent RNA binding complex SFiNX licenses piRNA-guided heterochromatin formation. Nat Struct Mol Biol 26(8):720–731. https://doi.org/10.1038/s41594-019-0270-6
15. Ishii T, Ishikawa M, Fujimori K, Maeda T, Kushima I, Arioka Y, Mori D, Nakatake Y, Yamagata B, Nio S, Kato TA, Yang N, Wernig M, Kanba S, Mimura M, Ozaki N, Okano H (2019) In vitro modeling of the bipolar disorder and schizophrenia using patient-derived induced pluripotent stem cells with copy number variations of PCDH15 and RELN. eNeuro 6(5):ENEURO.0403-0418. https://doi.org/10.1523/eneuro.0403-18.2019

Part III

Methods to Study Nuclear Regulation by Other Non-coding RNAs

Chapter 10

Whole-Mount RNA FISH Combined with Immunofluorescence for the Analysis of the Telomeric Ribonucleoproteins in the *Drosophila* Germline

Valeriya Morgunova, Maria M. Sukhova, and Alla Kalmykova

Abstract

The RNA fluorescence in situ hybridization (FISH) technique combined with immunostaining is a powerful method to visualize a specific transcript and a protein of interest simultaneously. Although whole-mount RNA FISH is routinely used to determine RNA intracellular localization, a detailed picture of RNA distribution in complex tissues remains a challenge. The main problem is the various permeability of morphologically different cells within a tissue. We overcome this challenge by developing an approach based on differential permeabilization treatment of tissue specimens. We have tested and optimized conditions for RNA FISH combined with immunofluorescent staining (RNA FISH/IF) to detect the maternal telomeric retrotransposon *HeT-A* RNPs in the *Drosophila* ovaries and syncytial embryos. Methods described here are applicable to a broad variety of biological tissue specimens.

Key words Combined RNA FISH and immunofluorescence, Permeabilization, RNP, Telomere, Retrotransposon *HeT-A*, *Drosophila*, Germline, Oogenesis, Embryogenesis, piRNA pathway

1 Introduction

The study of RNA biogenesis by visualization of transcripts and their protein partners is an essential step in both fundamental and biomedical research. The complex biogenesis of the RNA molecules and their subcellular localization are regulated by RNA-binding proteins. During its lifecycle, RNA localizes in both the nucleus and the cytoplasm as a part of ribonucleoprotein (RNP) complexes. A combination of RNA fluorescent in situ hybridization and immunofluorescence (RNA FISH/IF) visualizes a specific transcript and a protein of interest simultaneously. Although whole-mount RNA FISH is routinely used to determine RNA intracellular localization, a detailed picture of RNA distribution in the tissues with complex morphology remains a challenge. The main problem is the various permeability of morphologically different cells within

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_10,

a tissue. As a result, applying a single permeabilization protocol for a whole tissue sample may allow RNA visualization in a limited number of cells creating ambiguity in data interpretation. We overcome this challenge by developing an approach based on differential permeabilization treatment of tissue specimens. We designed and optimized RNA FISH/IF protocols for the detection of the telomeric retrotransposon *HeT-A* RNPs in the *Drosophila* ovaries and early embryos.

In recent years, emerging evidence suggests that telomeric repeats are transcribed to produce telomeric repeat-containing RNA (TERRA) which plays an essential role in the regulation of telomere homeostasis in different species [1]. Studying the biogenesis and localization of telomeric RNAs can help to unveil their yet unknown functions and cellular targets. *Drosophila* telomeres are composed of LINE (long interspersed nuclear element) retrotransposons, among which *HeT-A* is the most abundant [2]. We study the biogenesis and protein partners of the *Drosophila* telomeric repeat *HeT-A* transcripts during the oogenesis and early development upon telomere dysfunction [3, 4]. Telomere dysfunction caused by the Piwi-interacting RNA (piRNA) loss is accompanied by the telomere repeat overexpression. *HeT-A* overexpression can be used as a readout of the disorders of piRNA pathway and other factors regulating telomere integrity [5]. Moreover, studying the biogenesis of the telomeric transcripts could provide a link between dysfunctional telomeres and various cellular pathways. *HeT-A* transcripts encode the RNA-binding protein *HeT-A* Gag which interacts with *HeT-A* RNA forming the *HeT-A* RNPs. To reveal *HeT-A* RNPs and their partners in different cellular compartments, we apply an approach based on differentiated permeabilization treatments. Here, we present an RNA FISH/IF protocol for the detection of *HeT-A* RNPs on whole-mount *Drosophila* ovaries and syncytial embryos. The *HeT-A* RNPs have proven to be an outstanding model for RNA FISH/IF methodological research owing to their sophisticated biogenesis at these developmental stages [3, 4]. Indeed, *HeT-A* RNPs are detected both in the nucleus and in the cytoplasm at different stages of oogenesis; moreover, nucleic *HeT-A* RNPs have spheric shape whereas cytoplasmic RNPs form large aggregates [3].

Drosophila ovary is a suitable in vivo model to study RNA biogenesis in a complex tissue. *Drosophila* ovaries are composed of ovarioles (Fig. 1a); each of them contains the germarium and the series of developing egg chambers. Germarium contains germ and somatic stem cells and their derivatives. The egg chamber comprises the cluster of the germ cells—one oocyte and 15 nurse cells surrounded by somatic follicle cells. During oogenesis, the egg chamber is processed through 14 stages of oogenesis resulting in the formation of a mature egg. Germarium cells are more accessible to antibodies and RNA probes than egg chamber cells (Fig. 1b).

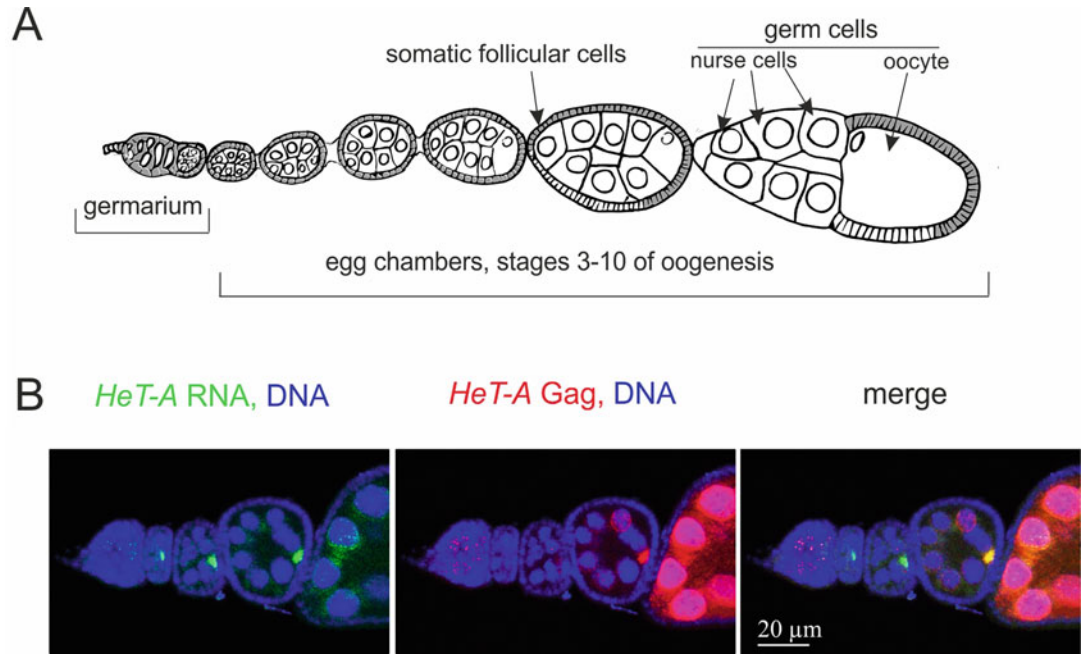


Fig. 1 RNA and protein detection in the *Drosophila* ovaries. (**a**) Scheme of the *Drosophila* ovariole, the structural unit of the ovary. (**b**) *HeT-A* RNA FISH (green) and *HeT-A* Gag immunostaining (red) in the fly ovaries upon piRNA pathway disruption (*spindle-E* knockdown, *spnE*_KD). DNA is stained with DAPI (blue). For permeabilization, ovaries were incubated in 0.6% Triton X-100. Accumulation of *HeT-A* RNA and Gag protein in the oocyte (arrows) is detected only at the early stages of oogenesis. Nonspecific staining at later stages (to the right) indicates that the experimental conditions are not appropriate for these stages and should be optimized

Moreover, secretion of the vitelline membrane proteins at later stages of oogenesis creates a water-impermeable layer resulting in poor access of the reagents to the adjacent cells. Therefore, different permeabilization strategies for RNA and protein visualization during different oogenesis stages are desirable. In addition, specific permeabilization conditions should be determined for the detection of RNPs in the nuclei. This approach generates an integral overview of maternal RNP localization during *Drosophila* oogenesis and embryogenesis before zygotic activation. We encourage researchers to use the experimental tips described here for optimization of preparing whole-mount tissue samples for RNA FISH/IF. The described methods are applicable to a broad variety of biological tissue specimens.

2 Materials

2.1 Riboprobe Labeling

1. Digoxigenin (DIG) RNA labeling mix (Roche).
2. Template DNA (linearized plasmid DNA or PCR product carrying T7 promoter in the appropriate orientation) (*see* **Note 1**).
3. TranscriptAid T7 High Yield Transcription Kit (ThermoFisher) or similar kit for in vitro transcription.
4. RNaseZAP or similar RNase decontamination solution.

2.2 Tissue Preparation and Fixation

1. Dissection buffer: PBS (phosphate-buffered saline), pH 7.4, 0.01% Tween 20, 1× Protease Inhibitor Cocktail (PIC). PIC should be EDTA-free, e.g., cOmplete™, prepared according to the manufacturer's instructions.
2. PBT: 1×PBS, 0.1% Tween 20.
3. 4% formaldehyde solution in 1× PBS (freshly prepared from paraformaldehyde). Adjust pH to 8.0 using sodium hydroxide solution.
4. Methanol.
5. Heptane.
6. Methanol:PBT solutions. Volume-per-volume (v/v) solutions in the following ratios are used for rehydration of methanol-fixed tissues: 7:3, 1:1, and 3:7.
7. Tube rotator.

2.3 Permeabilization

1. 10 mg/ml Proteinase K (Promega), in H_2O solution (store in aliquots at −20 °C).
2. Permeabilization solution: 0.6% Triton X-100 in 1× PBT. Add 1× PIC just before use.
3. 2 mg/ml glycine solution in PBT.

2.4 RNA FISH

1. 1× hybridization buffer (HB): 50% non-ionic formamide, 5× SSC, 0.1% Tween 20, 100 μg/ml Herring Sperm DNA (Promega), 50 μg/ml heparin. Store 2× HB without formamide in aliquots at −20 °C. Before experiment, mix 2× HB (lacking formamide) with deionized formamide (1:1) in a volume that is required for the number of samples to be processed. We recommend deionizing formamide using AG 501-X8 Resin (Bio-Rad) and storing in aliquots at −20 °C.
2. HB:PBT solution: 1:1 v/v solution of HB and PBT.
3. Hybridization wash buffer #1: 2× SSC, 0.1% Tween20 in H_2O.
4. Hybridization wash buffer #2: 0.2× SSC, 0.1% Tween20 in H_2O.

2.5 Immunostaining

1. Image-iT FX signal enhancer (Invitrogen).
2. Ovary blocking buffer: 5% BSA, 0.3% Triton X-100 in PBT.
3. Embryo blocking/staining buffer: 0.3% Triton X-100, 1× PIC in PBT.
4. Ovary staining buffer: 3% BSA, 0.3% Triton X-100, 1× PIC in PBT.
5. Immunostaining wash buffer: 0.3% Triton X-100 in PBT.
6. Anti-Digoxigenin-Fluorescein antibody (Roche) and anti-Fluorescein/Oregon Green, Alexa Fluor 488 Polyclonal Antibody (Invitrogen).
7. Primary antibodies for the detection of proteins of interest.
8. Secondary antibodies conjugated to fluorochromes. We routinely use Alexa 546- and Alexa 647-tagged secondary antibodies with minimal cross-reactivity to IgG from non-target species (Jackson ImmunoResearch).
9. 1 mM DAPI (4′,6-diamidino-2-phenylindole) in H_2O; store in aliquots at −20°C.
10. Vectashield mounting medium (Vector) or similar antifade mounting medium.

3 Methods

3.1 Riboprobe Preparation

At this step, work in RNase-free conditions: treat the surfaces and pipettes with RNaseZAP or similar RNase decontamination solution.

1. Carry out in vitro transcription using DIG RNA labeling mix and the TranscriptAid T7 High Yield Transcription Kit or similar kit according to the manufacturer's instructions.
2. Dissolve RNA riboprobe in 1× HB. The final concentration of DIG-labeled RNA in this solution should be ~100–200 ng/μl. Heat DIG-labeled riboprobe solution at 80 °C for 3 min and place on ice. Stock RNA probe can be stored at −20 °C for a year (*see* **Note 2**).

3.2 Preparation and Fixation of Drosophila Ovaries

Before starting this step, estimate the number of samples you are going to process with different permeabilization protocols. About 20–30 pairs of ovaries per sample are required.

1. Dissect the ovaries from the appropriate number of flies and store in a 1.5-ml Eppendorf tube in the dissection buffer on ice during isolation (up to 2 h). For dissection technique of *Drosophila* ovaries, refer to the detailed protocol described elsewhere [6]. Three-day-old flies fed on yeast for 24 h are used for ovary preparation. It is recommended to tease apart the ovarioles with needles or tweezers during dissection.

2. Wash ovaries with PBT. Add freshly prepared 4% formaldehyde solution to the samples, gently rotate the tubes several times to prevent forming a lump of tissues at the bottom, and place the tubes on their side on the table. Fix ovaries for 20 min at room temperature (RT) without rotation.
3. Wash ovaries with PBT for 5 min three times on a tube rotator. After this step, tissues can be stored in PBT containing 1× PIC at +4 °C overnight.

3.3 Preparation and Fixation of Drosophila 0–2-h-Old Embryos

For *Drosophila* embryo collection and dechorionation, see protocols [7, 8].

1. Wash the dechorionated embryos with 500 μl methanol by gentle rotation several times, remove methanol, then add 500 ml methanol, gently rotate the tubes, place the tubes on their side on the table and incubate 10 min at RT for fixation.
2. Add an equal volume of heptane and incubate 20–30 min on a tube rotator at high speed till most of the devitellinized embryos sink to the bottom of tubes. Remove the upper heptane layer and embryos remaining at the interphase.
3. Rinse settled (devitellinized) embryos with methanol for 5 min at RT.
4. Add methanol. At this point, embryos can be stored at −20 °C in methanol for about 2 weeks.
5. Rehydrate embryos at RT on a rotator with successive 5-min washes in methanol:PBT (7:3), methanol:PBT (1:1), methanol:PBT (3:7), PBT.
6. Re-fix in 4% formaldehyde as described in Subheading 3.2, **steps 2** and **3**.

3.4 Pretreatment

All steps are performed in 1.5-ml Eppendorf tubes at RT on a tube rotator (*see* **Note 3** for tip on using an alternative way of tissue handling).

3.4.1 For the Detection of the Nuclear RNA and Proteins at Mid to Late Stages of Oogenesis (See Note 4)

Figure 2 shows an example of the nuclear RNP detection in the *Drosophila* ovaries.

1. Incubate the appropriate aliquot of tissues with 50 μg/ml proteinase K solution in 1× PBS for 6–8 min (*see* **Note 5**).
2. Remove proteinase K solution and stop the digestion with 2 mg/ml glycine in PBT. Incubate for 2 min.
3. Wash two times for 5 min in PBT.
4. Re-fix tissues during 20 min in 4% PFA in 1× PBS.
5. Wash two times for 5 min in PBT.

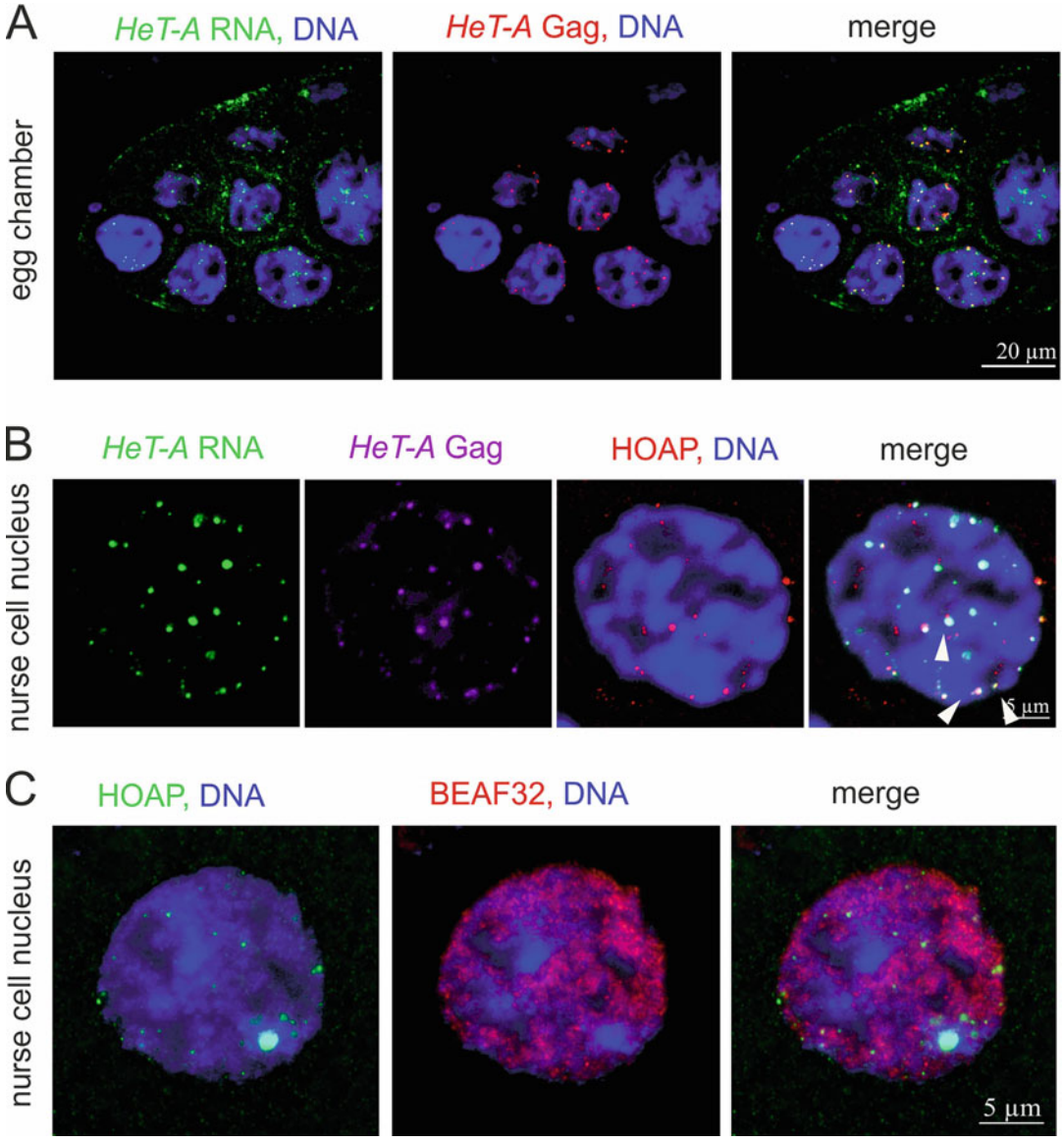


Fig. 2 Detection of the nuclear RNA and proteins at mid-stages of *Drosophila* oogenesis. (**a**) *HeT-A* RNA FISH (green) and *HeT-A* Gag (red) immunostaining of the *Drosophila* ovaries in *spnE*_KD. A fragment of a stage 6 egg chamber is shown. (**b**) *HeT-A* RNA FISH (green) combined with *HeT-A* Gag (magenta) and HOAP (red) immunostaining of the *spnE*_KD ovaries. An enlarged nucleus of the stage 6 germline nurse cell is shown. Telomeric localization of *HeT-A* RNP is shown by arrowheads. (**c**) BEAF32 (red) and HOAP (green) immunostaining in the nurse cell nucleus in the wild type flies. (**a**–**c**) For permeabilization, ovaries were incubated with proteinase K solution during 6 min. DNA is stained with DAPI (blue)

3.4.2 For the Detection of the Cytoplasmic RNA and Proteins at Mid-Stages of Oogenesis (See ***Note 6***)

Figure 3 shows the cytoplasmic RNPs in the *Drosophila* ovaries.

1. Incubate ovaries in permeabilization solution for 20–30 min.
2. Wash ovaries with PBT for 5 min three times.

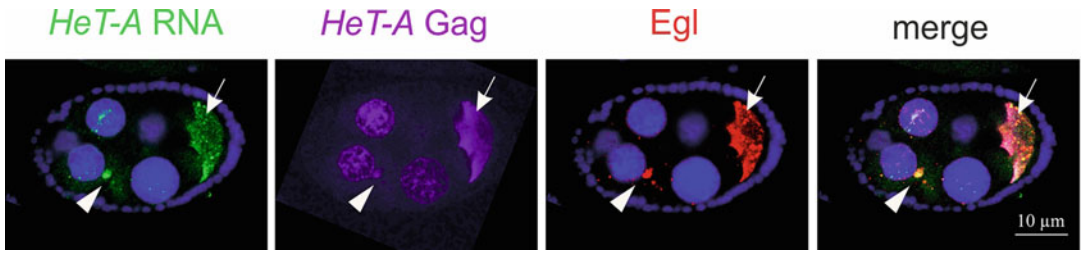


Fig. 3 Detection of the cytoplasmic RNA and proteins at mid-stages of oogenesis. *HeT-A* RNA FISH (green) combined with immunostaining of *HeT-A* Gag (magenta) and Egalitarian (Egl) RNA-binding protein (red) in the mid-stage of oogenesis. Aggregates of *HeT-A* RNP and Egl in the cytoplasm of the nurse cell are shown by arrowhead. Accumulation of *HeT-A* RNPs and Egl is observed in the oocyte (arrow). For permeabilization, ovaries were incubated in 0.6% Triton X-100 during 30 min. DNA is stained with DAPI (blue)

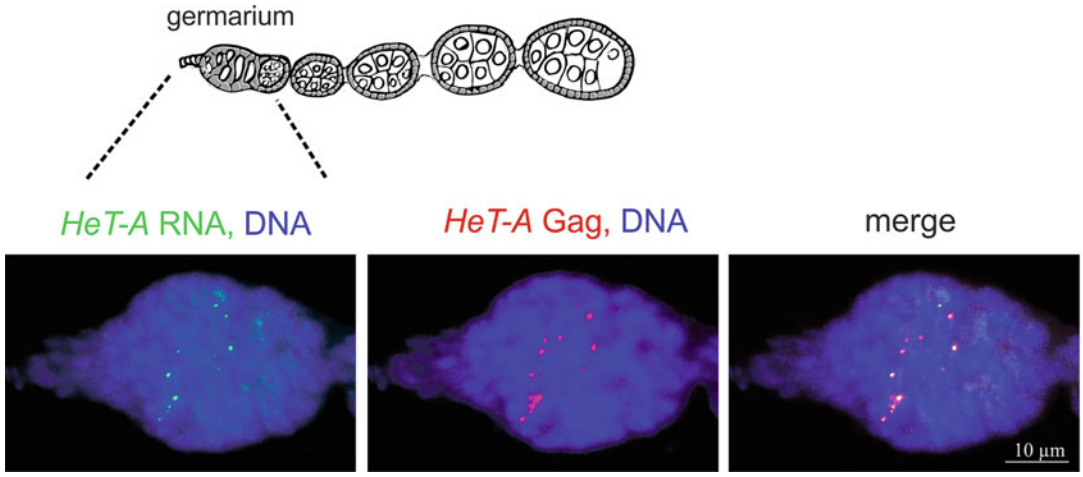


Fig. 4 Detection of the *HeT-A* RNPs in the germarium. Scheme of the *Drosophila* ovariole is shown above. The germarium region corresponding to the confocal images is indicated by dotted lines. *HeT-A* RNA FISH (green) and *HeT-A* Gag immunostaining (red) in the fly ovaries upon piRNA pathway disruption (*spnE*_KD). DNA is stained with DAPI (blue). For permeabilization, ovaries were incubated in 0.6% Triton X-100 during 10 min

*3.4.3 For the Detection of RNA and Proteins at Early Stages of Oogenesis in the Germarium Region (*See *Also the Protocol [9])*

Figure 4 shows nuclear RNPs in the germarium cells of the *Drosophila* ovaries.

1. Incubate ovaries in permeabilization solution for 15 min.
2. Wash ovaries with PBT three times for 5 min each.

3.4.4 For the Detection of RNA and Proteins in 0- to 2-h-Old Drosophila *Embryos*

Figure 5 demonstrates *HeT-A* RNPs around centrosomes in the *Drosophila* syncytium embryos.

1. Incubate embryos in permeabilization solution 0.6% Triton X-100, 1× PBT, 1× PIC for 20 min (*see* **Notes 7** and **8**).
2. Wash ovaries in PBT three times for 5 min each.

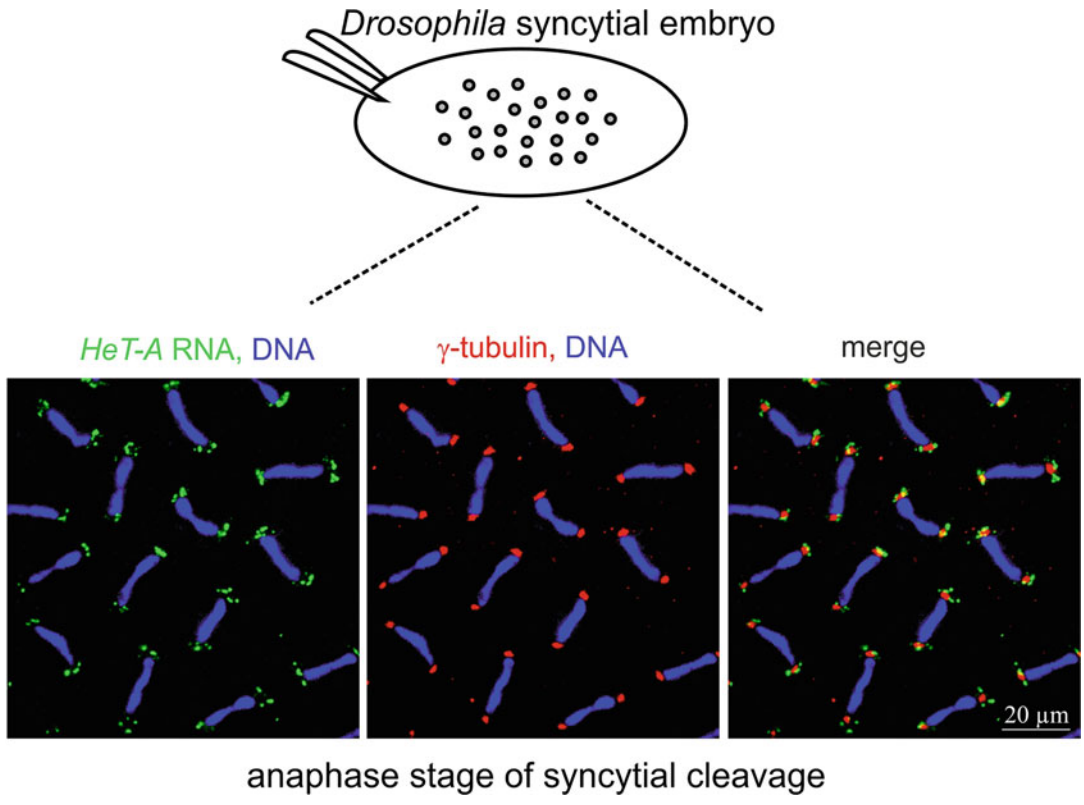


Fig. 5 Detection of the maternal RNA and proteins in the early *Drosophila* embryos. Scheme of the *Drosophila* early embryo is shown above. The syncytium region corresponding to the confocal images is indicated by dotted lines. *HeT-A* RNA FISH (green) and γ-tubulin (red) immunostaining in the syncytial *Drosophila* embryos upon piRNA pathway disruption (*spnE*_KD). DNA is stained with DAPI (blue). For permeabilization, 0- to 2-h old embryos were incubated in 0.6% Triton X-100 during 20 min

3.5 RNA FISH

1. Wash samples in 1 ml of HB:PBT (1:1) solution for 15 min.
2. Preincubate ovaries in HB at 55 °C for 1–3 h.
3. Hybridization is performed in 200–400 μl of HB containing 2.5–10 ng/μl DIG-labeled riboprobe from Subheading 3.1 at 55 °C for 16–18 h.
4. Wash three times for 30 min each in HS at 55 °C.
5. Wash 15 min in HB:PBT (1:1) at 55 °C.
6. Wash two times for 15 min each in 2× SSC, 0.1% Tween20 at 55 °C.
7. Wash two times for 15 min each in 0.2× SSC, 0.1% Tween20 at 55 °C.
8. Wash two times for 15 min each in PBT at RT.

3.6 RNA Detection and Immunostaining

1. Replace PBT with Image-iT FX (Invitrogen) and incubate for 30 min at RT.
2. Wash samples in PBT three times for 5 min each at RT.
3. For **ovary** samples: Block in PBT containing 5% BSA, 0.3% Triton X-100 for 1 h.
   For **embryos**: Incubate in PBT, 0.3% Triton X-100, 1× PIC for 30 min (*see* **Note 8**).
4. Incubate **ovaries** with anti-DIG-FITC antibody and with primary antibodies specific for the protein of interest in PBT containing 3% BSA, 0.3% Triton X-100, 1× PIC at +4°C on a rotating wheel overnight. For **embryos**, use PBT containing 0.3% Triton X-100, 1× PIC.
5. Wash samples with PBT, 0.3% Triton X-100 three times for 5 min each at RT.
6. Wash samples with PBT three times for 5 min each at RT.
7. For **ovary** samples: Block in PBT containing 5% BSA, 0.3% Triton X-100 for 1 h.
   For **embryos**: Incubate in PBT, 0.3% Triton X-100, 1× PIC for 30 min.
8. Incubate ovaries with anti-Fluorescein/Oregon Green, Alexa Fluor 488 Polyclonal Antibody (Invitrogen) and fluorophore-conjugated secondary antibodies in PBT containing 3% BSA, 0.3% Triton X-100, 1× PIC at +4°C on a rotating wheel overnight. For **embryos**, use PBT containing 0.3% Triton X-100, 1× PIC.
9. Wash samples with PBT, 0.3% Triton X-100 two times for 5 min at RT.
10. Wash samples with PBT two times for 5 min each at RT.
11. Incubate in 5 μM DAPI in PBT for 10 min at RT.
12. Wash samples in PBT for 5 min at RT.
13. Pipette ovaries/embryos onto a slide, remove PBT, add 60 μl of Vectashield mounting medium (Vector) and gently arrange the tissues/embryos in the medium by needle.
14. Lower a coverslip onto the tissues. Seal the coverslip to the slide with nail polish. After mounting on a slide, wait 5–24 h before taking images. Keep the slides in the dark at +4 °C during this period.
15. Analyze samples by confocal microscopy.

4 Notes

1. PCR-amplified fragments containing the T7 RNA polymerase promoter, or linearized plasmid containing cloned DNA fragment and RNA polymerase promoters, may be used as templates for the synthesis of riboprobes. Make sure or verify that the T7 RNA polymerase promoter is in a correct orientation relative to the DNA template to produce desirable single-strand RNA probe. Antisense riboprobe will hybridize to the mRNA. The "sense" (coding) sequence is the same as the mRNA.
2. Optimal length of the riboprobe is 100–200 nt. For longer RNA probes, the post-synthesis probe fragmentation by alkaline hydrolysis is recommended. Briefly, add 30 μl of 200 mM Na_2CO_3 and 20 μl of 200 mM $NaHCO_3$ solutions to the 50 μl of aqueous RNA probe solution, mix and incubate at 60 °C. The time of incubation depends on the RNA length and estimates as ~60 min for 500- to 1000-nt long RNAs. Then add 5 μl of 10% acetic acid and 11 μl of 3 M sodium acetate. Precipitate RNA with 2.5 volume of ethanol at −20 °C.
3. Historically, we perform all subsequent procedures using handmade devices, which saves time and allows preserving all tissues in the sample by the end of the lengthy protocol. Tissues are put into handmade polypropylene tubes attached to minimeshes and placed in 24-well culture plates (Fig. 6). In the course of treatment, these mesh tubes are moved (by tweezers) from well to well filled with corresponding solutions. To make mesh tubes, we cut off 2 cm sections of a disposable 2 ml polypropylene pipette, polish the cut surface, melt the rim of the tube over a flame or by soaking in chloroform, and weld to nylon mesh. For details, see similar procedure described elsewhere [10]. Several tubes can be placed in one well. The

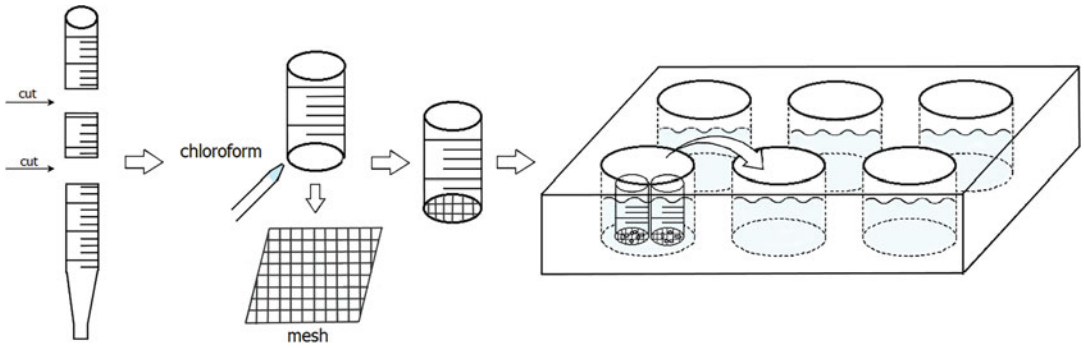


Fig. 6 The handmade mesh tubes can be used as containers for biological samples in the course of multi-step RNA FISH/IF procedure

minimal volume of solution per well is 400 μl. During hybridization, the plate with mini-meshes is placed in a closed container containing wet filter paper to prevent evaporation.

4. Be aware that the cytoplasmic proteins will be removed while the cytoplasmic RNAs may be mislocalized owing to the disrupted protein–RNA links.
5. Proteinase K efficiently removes cytoplasm and nuclear proteins. The time of treatment may depend on proteinase batch. Several time points should be tested to find treatment time producing the best signal for any control nuclear protein. In Fig. 2c, immunostaining of the insulator protein BEAF32 (Developmental Studies Hybridoma Bank) and telomere-specific HOAP protein (home-made antibodies) is shown. Proteinase incubation was 8 min.
6. Be aware that the nuclear RNA and proteins are not accessible if you apply this method. For the detection of the cytoplasmic RNA and proteins at later stages of oogenesis (stages 8–10) proceed as in Subheading 3.4.1, but the incubation time with proteinase K is 2–4 min in this case.
7. Staining of RNA and proteins within the interior portion of the egg during first mitotic divisions might benefit from protease treatment. If you are aiming for such experiment, incubate 0- to 1-h old embryos with 50 μg/ml proteinase K solution in 1× PBS during 2–5 min.
8. It is not recommended to incubate *Drosophila* embryos in BSA-containing solutions. Using BSA or goat serum for blocking of early embryo sample leads to increased nonspecific background that considerably reduces the image quality.

Acknowledgments

We would like to thank Claude Maisonhaute for generously sharing his experience in using handmade tubes for the whole-mount RNA FISH experiments. This work was supported by the Russian Science Foundation (22-14-00006 to A.K.).

References

1. Bettin N, Oss Pegorar C, Cusanelli E (2019) The emerging roles of TERRA in telomere maintenance and genome stability. Cell 8(3): 246. https://doi.org/10.3390/cells8030246
2. Casacuberta E (2017) Drosophila: retrotransposons making up telomeres. Viruses 9(7): 192. https://doi.org/10.3390/v9070192
3. Kordyukova M, Morgunova V, Olovnikov I, Komarov PA, Mironova A, Olenkina OM, Kalmykova A (2018) Subcellular localization and Egl-mediated transport of telomeric retrotransposon HeT-A ribonucleoprotein particles in the drosophila germline and early embryogenesis. PLoS One 13(8):e0201787. https://doi.org/10.1371/journal.pone.0201787
4. Morgunova V, Akulenko N, Radion E, Olovnikov I, Abramov Y, Olenina LV, Shpiz S, Kopytova DV, Georgieva SG, Kalmykova A (2015) Telomeric repeat silencing in germ cells is essential for early development in drosophila. Nucleic Acids Res 43(18): 8762–8773. https://doi.org/10.1093/nar/gkv775
5. Cacchione S, Cenci G, Raffa GD (2020) Silence at the end: how drosophila regulates expression and transposition of Telomeric Retroelements. J Mol Biol 432(15):4305–4321. https://doi.org/10.1016/j.jmb.2020.06.004
6. Wong LC, Schedl P (2006) Dissection of drosophila ovaries. J Vis Exp 1:52
7. Rothwell WF, Sullivan W (2007) Drosophila embryo collection. CSH Protoc 2007:pdb prot4825. https://doi.org/10.1101/pdb.prot4825
8. Rothwell WF, Sullivan W (2007) Drosophila embryo dechorionation. CSH Protoc 2007: pdb prot4826. https://doi.org/10.1101/pdb.prot4826
9. Lie-Jensen A, Haglund K (2016) Antibody staining in drosophila Germaria. Methods Mol Biol 1457:19–33. https://doi.org/10.1007/978-1-4939-3795-0_3
10. Rand MD (2014) A method of permeabilization of drosophila embryos for assays of small molecule activity. J Vis Exp 89:51634. https://doi.org/10.3791/51634

Chapter 11

CRISPR-Mediated Activation of Transposable Elements in Embryonic Stem Cells

Akihiko Sakashita, Masaru Ariura, and Satoshi H. Namekawa

Abstract

Mounting evidence has established that subsets of transposable elements (TEs) function as gene regulatory elements in a cell type- and species-specific manner. Here we describe an in vitro system to ectopically activate TEs using CRISPR-mediated activation (CRISPRa) for functional studies in mouse embryonic stem cells (ESCs). We established a stable mouse CRISPRa ESC line, in which expression of guide RNA enables the activation of TE-derived enhancers and the expression of their adjacent genes. We show an example of ectopic activation of TE-derived enhancers that function in male meiosis, as well as the expression of adjacent germline genes in ESCs. This system can also be applied to functional studies of TEs that are not active in ESCs.

Key words CRISPR activation, Transposable elements, Endogenous retroviruses, Embryonic stem cells, Germ cells, Spermatogenesis, Meiosis

1 Introduction

Approximately half of the mammalian genome is occupied by transposable elements (TEs), which are remnants of ancestral virus infections and horizontal gene transfers [1, 2]. TEs are classified into two main categories: those with long terminal repeats (LTRs), such as endogenous retroviruses (ERVs), and those that lack such repeat sequences, i.e., long interspersed nuclear elements (LINEs). Since spontaneous transposition of TEs poses a threat to host genome stability, they are robustly suppressed by heterochromatic marks in various somatic cells and tissues. The reinvigoration of several TE families is linked to such human diseases as cancer and neurodegenerative and autoimmune disorders [3].

The original version of this chapter was revised. The correction to this chapter is available at https://doi.org/10.1007/978-1-0716-2380-0_23

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_11,

Although TEs were previously thought to be "junk," "selfish," and (or) "parasitic" DNA elements, many recent studies have focused on their "co-opted" functions in the regulation of host genome architecture and gene expression [4, 5]. Many co-opted TEs are capable of modifying gene expression in a spatiotemporal-specific manner by acting as cis-regulatory elements, such as alternative promoters and enhancers. Several TE families carry CTCF-binding motifs, and these TE-derived loop anchoring motifs contribute to the high-order chromosomal structure in a variety of human and mouse cell types [6]. We recently demonstrated that specific types of evolutionarily young ERVs act as enhancers to drive the expression of germline genes in male meiosis, leading to the establishment of species-specific transcriptomes in the mammalian germline [7]. These ERV enhancers carry binding sites of a male germline-specific transcription factor, A-MYB (MYBL1), which activates these ERV enhancers [7]. These studies have collectively demonstrated that cis-regulatory functions of co-opted TEs drive cell type- and species-specific transcriptomes and genome architectures.

Here, we describe a detailed protocol of an in vitro system that ectopically activates TEs using CRISPR-mediated activation (CRISPRa) for functional studies in mouse embryonic stem cells (ESCs). We describe how to establish a stable, mouse ESC line capable of CRISPRa, in which nuclease-null Cas9 (dCAS9) has been fused to a tripartite activator domain VP64-p65-RTA (VPR) [8]. We further describe a method to introduce a multiplex guide RNA (gRNA) expression vector to this cell line via lentiviral gene transfer; this gRNA expression vector targets a consensus sequence of a type of ERV enhancers (interspersed RLTR10B2). Using this model, we can ectopically induce activation of RLTR10B2 enhancers *en-masse* in ESCs [7], mimicking epigenetic gene regulation of the germline in ESCs. Thus, this model serves as a valuable tool to investigate functions of TEs as gene regulatory elements in vitro. Of note, using the above-mentioned ES cells, we can also perform CRISPRa of specific copies of TEs by transient transfection of synthesized single gRNAs [7].

This system can easily be modified with alternative gRNA sequences. For example, this CRISPRa-mediated activation system can be targeted to other genomic sites that enrich TEs, such as pre-pachytene piRNA clusters, the large portion (approximately 80%) of which are derived from TEs [9]. Notably, due to limited access to mammalian embryonic germ cells, the regulatory mechanisms to drive primary piRNA transcripts remain largely unknown. This system can be applied to study the transcriptional network underlying pre-pachytene piRNA clusters in combination with biochemical approaches.

2 Materials

2.1 Cell Lines

1. Mouse embryonic stem cells (ESCs). We have used Wild-type J1 male derived from male agouti 129S4/SvJae embryos as previously described [10], but this protocol is applicable to other ESCs.
2. Human embryonic kidney 293 cells-expressing SV40 large T antigen (HEK293T; ATCC CRL-11268).

2.2 Reagents

2.2.1 Plasmids (See Note 1)

1. PB-TRE-dCas9-VPR (#63800, Addgene).
2. pSpCas9(BB)-2A-Puro (pX459) V2.0 (#62988, Addgene).
3. pMD2.G (#12259, Addgene).
4. psPAX2 (#12260, Addgene).
5. pLV-U6-gRNA-UbC-DsRed-P2A-Bsr (#83919, Addgene).
6. CAG-PiggyBac transposase (pCyL43) plasmid.
7. CMV-*Mybl1* expression plasmid (#MG225161, OriGene).

2.2.2 Media for Cell Culture and Passaging

1. Mouse ESC medium: 15% FBS, 25 mM HEPES, 1× GlutaMAX, 1× MEM non-essential amino acids solution, 1× penicillin/streptomycin, 0.055 mM β-mercaptoethanol in Dulbecco's modified Eagle's medium (DMEM) with high glucose (4.5 g/L), containing 2i (1 μM PD325901, LC Laboratories; and 3 μM CHIR99021, LC Laboratories) and 1300 U/mL LIF.
2. MEF medium: 10% FBS, 1× penicillin/streptomycin, 1 mM sodium pyruvate, 1× MEM non-essential amino acids solution, 1× GlutaMAX in DMEM with high glucose (4.5 g/L).
3. 0.25% (w/v) trypsin-EDTA solution (Thermo Fisher Scientific) (*see* **Note 2**).

2.2.3 Other Reagents

1. Lipofectamine 3000 Transfection Reagent (Thermo Fisher Scientific).
2. Phosphate-Buffered Saline (10×, pH 7.4, Thermo Fisher Scientific).
3. Opti-MEM Reduced Serum Medium (Thermo Fisher Scientific).
4. Forskolin (Sigma).
5. 0.01% Poly-L-Lysine Solution (Sigma).
6. Gelatin from porcine skin powder (Sigma).
7. Lenti-X Concentrator (Clontech).
8. Hanks' Balanced Salt Solution, calcium, magnesium (HBSS (+)), no phenol red (Thermo Fisher Scientific).
9. Polybrene Infection/Transfection Reagent (Thermo Fisher Scientific).

10. ViralPlus Transduction Enhancer (ABM).
11. Hygromycin B Gold (InvivoGen).
12. Blasticidin S HCl (Thermo Fisher Scientific).
13. Doxycycline (Dox, Clontech).
14. RNA extraction kit (e.g., RNeasy Mini Kit, QIAGEN).
15. DNA purification kit (e.g., QIAquick PCR Purification Kit, QIAquick Gel Extraction Kit and QIAEX II Gel Extraction Kit, QIAGEN).
16. cDNA synthesis kit (e.g., SuperScript IV First-Strand Synthesis System, Thermo Fisher Scientific).
17. Fast SYBR Green Master Mix (Thermo Fisher Scientific).
18. Tween 20 (Sigma).
19. Tris base (Sigma).
20. Sodium dodecyl sulfate (Sigma).
21. Glycerol (Sigma).
22. β-Mercaptoethanol (Thermo Fisher Scientific).
23. Bromophenol Blue (Sigma).
24. 5× Power Blotter 1-Step™ Transfer Buffer (Thermo Fisher Scientific).
25. CRISPR-Cas9 Antibody, N-Terminus, clone number 7A9-3A3 (NOVUS Biologicals).
26. Anti-Mouse IgG (H + L) Goat IgG Fab' – HRP (IBL).
27. ECL Prime Western Blotting Detection Reagent (MERCK).
28. ECL Western Blotting Detection Reagents (MERCK).
29. KOD Xtreme Hot Start DNA Polymerase (MERCK).
30. Carbenicillin disodium (Sigma).
31. LB Broth (Sigma).
32. LB Broth with agar (Sigma).
33. BbsI-HF (New England BioLab).
34. BstBI (New England BioLab).
35. BsaBI (New England BioLab).
36. Rapid DNA Ligation Kit (MERCK).
37. NEBuilder HiFi DNA Assembly Master Mix (New England BioLab).
38. NEB Turbo Competent E. coli (New England BioLab).
39. Plasmid extraction kit (e.g., QIAprep Spin Miniprep Kit (QIAGEN) and PureLink™ HiPure Plasmid Midiprep Kit, Thermo Fisher Scientific).

2.3 Equipment

1. Standard consumables and equipment for cell culture.
2. Fluorescence microscope (e.g., EVOS M7000 Imaging System, Thermo Fisher Scientific).
3. Fluorescence-activated cell sorting (FACS) instrument: SH800S cell sorter (SONY) equipped with 405-, 488-, 561-, and 638-nm laser was used to excite DsRed. FL3 (617/30) emission detector was used to filter fluorescence.
4. Sony Sorting Chip-100 μm (SONY).
5. General electrophoresis system for agarose and polyacrylamide gel electrophoresis.
6. Gel imaging system (e.g., Amersham Imager 680, GE Healthcare).
7. Power Blotter–Semi-dry Transfer System (Thermo Fisher Scientific).
8. qPCR system (e.g., StepOnePlus Real-Time PCR System, Thermo Fisher Scientific).
9. Lenti-X GoStix Plus (Clontech).
10. Spectrophotometer (e.g., NanoDrop, Thermo Fisher Scientific).
11. PCR Thermal Cycler (e.g., ProFlex PCR System, Thermo Fisher Scientific).

2.4 Software

1. CRISPOR (http://crispor.tefor.net).
2. Basic Local Alignment Search Tool (BLAST, https://blast.ncbi.nlm.nih.gov/Blast.cgi).
3. HISAT2 v2.2.1 (http://daehwankimlab.github.io/hisat2/).
4. Samtools v1.3.1 (http://www.htslib.org/).
5. SUBREAD v2.0.1 (http://subread.sourceforge.net/).
6. R v4.0.3 (https://www.r-project.org/).
7. DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html).
8. IGV v2.9.2 (https://software.broadinstitute.org/software/igv/).

2.5 Reagent Setup

1. 3 mM (1000×) CHIR99021: Dissolve 5 mg of CHIR99021 in 3.58 mL of DMSO. It can be stored at −20 °C.
2. 1 mM (1000×) PD325901: Dissolve 5 mg of PD325901 in 10.37 mL of DMSO. It can be stored at −20 °C.
3. 1 mg/mL (1000×) Dox: Dissolve 100 mg of Dox in 100 mL of ddH_2O and sterilize by filtration. It can be stored at −20 °C.
4. 0.05% (w/v) trypsin-EDTA solution: Dilute 2 mL of 0.25% (w/v) trypsin-EDTA solution with autoclaved PBS (−) to final volume of 10 mL. It can be stored at 4 °C.

5. 0.002% (w/v) poly-L-lysine solution: Dilute 10 mL of 0.01% (w/v) poly-L-lysine solution with autoclaved ddH_2O to a final volume of 50 mL. It can be stored at 4 °C.
6. 0.2% (w/v) gelatin solution: Dissolve 2 g of gelatin in 1 L of ddH_2O and sterilize by autoclave. It can be stored at RT.
7. 50 mg/mL (500×) Carbenicillin: Dissolve 1 g of Carbenicillin disodium in 20 mL of ddH_2O. It can be stored at −20 °C.
8. LB agar plate: Dissolve 40 g of LB Broth with agar in 1 L of ddH_2O and sterilize by autoclave. After cooling to approximately 50 °C, add 2 mL of 50 mg/mL Carbenicillin to LB Broth with agar and dispense into sterile Petri dishes. It can be stored at 4 °C for 1–2 months.
9. LB Broth medium: Dissolve 20 g of LB Broth in 1 L of ddH_2O and sterilize by autoclave. It can be stored at 4 °C.
10. 10 mM Forskolin: Dissolve 10 mg of Forskolin in 2.44 mL of DMSO. It can be stored at −20 °C.
11. 0.1% (v/v) PBS-T: Dilute 1 mL of Tween 20 detergent with autoclaved PBS (−) to final volume of 1 L. It can be stored at 4 °C.
12. 1 M Tris-HCl (pH 6.8): Dissolve 60.6 g of Tris in 400 mL of ddH_2O and adjust pH to 6.8 with hydrochloric acid. Make up to final volume of 500 mL with ddH_2O and sterilize by autoclave. It can be stored at RT.
13. 4× Laemmli SDS sample buffer (250 mM Tris-HCl (pH 6.8), 8% Sodium dodecyl sulfate, 40% Glycerol, 0.4% β-Mercaptoethanol, and 0.04% Bromophenol Blue in ddH_2O): Mix 5 mL of 1 M Tris-HCl (pH 6.8), 1.6 g of Sodium dodecyl sulfate, 8 mL of glycerol, 80 μL of β-Mercaptoethanol, and 8 mg of Bromophenol Blue, and make up to a final volume of 20 mL with ddH_2O. It can be stored at −20 °C.
14. 1× Laemmli SDS sample buffer (62.5 mM Tris-HCl (pH 6.8), 2% Sodium dodecyl sulfate, 10% Glycerol, 0.1% β-Mercaptoethanol and 0.01% Bromophenol Blue in ddH_2O): Mix 250 μL of 4× Laemmli SDS sample buffer and 750 μL of ddH_2O. Prepare freshly before use.

3 Methods

3.1 Generation and Validation of CRISPRa Transgenic ESCs

We have successfully generated CRISPRa ESC lines with the use of the PiggyBac (PB) Transposon vector system. As shown in Fig. 1a, this PB vector/PB-TRE-dCas9-VPR carries dCas9 proteins fused to the tripartite activator domain consisting of VPR, which shows higher activity compared to the dCas9-Synergistic Activation

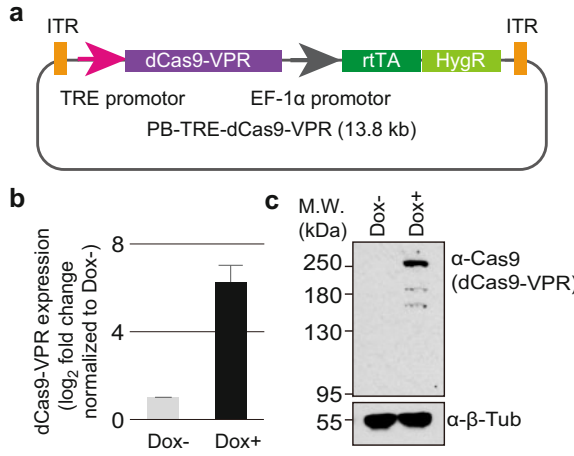


Fig. 1 Generation of CRISPRa transgenic ESC lines. (**a**) Schematic representation of the Dox-inducible dCas9-VPR expression vector used for the generation of CRISPRa ESCs with the PiggyBac transposon system. dCas9-VPR transgene is driven by TRE promoter upon Dox induction. Tripartite activator domain (VP64-p65-Rta (VPR)) was fused to the C-terminal region of dCas9. *ITR* inverted terminal repeat, *TRE* tetracycline responsive element, *rtTA* reverse tetracycline trans-activator, *HygR* hygromycin resistance gene. (**b**) qRT-PCR analysis of dCas9-VPR mRNA level in CRISPRa ESCs, which were cultured in the presence (+) or absence (−) of Dox for 24 h. The relative expression of dCas9-VPR in Dox+ ESCs was obtained based on the ratio of the normalized value of the Dox- cells. Error bar represents mean ± S.E.M. (**c**) Western blotting analysis of dCas9-VPR protein in CRISPRa ESCs, which were cultured in the presence (+) or absence (−) of Dox for 24 h. dCas9-VPR protein was detected with mouse α-Cas9 monoclonal antibody (7A9-3A3, NBP2-36440, NOVUS Biologicals). β-Tubulin (Tub) was used as a loading control

Mediator (SAM) system, controlled by tetracycline responsive element (TRE). Further, EF-1α promoter drives reverse tetracycline trans-activator (rtTA) and hygromycin resistance gene (Fig. 1a). In comparison with other gene delivery systems, such as lentiviral vector, the PB system is the best strategy to integrate the dCas9-VPR transgene into the host genome.

1. Cover the bottom of each well in a six-well plate with a thin layer of 0.2% (w/v) gelatin solution (1 mL for each well). Incubate for 20 min in a CO_2 incubator. During the incubation, harvest exponentially growing ESCs using 0.25% trypsin-EDTA solution. After incubation, aspirate off 0.2% (w/v) gelatin solution and wash the well once with 2 mL of PBS (−). Seed at least three wells: one for PB-TRE-dCas9-VPR plus CAG-PiggyBac transposase (pCyL43), one for donor only, and one for non-transfected control.

Table 1
Lipofectamine 3000 master components for the generation of CRISPRa ESCs.

Tube A	
Component	**Volume**
Lipofectamine 3000 Reagent	7 μL
Opti-MEM reduced serum medium	150 μL

Tube B			
	Volume		
Component	**Donor + transposase**	**Donor only**	**Non-transfected control**
P3000 reagent	6 μL	6 μL	6 μL
PB-TRE-dCas9-VPR (#63800, Addgene)	1.8 μg	1.8 μg	–
CAG-PiggyBac transposase (pCyL43)	800 ng	–	–
Opti-MEM reduced serum medium	150 μL	150 μL	150 μL

2. Seed 2×10^5 of ESCs onto each coated well of the six-well plate; culture overnight with 2 mL of ESC medium.
3. The next day, mix the components from Table 1 in two (A and B) 1.5 mL microcentrifuge tubes.
4. Combine an equal amount of Tube A and Tube B in **step 3** and mix well by pipetting.
5. Incubate at room temperature for 5 min.
6. After incubation, the total volume of the liposome-DNA complex is added dropwise to adherent ESC colonies.
7. At 8–12 h posttransfection, refresh media and allow ESC colonies to expand for 2 days.
8. At 48 h posttransfection, seed the transfected ESCs onto a 100 mm dish coated with 0.2% (w/v) gelatin solution by trypsinization; culture overnight with 10 mL of ESC medium.
9. To select dCas9-VPR integrant-containing ESCs, refresh with media containing 200 μg/mL hygromycin B gold, and allow ESC colonies to expand in ESC media containing 200 μg/mL of hygromycin B gold for 10 days or until all cells in the non-transfected control have died (*see* **Note 3**).
10. After selection with hygromycin B gold, isolate the monoclonal ESC population with the highest expression of dCas9-VPR, pick 12 or more ESC colonies as follows:

(a) To pick colonies easily, seed hygromycin-resistant ESCs onto a 100 mm dish coated with 0.2% (w/v) gelatin solution, at a density of 0.5–2 × 10^4 cells/100 mm dish, and allow ESC colonies to expand for few (2–3) days in ESC medium containing 200 μg/mL of hygromycin B gold.

(b) Cover the bottom of each well of a 96-well flat-bottom plate with a thin layer of 0.2% (w/v) gelatin solution (100 μL for each well) and incubate for 20 min in a CO_2 incubator. After incubation, wash the well once with 100 μL of PBS (−).

(c) Aliquot 20 μL of 0.05% (w/v) trypsin-EDTA solution per well to a 96-well round-bottom plate. Then, carefully pick a healthy ESC colony from "**step (a)**" with a yellow pipette tip and P20 pipette (set to "2 μL") under the stereomicroscope and transfer it into the well of the 96-well round-bottom plate with 0.05% (w/v) trypsin-EDTA solution.

(d) Incubate the cells in a CO_2 incubator for 10–20 min until the colony has de-compacted. (Occasionally tap the plate to ensure de-compaction and initial dissociation).

(e) Add 100 μL of ESC media containing 200 μg/mL of hygromycin B gold to each well, and vigorously pipette up and down 15–20 times using a multichannel pipette to break up colony into a single-cell suspension.

(f) Transfer the single-cell suspension to a 96-well flat-bottom plate coated with 0.2% (w/v) gelatin solution (from "**step (b)**"). Replace media every day and passage expanded colonies as necessary.

11. Once enough clones have been expanded, confirm stable expression of dCas9-VPR integrant-transgene upon addition of Dox (final concentration: 1 μg/mL). Apply several methods such as quantitative RT-PCR analysis (**step (a)**, Fig. 1b) and Western blotting (**step (b)**, Fig. 1c).

(a) Quantitative RT-PCR analysis.

After 24 h of induction with Dox, total RNA is isolated using RNeasy Mini Kit (QIAGEN) with DNase treatment on the column, according to manufacturer's instruction. Reverse transcribe equal volumes of purified RNA with an oligo (dT) primer to synthesize the cDNA using SuperScript IV First-Strand Synthesis System (Thermo Fisher Scientific) according to the manufacturer's instruction. Carry out real-time quantitative PCR using the following cycling conditions: 95 °C for 2 min followed by 40 cycles each of 95 °C for 15 s and 60 °C for 1 min. We used the Fast SYBR Green Master Mix

Table 2
List of primer sequences used in the qRT-PCR analysis

Primer ID	Sequence (5′ → 3′)
Cas9m4_Fw	CATCAGTCAATTACGGGGCTCTA
VPR_Rv	ATCAGCATGTCCAGGTCGAAATC
Hprt_RT_Fw	AGCCCCAAAATGGTTAAGGTTG
Hprt_RT_Rv	TTGCAGATTCAACTTGCGCTCA

(Thermo Fisher Scientific) on a StepOnePlus Real-Time PCR System (Thermo Fisher Scientific). Analyze the levels of expression of a dCas9-VPR using the ΔΔ*C*t method and normalize to the standard internal gene, *Hprt* (Fig. 1b). The specific primer set used for qRT-PCR is listed in Table 2.

(b) Western Blotting Analysis.

After 24 h induction with Dox, directly lyse cells in 200 μl of 1× Laemmli SDS sample buffer and sonicate with Bioruptor at High setting, for 10 cycles each of 30 s with 30 s intervals. Separate total proteins by 7.5% SDS-PAGE and blot on a Protran Nitrocellulose Membranes (0.45 μm pore size, GE Healthcare) via Power Blotter-Semi-dry Transfer System (Thermo Fisher Scientific). Incubate the blotted membrane with the following primary antibody (Mouse α-Cas9 monoclonal antibody: 1/2000, NBP2-36440, NOVUS Biologicals or Mouse α-β-Tubulin monoclonal antibody) overnight at 4 °C. After washing with 0.1% (v/v) PBS-T, apply a horseradish peroxidase (HRP)-coupled goat anti-mouse IgG antibody as a secondary antibody treatment (for 30 min at room temperature with gentle rocking).

After washing with 0.1% (v/v) PBS-T, use ECL Prime (for dCas9-VPR signal) or ECL (for α-β-tubulin signal) Western Blotting Detection Reagent to detect signals (Fig. 1c).

12. Verified clones can then be expanded and used for subsequent experiments and analysis. We termed clone with the highest expression level of dCas9-VPR as "CRISPRa ESCs."

3.2 Design of RLTR10B2-Targeting gRNA and Construction of a gRNA-Expressing Lentiviral Vector

3.2.1 Design the gRNA Oligos and Clone Them into the pSpCas9(BB)-2A-Puro (pX459) V2.0 Plasmid

1. Download the consensus sequence of RLTR10B2 from RepBase website (https://www.girinst.org/). Use an online tool (CRISPOR [11]: http://crispor.tefor.net/) to find optimal gRNA sequences. Select five potential gRNA sequences from the output (Fig. 2a).
2. To identify potential binding/targeting sites for RLTR10B2-targeting gRNA in silico, blast the gRNA target sequence candidate to the mouse reference genome (GRCm38/mm10) using BLASTN with the following commands (Fig. 2b):

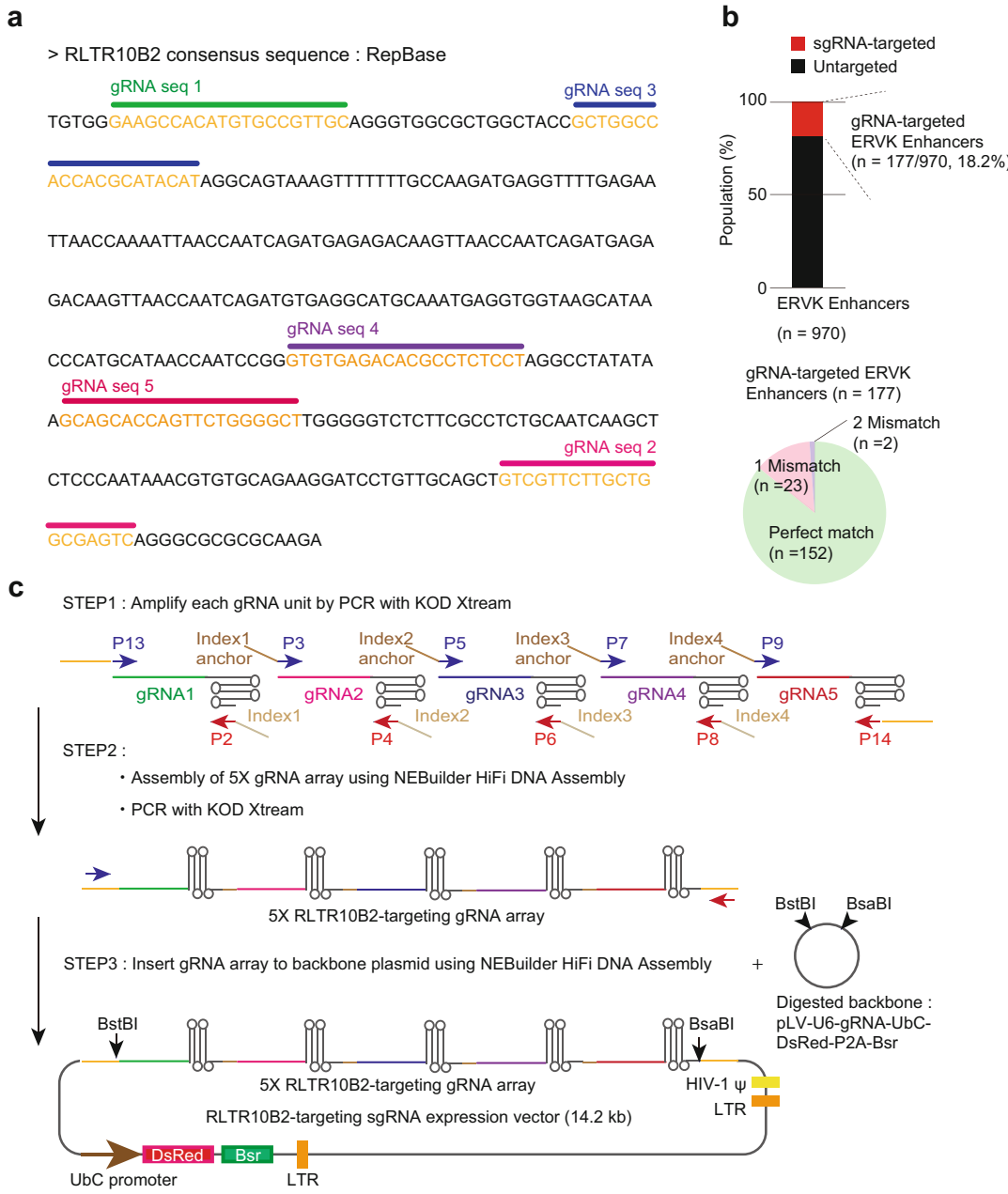


Fig. 2 Design of RLTR10B2-targeting sgRNA and construction of a multiplex gRNA expression vector. (**a**) RLTR10B2 was selected as a target to manifest the feasibility of dCas9-VPR-mediated activation of transposable elements. Individual single-guide (sg) RNAs were designed on the consensus sequence of RLTR10B2 from RepBase (https:/www.girinst.org/repbase/), and highlighted in orange. (**b**) Top; bar chart shows the proportion of targeted (red) and untargeted (black) ERVK enhancer in meiosis by RLTR10B2-targeting gRNA. Bottom; a pie chart indicates the distribution of perfectly matched, 1- and 2-mismatched genomic sequences of ERVK enhancer to 5× RLTR10B2-targeting gRNA array. (**c**) Schematic representations of an assembly method for multiplex sgRNA array and construction of RLTR10B2-targeting sgRNA expression plasmid. STEP1: subcloned transcriptional unit of each gRNA was amplified by PCR, a specific primer set containing 20 nucleotides of homologous sequence at the 5′ end of the reverse (index) and forward (index anchor) primers. The forward and reverse primers (P) hybridize to constant regions (U6 promoter and gRNA

(a) Extract the fasta sequences of ERVK enhancers from the reference genome (mm10.fa) using the bedtools "getFasta" subcommand.

```
bedtools getfasta -s -fi mm10.fa -bed ERVK_Enhancer.bed > ERVK_Enhancer.fa
```

(b) Prepare formatted database files using blastn "makeblastdb" subcommand, which converts a FASTA file into a set of indexed binary files.

```
makeblastdb -in ERVK_Enhancer.fa -dbtype nucl -out ERVK_Enhancer_DB -parse_seqids
```

(c) Align and compare query DNA sequence (gRNA_candidate.fa) with the database of sequence (*ERVK_Enhancer_DB)*.

```
blastn -word_size 14 -outfmt 6 -strand plus -db ERVK_Enhancer_DB -query gRNA_candidate.fa -out {OUTPUT_NAME}.txt
```

3. Order both sense and antisense gRNA oligos for each site with the overhang sequence (Sense: 5′-CACCN…N-3′, Antisense: 3′-N…NCAAA-5′) for cloning into the BbsI site in pSpCas9 (BB)-2A-Puro (pX459) V2.0 gRNA cloning vector (hereafter designated as "pX459"), to any favorite supplier (we usually use IDT) with standard desalting. The specific gRNA oligos used for this study are listed in Table 3.
4. Resuspend each oligo in the final concentration of 100 μM with ddH_2O.
5. Anneal each pair of sense and antisense oligos by mixing them separately as follows (Table 4) and then incubating the mixture in a PCR machine; 95 °C for 5 min, followed by a ramp down to 20 °C at 1 °C per min. Keep at room temperature until the next step.
6. Meanwhile, digest 2.5 μg of pX459 plasmid with BbsI-HF in 1× CutSmart Buffer at least 3 h at 37 °C (*see* **Note 4**).
7. Purify digested pX459 using the QIAquick PCR Purification Kit and elute the linearized products in Buffer EB (or 10 mM Tris-HCl); measure the concentration of recovered plasmid DNA by NanoDrop.

Fig. 2 (continued) scaffold, respectively) of transcriptional units of gRNA. STEP2: amplified transcriptional units of gRNA are mixed and assembled by NEBuilder HiFi DNA Assembly Cloning Kit. Constructed full-length 5× RLTR10B2 gRNA array is amplified by PCR. STEP3: Assembled 5× RLTR10B2-targeting sgRNA array was inserted among the BstBI and BsaBI sites of the linearized pLV-U6-gRNA-UbC-DsRed-P2A-Bsr plasmid

Table 3
List of gRNA oligos used in this study. CACC and AAAC on both ends are overhangs for cloning into BbsI site (highlighted in red)

Oligo ID	Sequence (5'→3')
RLTR10B2_gRNA_1: Sense	caccGAAGCCACATGTGCCGTTGC
RLTR10B2_gRNA_1: Anti_sense	aaacGCAACGGCACATGTGGCTTC
RLTR10B2_gRNA_2: Sense	caccGTCGTTCTTGCTGGCGAGTC
RLTR10B2_gRNA_2: Anti-sense	aaacGACTCGCCAGCAAGAACGAC
RLTR10B2_gRNA_3: Sense	caccGCTGGCCACCACGCATACAT
RLTR10B2_gRNA_3: Anti-sense	aaacATGTATGCGTGGTGGCCAGC
RLTR10B2_gRNA_4: Sense	caccGTGTGAGACACGCCTCTCCT
RLTR10B2_gRNA_4: Anti-sense	aaacAGGAGAGGCGTGTCTCACAC
RLTR10B2_gRNA_5: Sense	caccGCAGCACCAGTTCTGGGGCT
RLTR10B2_gRNA_5: Anti-sense	aaacAGCCCCAGAACTGGTGCTGC

Table 4
Components for annealing gRNA oligos

Component	Volume (μL)
100 μM sense oligo	1
100 μM antisense oligo	1
T4 polynucleotide kinase buffer (10×)	5
ddH_2O	43

8. Set up the following Ligation mixture (Table 5) for each annealed oligo and incubate at room temperature for 30 min (or more) (*see* **Note 5**).
9. Transform chemically competent *E. coli* (NEB Turbo Competent *E. coli* (High Efficiency)) with the ligated plasmid (from **step 7**): Add 2 μL of the ligation reaction mix to 25 μL of NEB Turbo Competent *E. coli* (High Efficiency) and incubate for 20 min on ice. Then, heat-shock for 35 s at 42 °C and immediately place on ice for 2 min. Resuspend *E. coli* in 100 μL of pre-warmed LB medium containing 100 μg/mL Carbenicillin. Plate *E. coli* on a pre-warmed LB agar plate containing 100 μg/mL Carbenicillin and incubate overnight at 37 °C.

Table 5
Components for ligation reaction mix

Component	Volume (μL)
0.2 pmol (1/20 diluted) annealed gRNA oligo	1
50 ng digested pX459	*X*
ddH_2O	$3 - X$
T4 DNA ligation buffer (2×)	5
T4 DNA ligase 5 U/μL	1

10. Pick 4 or more colonies for each annealed oligo and grow overnight in 2 mL of LB medium containing 100 μg/mL Carbenicillin with shaking.
11. Extract plasmid DNA from transformed *E. coli* using QIAprep Spin Miniprep Kit and confirm correct insertions of each gRNA oligo by Sanger sequencing using human U6 sequence primer: 5′-GACTATCATATGCTTACCGT-3′.

3.2.2 Assembly of 5×RLTR10B2-Targeting gRNA Array and Construction of Lentiviral gRNA Expression Vector

As shown in Fig. 2c, we utilized commercially available NEBuilder HiFi DNA Assembly Master Mix (E2621S, New England BioLab). This system enables rapid assembly of gene fragments for expression of multiplexed gRNA array without the use of restriction enzymes (Fig. 2c).

1. Design primer sets that amplify each gRNA transcriptional unit, consisting of a human U6 promoter, gRNA spacer, and scaffold sequences (Fig. 2c: STEP1). The forward and reverse primer sets hybridize to common U6 promoter and gRNA scaffold sequences, respectively, and contain randomized 20 nucleotides of homologous sequence at the 5′ end for NEBuilder assembly (Table 6).
2. As a template, use pX459 containing each gRNA spacer (Fig. 2a: gRNA seq 1–5). Amplify each gRNA transcriptional unit using a specific primer set and KOD Xtreme Hot Start DNA Polymerase (Fig. 2c STEP1, Table 6). PCR is carried out using the following cycling conditions: 95 °C for 2 min followed by 40 cycles each of 98 °C for 15 s and 68 °C for 1 min.
3. Analyze a small aliquot of each PCR by 2% agarose gel electrophoresis and confirm that PCR products are visible as a single specific band after electrophoresis.
4. Purify PCR product using the QIAquick PCR Purification Kit and elute in Buffer EB (or 10 mM Tris-HCl). Measure the concentration of recovered DNA product using NanoDrop.
5. Assemble each PCR-amplified gRNA transcriptional unit by mixing them with NEBuilder HiFi DNA Assembly Master

Table 6
List of primer sequence used in amplification of gRNA transcriptional unit

Primer ID	Sequence (5'→3')
P13 GA cl pLV-U6-gRNA-DsRed_Fw	caccatctttaattgcttcagaaacttcgaaGAGGGCCTATTTCCCATGATTC
P2; sgRNA1_Rv (+Index1)	ttggaaagctcgtcttagacACGCGCTAAAAACGGACTAGC
P3; sgRNA2_Fw (+index1 anchor)	gtctaagacgagctttccaaGAGGGCCTATTTCCCATGATTC
P4; sgRNA2_Rv (+Index2)	gatacttacagctaccactacACGCGCTAAAAACGGACTAGC
P5; sgRNA3_Fw (+index2 anchor)	gtagtggtagctgtaagtatcGAGGGCCTATTTCCCATGATTC
P6; sgRNA3_Rv (+Index3)	ggttagtcaacaatgtgtccaACGCGCTAAAAACGGACTAGC
P7; sgRNA4_Fw (+index3 anchor)	tggacacattgttgactaaccGAGGGCCTATTTCCCATGATTC
P8; sgRNA_Rv (+index4)	aggttactcgcactgttgaaACGCGCTAAAAACGGACTAGC
P9; sgRNA_Fw (+index4 anchor)	ttcaacagtgcgagtaacctGAGGGCCTATTTCCCATGATTC
P14 GA cl to pLV-U6-gRNA-DsRed_Rv	acatgatggtcattttgtaagattagatggaaatcACGCGCTAAAAACGGACTAGC

Mix as follow (Table 7) and incubate the mixture in a PCR machine at 50 °C for 90 min; then leave the unit at 4 °C until the next step (Fig. 2c STEP2).

6. As a template, directly use assembled products. Amplify full-length 5× RLTR10B2-targeting gRNA transcriptional unit (~2500 bp) using a specific primer set and KOD Xtreme Hot Start DNA Polymerase (Fig. 2c STEP2, Table 8). Carry out PCR using the following cycling conditions: 95 °C for 2 min followed by 40 cycles each of 98 °C for 15 s and 68 °C for 3 min.

Table 7
Components for assembly of gRNA transcriptional units

Component	Volume (μL)
50 ng/μL gRNA 1 transcriptional unit	1
50 ng/μL gRNA 2 transcriptional unit	1
50 ng/μL gRNA 3 transcriptional unit	1
50 ng/μL gRNA 4 transcriptional unit	1
50 ng/μL gRNA 5 transcriptional unit	1
ddH_2O	5
NEBuilder HiFi DNA assembly master mix	10

Table 8
List of primer sequence used in the amplification of full-length 5× RLTR10B2-targeting gRNA array (~2500 bp)

Primer ID	Sequence (5′ → 3′)
P15_Fw GA cl pLV-U6-gRNA-DsRed	CACCATCTTTAATTGCTTCAGAAAC
P16_Rv GA cl pLV-U6-gRNA-DsRed	ACATGATGGTCATTTTGTAAGATTAG

7. The total amount of PCR product is separated by 0.9% agarose electrophoresis. Excise the target DNA band near the ~2500 bp with a sharp scalpel under an LED light.
8. Purify PCR product using the Gel Extraction Kit and elute in Buffer EB (or 10 mM Tris-HCl). The concentration of recovered DNA product is measured by NanoDrop.
9. Meanwhile, digest 10 μg of pLV-U6-gRNA-UbC-DsRed-2A-Bsr plasmid with BstBI and BsaBI in 1× CutSmart Buffer. Digestion is carried out using the following cycling condition: 60 °C for 90 min and 65 °C for 90 min, followed by 80 °C for 20 min. Then leave at 4 °C until the next step.
10. The total amount of linearized pLV-U6-gRNA-UbC-DsRed-2A-Bsr plasmid is separated by 0.9% agarose electrophoresis. Excise target DNA band >11 kp with a sharp scalpel under an LED light.
11. Purify digested pLV-U6-gRNA-UbC-DsRed-2A-Bsr plasmid from gel using the QIAEX II Gel Extraction Kit and elute the linearized products in Buffer EB (or 10 mM Tris-HCl). The concentration of recovered plasmid DNA is measured by NanoDrop.

Table 9
Components for assembly of gRNA transcriptional units and linearized pLV-U6-gRNA-UbC-DsRed-2A-Bsr plasmid

Component	Volume (μL)
100 ng/μL linearized pLV-U6-gRNA-UbC-DsRed-2A-Bsr	0.5
5 ng/μL full-length 5× RLTR10B2-targeting gRNA transcriptional unit	6
ddH_2O	3.5
NEBuilder HiFi DNA assembly master mix	10

12. Assemble linearized pLV-U6-gRNA-UbC-DsRed-2A-Bsr plasmid (from **step 11**) and full-length 5× RLTR10B2-targeting gRNA transcriptional unit (from **step 8**) by mixing them with NEBuilder HiFi DNA Assembly Master Mix as follows (Table 9); then incubate the mixture in a PCR machine, 50 °C for 90 min. Leave at 4 °C until the next step (Fig. 2c STEP3).
13. Transform chemically competent *E. coli* (NEB Turbo Competent *E. coli* (High Efficiency)) with assembled 5× RLTR10B2-targeting gRNA-expressing lentiviral vector (from **step 12**): Add 2 μL of assembled product to 25 μL of NEB Turbo Competent *E. coli* (High Efficiency) and incubate for 20 min on ice. Then, heat-shock for 35 s at 42 °C and immediately place on ice for 2 min. Resuspend *E. coli* in 975 μL of pre-warmed SOC medium and incubate at 37 °C for 40 min. Centrifuge at 3000 rpm (800 × *g*) for 3 min and discard the supernatant. Resuspend *E. coli* in 100 μL of pre-warmed LB medium containing 100 μg/mL Carbenicillin. Then plate on a pre-warmed LB agar plate containing 100 μg/mL Carbenicillin and incubate overnight at 37 °C.
14. Pick 12 or more colonies and grow overnight in 2 mL of LB medium containing 100 μg/mL Carbenicillin with shaking.
15. Extract plasmid DNA from transformed *E. coli* using QIAprep Spin Miniprep Kit and confirm correct insertion of gRNA oligo by an analytical digest of plasmid with BstBI and BsaBI and Sanger sequencing using specific primer set as listed in Table 8.

3.3 Production of Lentiviral Particles, Harbor 5× RLTR10B2-Targeting gRNA Array by Transfecting HEK293T

We have successfully generated recombinant lentiviral particles that harbor 5× RLTR10B2-targeting gRNA array by transfecting HEK293T cells with the following plasmids: constructed 5× RLTR10B2-targeting sgRNA expression plasmid (with DsRed reporter and blasticidin S resistance genes), psPAX2 packaging vector, encodes Gag and Pol, and pMD2.G viral envelope expressing vector (**CAUTION**: Before starting experiments with a virus,

Table 10
Lipofectamine 3000 master components for the generation of lentiviral particles

Tube A	
Component	**Volume (μL)**
Lipofectamine 3000 Reagent	14
Opti-MEM reduced serum medium	300

Tube B	
Component	**Volume**
P3000 reagent	12 μL
psPAX2 (#12260, Addgene)	2.26 μg
pMD2.G (#12259, Addgene)	1.48 μg
pLV-U6-5XRLTR10B2-targeting-gRNA-UbC-DsRed-P2A-Bsr (from Subheading 3.2.2, **step 15**)	2.26 μg
Opti-MEM reduced serum medium	300 μL

ensure compliance with the relevant Biosafety office at your Institute, university, and/or government. The entire procedure must be carried out in a BSL2 laboratory. It requires users to undergo proper training involving transfection of the packaging cell line, harvesting viruses, and viral infection.)

1. Cover the bottom of the 60 mm dish with a thin layer of 0.002% (w/v) poly-L-lysine solution (2 mL for a dish) and incubate for 15 min at room temperature. During the incubation, harvest exponentially growing HEK293T cells using 0.25% trypsin-EDTA solution. After incubation, aspirate off 0.002% (w/v) poly-L-lysine solution and wash the well three times with 2 mL of ddH_2O (*see* **Note 6**).
2. Seed 3.8×10^6 of HEK293T cells onto coated 60 mm dish and culture overnight with 4 mL of MEF media.
3. The next day, transfect HEK293T cells with lentivirus producing plasmid using Lipofectamine 3000 reagent. Mix the components from Table 10 in two (A and B) 1.5-mL microcentrifuge tubes (*see* **Note 7**).
4. Combine an equal amount of Tube A and Tube B in **step 3** and mix well by pipetting.
5. Incubate at room temperature for 5 min.
6. After incubation, add the total volume of the liposome-DNA complex dropwise to adherent HEK293T cells.

7. At 24 h posttransfection, replace media with 3 mL of MEF medium containing 10 μM Forskolin.
8. Incubate the transfected cells for 48 h in a CO_2 incubator to produce lentiviral particles.
9. After incubation, carefully collect supernatant containing recombinant lentiviral particles and centrifuge at 12,000 × *g* for 1 min to remove cellular debris.
10. Transfer supernatant to new 15-mL conical tube.
11. By using a small aliquot of supernatant (~20 μL), determine the titer of recombinant lentiviruses with Lenti-X Gostix Plus cassette according to manufacturer's instructions ($\geq 4.6 \times 10^6$ IFU/mL).
12. For three volumes of remaining supernatant (~3 mL), add 1 volume (~1 mL) of Lenti-X Concentrator, and mix by gently pipetting up and down.
13. Incubate the mixture at 4 °C for at least 30 min (or overnight).
14. Centrifuge mixture at 1500 × *g* for 45 min at 4 °C (*see* **Note 8**). An off-white pellet will be visible after centrifuging.
15. Discard the supernatant (*see* **Note 9**).
16. Gently resuspend the pellet in 1/100th (~30 μL) of the original volume using HBSS(+).
17. The virus suspension can be used for immediate infection or stored long-term as a ~10 μL aliquot at −80 °C.

3.4 Generation and Validation of RLTR10B2-Targeting CRISPRa ESCs

To generate RLTR10B2-targeting CRISPRa ESCs, we transduced the CRISPRa ESCs (from Subheading 3.1) with a lentiviral vector containing 5× RLTR10B2-targeting gRNA array. Since the expression cassette of the lentiviral vector also contains Blasticidin-resistant genes and DsRed reporter genes for the selection of positive transformants, the degree of sgRNA expression is evaluated through observations of the red fluorescent reporter protein DsRed. Therefore, to enrich RLTR10B2-targeting CRISPRa ESCs, we used a two-step procedure: positive selection with Blasticidin S (1st) followed by FACS sorting of $DsRed^{High}$ cell population (2nd, Fig. 3).

1. One day before transduction, seed 1×10^6 CRISPRa ESCs (from Subheading 3.1) onto a 60 mm dish coated with 0.2% (w/v) gelatin solution and culture overnight with 4 mL of ESC medium containing 200 μg/mL hygromycin B gold.
2. The next day, replace the entire media with 4 mL of new media containing 10 μL of concentrated lentiviral particle, 8 μg/mL Polybrene and 1/100 diluted ViralPlus transduction enhancer. Replace media every day.

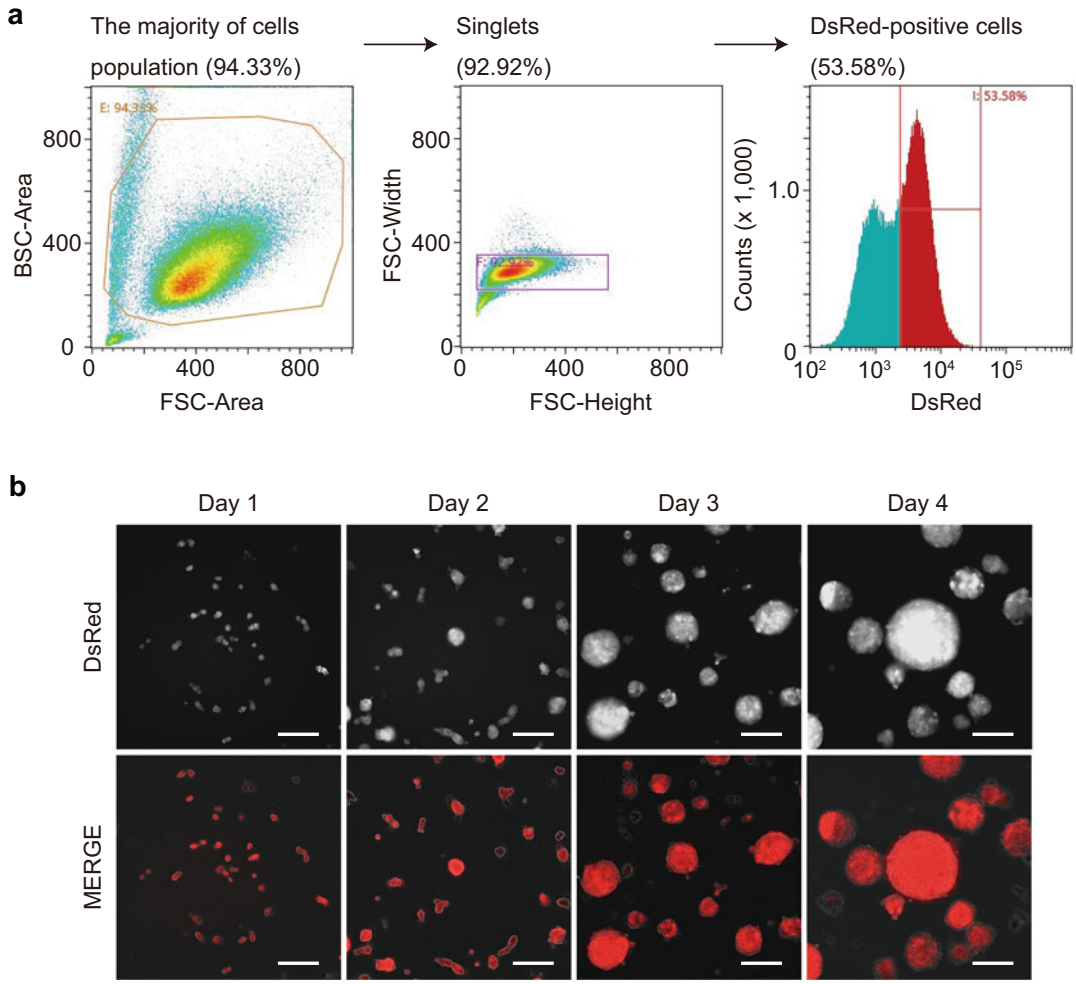


Fig. 3 Isolation of RLTR10B2-targeting CRISPRa ESCs through FACS. (**a**) Flow cytometric analysis of the RLTR10B2-targeting CRISPRa ESCs, based on DsRed and light scattering parameters. The first gate identified the general cell population and excluded debris and aggregates according to forward scatter (FSC) and back scatter (BSC), which are proportional to the cell size and cell granularity. Second, singlets were gated using FSC-Height vs. FSC-Width. Finally, for isolation of RLTR10B2 targeting CRISPRa ES cells, gaiting in the histogram (DsRed-Area vs Counts) was determined by the clear separation between DsRed positive cells and negative cells. (**b**) Representative images of RLTR10B2-targeting CRISPRa ESC colonies from day1 to day4 after cell sorting. Scale bar, 100 μm

3. Two days after transduction, start positive selection with Blasticidin S. the cells are cultivated for 4–6 days in ESC medium containing 200 μg/mL hygromycin B gold and 20 μg/mL Blasticidin S. Replace media every day (*see* **Note 10**).
4. Once colonies of transduced CRISPRa ESC have expanded, harvest cells using 0.25% trypsin-EDTA solution.
5. To remove cell aggregates, pass the cell suspension through a 35 μm nylon mesh cell strainer prior to FACS sorting.

6. Start cell sorting with SH800S cell sorter equipped with 405-, 488-, 561-, and 638-nm laser (Fig. 3a). Data analysis is done using SH800 Software. DsRed can be excited by green-yellow (561-nm) laser, and the emission spectrum is detected in FL3 (617/30) bandpass filter. Forward scatter (FSC) and back scatter (BSC) can be detected using 488 nm laser and FSC and BSC detector channels. Perform cell sorting using a Sony Sorting Chip-100 μm (LE-C3210, SONY) in purity sorting mode. Adjust the flow rate to ~3000 eps. The $DsRed^{High}$ cells are sorted into 1.5-mL tube containing 1 mL of ESC medium supplemented with 200 μg/mL hygromycin B gold and 20 μg/mL Blasticidin S.
7. After sorting, directly seed 3×10^6 $DsRed^{High}$ cells onto 100 mm dish coated with 0.2% (w/v) gelatin solution and cultivate with 10 mL of ESC medium containing 200 μg/mL hygromycin B gold and 20 μg/mL Blasticidin S.
8. Allow ESC colonies to expand for 4 days (Fig. 3b). Replace media every day.
9. We termed the newly established cell line "RLTR10B2-targeting CRISPRa ESCs" and used them for subsequent analysis.

3.5 Functional Evaluation of CRISPRa-Mediated Activation of Spermatogenic ERV Enhancers: RLTR10B2 in ESCs

Upon induction of Dox, dCas9-VPR is guided to interspersed genomic RLTR10B2 loci via multiplexed RLTR10B2 gRNAs and artificially activates them in ESCs, where they are normally repressed by H3K9me3 [12]. Recently, we have demonstrated A-MYB, a male germline-specific transcription factor, binds to a type of ERV enhancer, RLTR10B2, and drives the expression of spermatogenesis-specific transcripts in a meiosis-specific manner [7]. Indeed, expression of genes adjacent to ERV enhancers was synergistically induced by co-induction of A-MYB with RLTR10B2-targeting CRISPRa [7]. The effect of CRISPRa and A-MYB induction can be evaluated through RNA-sequencing analysis.

1. Cover the bottom of each well of a 24-well plate with a thin layer of 0.2% (w/v) gelatin solution (500 μL for each well) and incubate for 20 min in a CO_2 incubator. During the incubation, harvest exponentially growing RLTR10B2-targeting CRISPRa ESCs using 0.25% trypsin-EDTA solution. After incubation, aspirate off 0.2% (w/v) gelatin solution and wash the well once with 1 mL of PBS (−). Prepare at least twelve wells: Three biological replicates each in four groups: (a) cells not treated with Dox for control; (b) cells treated with Dox; (c) cells not treated with Dox and induced A-MYB expression; (d) cells treated with Dox and induced A-MYB expression.
2. Seed 2×10^5 of ESCs onto each coated well of a 24-well plate and culture overnight with 1 mL of ESC medium containing 200 μg/mL hygromycin B gold and 20 μg/mL Blasticidin S.

3. The next day, add Dox (final concentration: 1 μg/mL) to experimental groups (b) and (d).
4. After 24 h of Dox induction, transfect cells in groups (c) and (d) with A-MYB expression vector (CMV-*Mybl1* (MG225161, Addgene) or PGK-*Mybl1* [7]), using Lipofectamine 3000 transfection reagent according to manufacturer's instruction (refer to Subheading 3.1).
5. The following day, directly lyse cells in RLT buffer and extract total RNA using RNeasy Mini Kit (QIAGEN) with DNase treatment on the column, according to manufacturer's instruction. Elute total RNA in Buffer EB (or 10 mM Tris-HCl) and measure concentration by NanoDrop.
6. For our case, total RNAs were submitted to a core facility for library preparation and RNA-sequencing analysis.
7. The effect of CRISPRa and A-MYB-mediated activation at RLTR10B2 loci can be quantified by a comparison of transcriptome differences between each experimental group (a–d).

3.6 RNA-seq Data Processing and Analysis

Raw fastq reads were aligned with the mm10 genome using HISAT2 (version 2.2.1) and converted to bam files using Samtools (version 1.3.1). To quantify aligned reads on respective annotated transcript loci (NCBI RefSeq transcripts), we used the featureCounts function, part of the Subread package (version 2.0.1), with the default setting. To detect differentially expressed genes

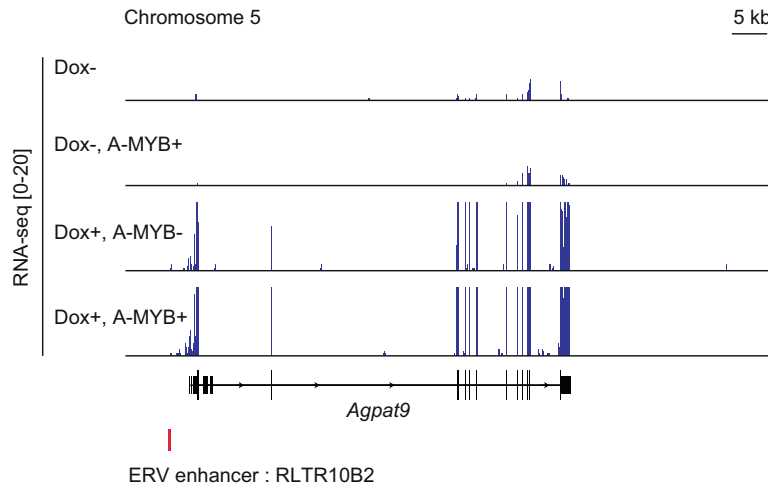


Fig. 4 Multiplexed activation of ERV enhancer by CRISPRa drives robust expression of adjacent genes. The representative track view shows RNA-seq signal in each condition of RLTR10B2-targeting CRISPRa ESCs. An ERV enhancer locus is highlighted in red. Effective activation of RLTR10B2 via CRISPRa system leads to robust expression of *Agpat9*, which is one of the preferentially expressed genes in pachytene spermatocytes

between two biological samples, a read count output file is input to the DESeq2 package (version 1.16.1). Use program functions DESeqDataSetFromMatrix and DESeq to compare each gene expression level between two biological samples. Differentially expressed genes were identified through two criteria: (1) ≥2-fold change and (2) binominal test ($P_{adj} < 0.01$; *P* values were adjusted for multiple testing using the Benjamini-Hochberg correction) in two stages, which are compared. To visualize read enrichments over representative genomic loci, TDF files were created from sorted BAM files using the IGVTools count function (Broad Institute). Figures of continuous tag counts over selected genomic intervals were created in the IGV browse (Fig. 4).

4 Notes

1. PB-TRE-dCas9-VPR (#63800), pSpCas9(BB)-2A-Puro (PX459) V2.0 (#62988), pMD2.G (#12259), psPAX2 (#12260), pLV-U6-gRNA-UbC-DsRed-P2A-Bsr (#83919) plasmids were obtained from Addgene (www.addgene.org). A CAG-PiggyBac transposase (pCyL43) plasmid was provided by the Wellcome Trust Sanger Institute. A CMV-*Mybl1* expression plasmid (MG225161) was obtained from Origene (https://www.origene.com/).
2. The expanded ESC colonies and confluent HEK293T cells were dissociated using 0.25% (w/v) trypsin-EDTA solution (Thermo Fisher Scientific) for passaging.
3. During drug screening, the population of dead cells increases in a dish. Wash out dead cells with PBS (−) prior to exchanging the medium, as necessary, because the growth of living cells is impaired in the presence of a large number of dead cells.
4. Longer or overnight incubation is recommended for complete digestion.
5. 0.05–0.2 pmol of annealed gRNA oligos are used for ligation.
6. Free poly-L-lysine is toxic and should be washed out thoroughly.
7. The transfection was performed at a ratio of 0.377 (sgRNA plasmid):0.377 (psPAX2 vector):0.247 (pMD2.G vector).
8. Use a swinging bucket, not fixed angle rotor. An off-white pellet will be visible after centrifuging.
9. Use a gel loading pipette tip to remove the last few drops of supernatant.
10. During drug-screening, the population of dead cells increases in the dish. Wash out dead cells as necessary with PBS (−) before exchanging the medium, because the growth of living cells are impaired in the presence of a large number of dead cells.

Acknowledgments

We thank Katie Gerhardt for editing the manuscript; Cincinnati Children's Hospital Medical Center for allowing us to develop this method when we were affiliated; and CCHMC Research Flow Cytometry Core for sharing FACS equipment, which is supported by NIH grant S10OD023410. This work was supported by Lalor Foundation Postdoctoral Fellowship and JSPS Overseas Research Fellowship to A.S. and NIH grants GM122776 and GM141085 to S.H.N.

References

1. Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46:21–42. https://doi.org/10.1146/annurev-genet-110711-155621
2. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C (2018) Ten things you should know about transposable elements. Genome Biol 19(1):199. https://doi.org/10.1186/s13059-018-1577-z
3. Payer LM, Burns KH (2019) Transposable elements in human genetic disease. Nat Rev Genet 20(12):760–772. https://doi.org/10.1038/s41576-019-0165-8
4. Thompson PJ, Macfarlan TS, Lorincz MC (2016) Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. Mol Cell 62(5):766–776. https://doi.org/10.1016/j.molcel.2016.03.029
5. Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18(2):71–86. https://doi.org/10.1038/nrg.2016.139
6. Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T (2020) Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol 21(1):16. https://doi.org/10.1186/s13059-019-1916-8
7. Sakashita A, Maezawa S, Takahashi K, Alavattam KG, Yukawa M, Hu YC, Kojima S, Parrish NF, Barski A, Pavlicev M, Namekawa SH (2020) Endogenous retroviruses drive species-specific germline transcriptomes in mammals. Nat Struct Mol Biol 27(10):967–977. https://doi.org/10.1038/s41594-020-0487-4
8. Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, Tuttle M, Iyer EPR, Lin S, Kiani S, Guzman CD, Wiegand DJ, Ter-Ovanesyan D, Braff JL, Davidsohn N, Housden BE, Perrimon N, Weiss R, Aach J, Collins JJ, Church GM (2015) Highly efficient Cas9-mediated transcriptional programming. Nat Methods 12(4):326–328. https://doi.org/10.1038/nmeth.3312
9. Iwasaki YW, Siomi MC, Siomi H (2015) PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem 84:405–433. https://doi.org/10.1146/annurev-biochem-060614-034258
10. Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69(6):915–926. https://doi.org/10.1016/0092-8674(92)90611-f
11. Concordet JP, Haeussler M (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res 46(W1):W242–w245. https://doi.org/10.1093/nar/gky354
12. He J, Fu X, Zhang M, He F, Li W, Abdul MM, Zhou J, Sun L, Chang C, Li Y, Liu H, Wu K, Babarinde IA, Zhuang Q, Loh YH, Chen J, Esteban MA, Hutchins AP (2019) Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. Nat Commun 10(1):34. https://doi.org/10.1038/s41467-018-08006-y

Chapter 12

Method for Evaluating Effects of Non-coding RNAs on Nucleosome Stability

Mariko Dacher, Risa Fujita, Tomoya Kujirai, and Hitoshi Kurumizaka

Abstract

In eukaryotic cells, genomic DNA is stored in the nucleus in a structure called chromatin. The nucleosome, the basic structural unit of chromatin consisting of DNA wound around a histone octamer, regulates access of transcription machinery to DNA. Nucleosome stability is thus tightly associated with gene expression. Recently, a class of non-coding RNAs was found to be directly associated with chromatin. Although these non-coding RNAs are reportedly important in genome regulation, the molecular mechanisms through which these RNAs act remain unclear. Here, we introduce a biochemical method to evaluate the effects of ncRNAs on nucleosome stability, using the breast cancer-associated ncRNA *Eleanor2* as an example. This method is useful for assessing the effects of different RNAs on chromatin stability and conformation.

Key words RNA transcription, Recombinant histones, Nucleosome reconstitution, Thermal stability assay, Chromatin

1 Introduction

In eukaryotic cells, genomic DNA is packaged within the nucleus in a highly compacted structure termed chromatin [1]. The fundamental structural unit of chromatin is the nucleosome core particle, a complex of approximately 146 base pairs of DNA wrapped around a histone octamer consisting of two H2A-H2B dimers and one H3-H4 tetramer [2]. The nucleosome is structurally stable and acts as a repressor of biological events that occur on genomic DNA, such as transcription [3–5].

In organisms across all domains of life, some genomic regions produce RNA transcripts that are not translated into proteins, which are termed non-coding RNAs (ncRNAs) [6]. Although the majority of these ncRNAs have yet to be characterized in depth, numerous transcripts reportedly play key roles in genome organization [7–13]. Among the ncRNAs, those that are localized in the nucleus regulate genomic functions, such as transcription

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_12,

and recombination [14–16]. Over the past decade, extensive studies have revealed that these ncRNAs are involved in the formation of phase-separated condensates, as well as in interactions with epigenetic factors such as chromatin remodelers, histone post-translational modifiers, and telomerase [17–22].

The ncRNAs described in this chapter are clusters of ncRNAs called Eleanors (Estrogen receptor-α (*ESR1*)1 locus enhancing and activating non-coding RNAs), which are highly expressed in long-term estrogen-deprivation (LTED) cells, a breast cancer model [23]. When breast cancer cells that are ER (estrogen receptor-α)-positive are cultured for a long period of time under hormone deprivation conditions, the expression of the *ESR1* gene encoding the ER is upregulated, resulting in LTED cells [24]. In the nucleus of LTED cells, Eleanor RNAs are transcribed from a 0.7 Mb genomic region containing the *ESR1* gene, and accumulate in *cis* in the transcribed chromatin region [23]. As a result, the expression of the *ESR1* gene is highly activated in these LTED cells. However, the molecular mechanism by which the accumulated Eleanor RNAs activate this transcription event remains unclear.

The ncRNAs that influence chromatin regulation, including gene expression, undoubtedly affect chromatin structure. However, the impact of the chromatin-associated ncRNAs on the structure of nucleosomes remains enigmatic. In this article, we introduce a method to analyze the effects of ncRNAs on nucleosome stability, using reconstituted nucleosomes. This method is useful to evaluate the state of nucleosomes in the presence of various types of non-coding RNAs, which may provide an opportunity to gain insight into the link between ncRNAs and chromatin.

2 Materials

2.1 Transcription of RNA In Vitro

1. T7 RiboMAX™ Express Large Scale RNA Production System kit (Promega).
2. Linearized DNA template containing T7 promoter, target sequence, and restriction site that produces blunt ends by a specific restriction enzyme (e.g., *Eco*RV, *Sca*I). This sequence is defined by the protocol user to generate ncRNAs specific to their interest.
3. 5:1 (v/v) Phenol:Chloroform.
4. 100% 2-propanol.
5. 70% ethanol.
6. 24:1 (v/v) Chloroform: Isoamyl alcohol.
7. 3 M CH_3COONa (included in the RiboMAX kit).
8. illustra™ MicroSpin™ G-25 columns or equivalent.

9. Mini Dialysis kit, 8 kDa cut-off, 250 μL or equivalent (Cytiva).
10. SequaGel UreaGel 19:1 Denaturing Gel System or equivalent.
11. Hi-Di formamide (Thermo Fisher).
12. Single-stranded RNA ladder suitable for the length of the RNA product (e.g., Low Range ssRNA Ladder).
13. Sterilized Milli-Q water.

2.2 Nucleosome Reconstitution and Purification

1. Dialysis membrane (MWCO 6000–8000 Da).
2. HiLoad 16/600 Superdex 200 prep grade column (Cytiva).
3. Superdex 200 gel filtration column (1.5 cm diameter × 20 cm height).
4. Syringe filter unit, 0.22 μm pore size.
5. Water-saturated butanol.
6. Model 491 Prep Cell (Bio-Rad) or equivalent, with the required equipment such as power supply, peristaltic pump, chart recorder, UV detector, and fraction collector.
7. Centrifugal concentrator, 10K MWCO.

2.3 Thermal Stability Assay of Nucleosomes in the Presence of RNAs

1. SYPRO® Orange Protein Gel Stain (Sigma-Aldrich).
2. Optical 96-well Reaction Plate or equivalent.
3. Optical Adhesive Film or equivalent.
4. Real-time PCR system.

2.4 General Equipment

1. Refrigerated microcentrifuges that can accommodate 1.5 mL tubes at speeds up to 16,000 rcf.
2. Magnetic stirrer.
3. Precision scale.
4. HPLC.
5. Gel imager that detects nucleic acids stained with ethidium bromide.
6. Microspectrophotometer.

2.5 Buffers

1. Guanidine denaturing buffer: 20 mM Tris–HCl, pH 7.5, 7 M guanidine–HCl, 20 mM 2-mercaptoethanol.
2. 2 M Refolding buffer (2MRB): 10 mM Tris–HCl (pH 7.5), 1 mM EDTA (pH 8.0), 2 mM 2-mercaptoethanol.
3. Urea-denaturing PAGE running buffer (1× TBE buffer): 90 mM Tris, 90 mM boric acid, 2 mM EDTA.
4. Reconstitution buffer-high (RB-high): 10 mM Tris–HCl (pH 7.5), 2 M KCl, 1 mM EDTA (pH 8.0), 1 mM dithiothreitol.

5. Reconstitution buffer-low (RB-low): 10 mM Tris–HCl (pH 7.5), 0.25 M KCl, 1 mM EDTA (pH 8.0), 1 mM dithiothreitol.
6. Nucleosome elution buffer: 20 mM Tris–HCl (pH 7.5), 1 mM dithiothreitol.
7. 0.2× TBE buffer: 18 mM Tris, 18 mM boric acid, 0.4 mM EDTA.

3 Methods

3.1 In Vitro RNA Transcription

1. Mix 5 μL of 1 μg/μL linearized template DNA with 50 μL of RiboMAX™ Express T7 2× Buffer and 35 μL of nuclease-free water in a 1.5 mL low protein binding tube.
2. Add 10 μL of T7 Express Enzyme mix, containing T7 RNA polymerase.
3. Incubate for 30 min at 37 °C, and then add 5 μL of DNaseI (1 U/μL) supplied with the kit to cleave the template DNA.
4. Add 5 μL of DNaseI (1 U/μL) included in the kit to cleave the template DNA.
5. Incubate for 15 min at 37 °C.
6. Add 200 μL of Milli-Q water to the RNA sample to prevent RNA loss during phenol-chloroform extraction.
7. Add an equal volume (300 μL) of phenol:chloroform (5:1) to the sample. Vortex thoroughly.
8. Centrifuge in a pre-chilled microcentrifuge at 16,000 rcf for 10 min at 4 °C.
9. Collect the upper aqueous phase containing the RNA (*see* **Note 1**) and transfer it to a new tube.
10. Repeat the previous step once to increase the purity of the RNA. Add 300 μL of chloroform:isoamyl alcohol (5:1) to the collected aqueous layer sample and vortex thoroughly.
11. Centrifuge in a pre-chilled microcentrifuge at 16,000 rcf for 10 min at 4 °C.
12. Collect the upper aqueous phase.
13. Add 30 μL of 3 M sodium acetate and 300 μL of 100% 2-propanol. Vortex thoroughly.
14. Carefully pipette the supernatant away from the white pellet containing RNA.
15. Wash the pellet with 100 μL of 70% ethanol and centrifuge at 16,000 rcf for 10 min at 4 °C.
16. Carefully pipette the supernatant away from the pellet.

17. Air dry the RNA pellet by opening the tube lid for 10 min to remove the remaining ethanol (*see* **Note 2**).
18. Resuspend the pellet well in 50 μL of Milli-Q water, using a pipette.
19. Purify the transcript using a MicroSpin G-25 column to remove any ribonucleotides that were not incorporated within the transcript or short DNAs fragmented by DNase I. Follow the manufacturer's instructions for use.
20. Transfer 50 μL of the purified RNA solution into a Mini Dialysis tube with a small disk of dialysis membrane incorporated within the cap.
21. Dialyze the sample in the dialysis tube in 400 mL of sterilized and pre-chilled Milli-Q water at 4 °C for at least 4 h at 4 °C with a magnetic stirrer.
22. Transfer the dialyzed sample to a 1.5 mL low adsorption tube.
23. Measure the RNA sample at OD_{260} and determine its concentration using a microspectrophotometer with the parameter 1 OD = 40 ng/μL.
24. To confirm that the transcripts are not degraded, fractionate the RNA samples by 6% urea-denaturing PAGE, prepared using the SEQUAGEL UreaGel system, in 1× TBE running buffer. For RNA sample preparation, mix 2 μL of 100 ng/μL transcribed RNA with 8 μL of Hi-Di formamide, then heat at 95 °C for 10 min followed by rapid cooling. Apply 5 μL of the heated sample to the gel. Use an appropriate single-stranded RNA ladder to estimate the length of the obtained RNA transcript.
25. After the electrophoresis, stain the gel with EtBr and then take a picture of the gel with an imager. If the target product appears as a single band, then the RNA sample has not been degraded (Fig. 1).
26. Store the purified sample at −30 °C.

3.2 Nucleosome Preparation for Thermal Stability Assay

The preparation method for the reconstituted nucleosomes used in the experiments cited in this paper is described in the following section.

3.2.1 Assembly of Histone H2A-H2B Dimer and H3-H4 Tetramer

1. First, measure the weight of each lyophilized core histone sample. Equimolar amounts of histones are combined; in this case, 1 mg of H2A with 1 mg of H2B, and 1.4 mg of H3 with 1 mg of H4 were mixed together for the H2A-H2B dimer and H3-H4 tetramer reconstitutions, respectively (*see* **Note 3**).

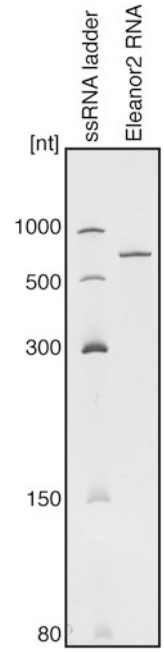


Fig. 1 Purification of *Eleanor2* RNA. Purified *Eleanor2* RNA was analyzed by 6% 1× TBE urea-denaturing PAGE with ethidium bromide staining

2. Add 2 mL and 2.4 mL of guanidine denaturing buffer to the H2A-H2B and H3-H4 mixtures, respectively, to obtain final concentrations of 1 mg/mL.
3. Allow the histones to dissolve sufficiently, by agitating the mixtures for 1.5 h at 4 °C.
4. Insert the samples into 6–8 kDa dialysis membrane. Dialyze for at least 4 h at 4 °C against 500 mL of pre-chilled 2 M refolding buffer (2MRB) with stirring, for the histones to refold while the denaturing agent is removed.
5. Dialyze against three changes of 500 mL of 2MRB for at least 4 h each at 4 °C, to eliminate all traces of guanidine. Precipitated protein should be removed by centrifugation.

3.2.2 Purification of H2A-H2B Dimer and H3-H4 Tetramer by Size Exclusion Chromatography

1. Collect and filter the samples with a 0.22 μM pore size filter. Afterwards, load the samples on a Superdex 200 gel filtration column (1.5 cm diameter × 20 cm height) or a HiLoad 16/600 Superdex 200 prep grade column connected to an HPLC or an equivalent instrument, previously equilibrated with 0.2 μm-filtered 2MRB.
2. Elute the samples using the 2MRB refolding buffer.
3. To evaluate the stoichiometry of the histones, analyze the eluted fractions by 16% SDS-PAGE. Collect and pool the fractions that contain equal molar amounts of the two histone combinations.

Table 1
Extinction coefficients for canonical human histones

Histone	Experimental value (M^{-1} cm^{-1})	Calculated value (M^{-1} cm^{-1})
H2A	4215	4470
H2B	3101	7450
H3.1	1782	4470
H4	4030	5960

4. Concentrate the fractions to about 100 μL by ultrafiltration at 7500 rcf, using a centrifugal concentrator (10 kDa pore size) and a pre-chilled centrifuge.
5. Measure the absorbances of the H2A-H2B dimer and the H3-H4 tetramer at OD_{280} using a microspectrophotometer. Use 2MRB as a reference.
6. Determine the molar concentration of the histone complexes using the absorbance value of OD_{280} and the extinction coefficient of each histone (Table 1).
7. Flash freeze the histone complexes in liquid nitrogen and store them at −80 °C.

3.2.3 Nucleosome Reconstitution from Purified Histone Complexes

To reconstitute nucleosomes in vitro, we used a palindromic 146 base-pair α-satellite DNA fragment [25].

Reconstitution of the Nucleosome

The salt dialysis method is the traditional method employed for nucleosome reconstitution. Here, we report the process of nucleosome assembly using H2A-H2B dimers and H3-H4 tetramer (*see* **Note 4**).

Small-Scale Reconstitution

1. Determine the appropriate molar ratios of H2A-H2B dimer and H3-H4 tetramer to DNA to be used in the small-scale reconstitution. This step is very important because it will help to optimize the yields of correctly reconstituted nucleosomes in large-scale reconstitutions. In this experiment, the reconstitution is performed with the refolded H2A-H2B dimer and H3-H4 tetramer, so the three components need to be titrated. Calculate the amounts of histone complexes to mix with approximately 50 μg of DNA, with trial ranges of 3.2 to 4 (e.g., 3.4, 3.6, 3.8), and 3 to 3.5 (e.g., 3, 3.2, 3.4) for H2A-H2B dimer and H3-H4 tetramer, respectively (*see* **Note 5**). Adjust the salt concentration to 2 M KCl using 4 M KCl and the final concentration of DNA to 0.65 mg/mL with water.

2. Mix all substances together, with the histone complexes added last.
3. Transfer the sample to a dialysis tube and dialyze the sample against 400 mL of RB-high at 4 °C with stirring.
4. Calibrate the peristaltic pump to a flow rate of 0.8 mL/min and start the reconstitution process by continuously replacing the dialysis solution with RB-low buffer containing 250 mM KCl (1.6 L), which will lead to a decrease in the KCl concentration, while stirring.
5. Dialyze the sample against 400 mL of RB-low for at least 4 h, once the gradient has finished. Store the samples at 4 °C.
6. Heat the sample at 55 °C for 2 h to stabilize the nucleosome positioning.
7. Evaluate the correct assembly of the nucleosome by 6% 0.2× TBE native-PAGE (*see* **Note 6**).

Large-Scale Reconstitution

1. Calculate the amounts of DNA and histone complexes according to the optimal ratio deduced from the small-scale reconstitution.
2. Combine water, 4 M KCl, DNA, H2A-H2B dimers and H3-H4 tetramers together, under the same conditions as in the small-scale reconstitution.
3. Dialyze the sample against RB-high, and assemble the nucleosome using the salt dialysis method, described above in Subheading "Small-Scale Reconstitution", **steps 3–5**, followed by the incubation at 55 °C.

Purification of Nucleosomes

Reconstituted nucleosomes are purified by preparative native-PAGE using a Bio-Rad Prep Cell apparatus, according to the following procedure. The sample mixture is separated based on the migration speed through the gel, allowing the recovery of reconstituted nucleosomes while excluding the excess free histones and DNA.

1. Prepare a 6.5 cm high 6% polyacrylamide gel with 0.2× TBE, using a Model 491 Prep Cell apparatus according to the manufacturer's instructions.
2. Once the gel is poured, overlay the top of the gel with water-saturated butanol as soon as possible (*see* **Note 7**).
3. Prepare 20 mM Tris–HCl (pH 7.5) buffer containing 1 mM DTT for sample elution, and 0.2× TBE buffer for running buffer, and store at 4 °C.
4. Once the gel has polymerized, rinse the gel surface with ultrapure water.
5. Prepare the Prep Cell system according to the manufacturer's instructions. Set up a dialysis membrane as a filter.

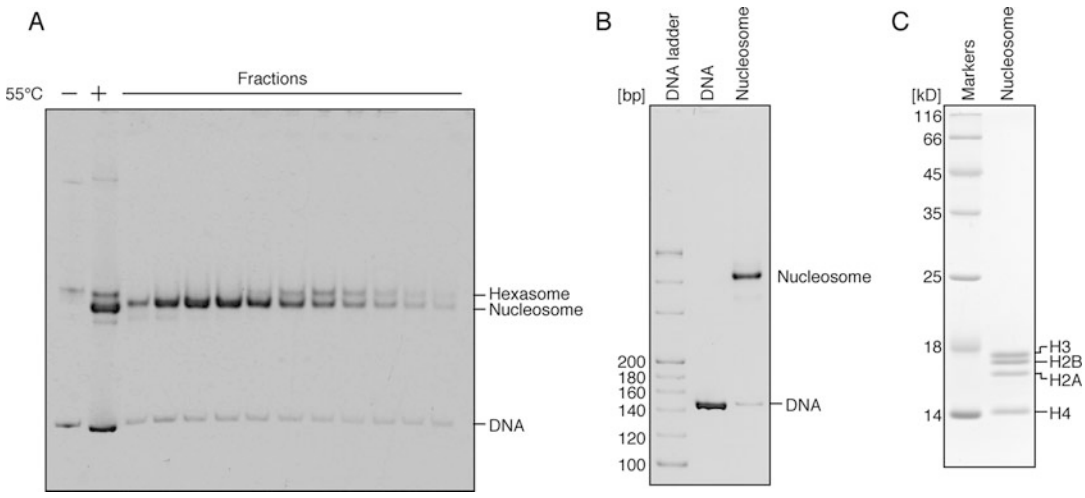


Fig. 2 Reconstitution of the human canonical nucleosome. (**a**) Bands corresponding to the reconstituted nucleosomes that were incubated (+) or not (−) at 55 °C are shown on the left side of the gel. The reconstituted nucleosomes were purified by a Prep Cell apparatus (6.5 cm height, 6% acrylamide, 0.2× TBE). The eluted nucleosome peak fractions were analyzed by 0.2× TBE nondenaturing 6% PAGE with ethidium bromide staining and are represented on the right side of the gel. Bands corresponding to hexasome, nucleosome, and naked DNA are indicated. (**b**) The obtained nucleosome was analyzed by 0.2× TBE nondenaturing 6% PAGE with ethidium bromide staining. (**c**) The histone contents of the purified nucleosomes were analyzed by 18% SDS-PAGE with Coomassie Brilliant Blue (CBB) staining

6. Pre-run the gel column 1 h at 10 W, at a flow rate of 1 mL/min.
7. Add 30% sucrose to the sample mixture to a 5% final concentration. Save 10 μL for gel analysis.
8. Load the nucleosome sample on the gel column and run at 10 W for approximately 2 h, at a flow rate 1.5 mL/min. Collect 1.5 mL/fraction.
9. Check the eluted fractions by measuring the absorbance at 260 nm.
10. Analyze the eluted fractions on a 6% 0.2× TBE native gel (Fig. 2a).
11. Collect and pool the fractions that contain the reconstituted nucleosomes and concentrate with a centrifugal concentrator (30 kDa pore size).
12. Measure the absorbance value at OD_{260}, since the nucleosome concentration depends on the DNA concentration in the nucleosome.
13. Evaluate the quality of the final product after concentration on a 6% 0.2× TBE native gel and by 18% SDS-PAGE, as shown in Fig. 2b, c.

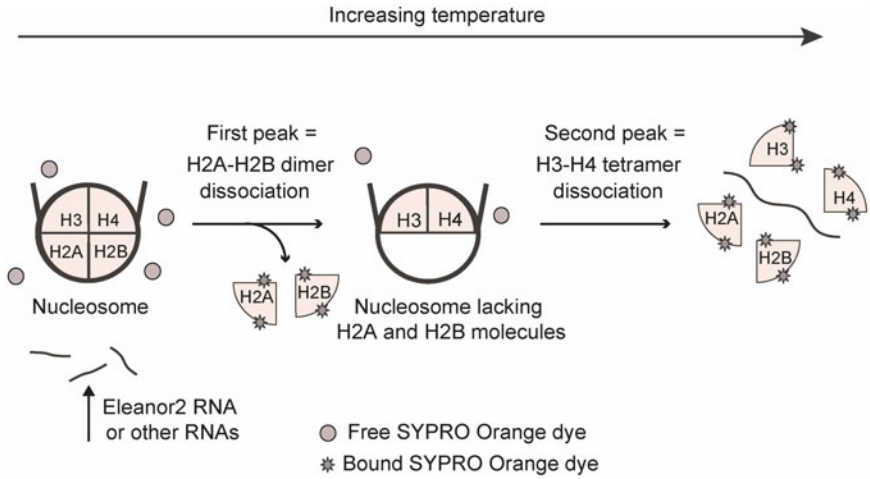


Fig. 3 Schematic representation of nucleosome disruption during the thermal stability assay using SYPRO Orange, a fluorescent dye. When SYPRO Orange binds to the hydrophobic residues of the denatured protein, it emits fluorescence that is monitored as the histones dissociate from the nucleosomes or the nucleosomes lacking H2A and H2B molecules

3.3 Thermal Stability Assay of the Nucleosome in the Presence of Eleanor2 RNA

The thermal stability assay allows the protein unfolding process to be monitored by increasing the temperature in the presence of SYPRO Orange, a fluorescent dye that preferentially binds to unfolded proteins [26] (Fig. 3). The thermal stability assay that we have developed to evaluate the impact of the Eleanor2 RNA on the physical properties of the nucleosomes is described in the following section.

1. Mix 9 μL of 2.78 μM nucleosome, 7 μL of 3.57 μM Eleanor2 RNA (this can be substituted with the ncRNA of the user's interest purified in step (1), and 2 μL of 1 M NaCl in a 1.5 mL low-absorption tube. As a background for quantitative analysis, mix 9 μL of nucleosome elution buffer, 7 μL of MilliQ-water, and 2 μL of 1 M NaCl (*see* **Note 5**). Prepare a control sample containing MilliQ-water instead of Eleanor2 RNA.
2. Prepare a final solution of 50× SYPRO Orange solution by diluting 1 μL of SYPRO Orange (initial concentration 5000×) in 99 μL of nucleosome elution buffer (*see* **Notes 8** and **9**).
3. Add 2 μL of freshly diluted SYPRO Orange to 18 μL of the nucleosome sample and mix well.
4. Load 19 μL of the sample solution into a 96-well plate and seal tightly with an adhesive film.
5. Briefly centrifuge the 96-well plate containing the sample solution.
6. Acquire the fluorescence signals with a StepOnePlus™ Real-Time PCR system (Applied Biosystems) with continuous fluorescent measurement, starting at 26 °C and ending at 95 °C (ramping rate of 1 °C/min).

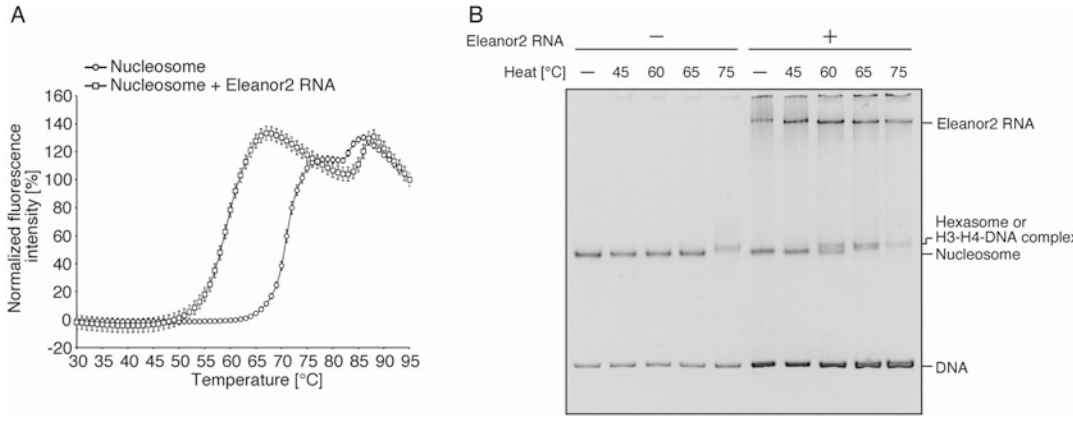


Fig. 4 Thermal stability assay of nucleosomes in the presence of *Eleanor2* RNA. (**a**) Normalized fluorescence intensity curves of the thermal dissociation of canonical human nucleosomes in the presence (□) or absence (□) of *Eleanor2* RNA. The first and second peaks correspond to the dissociations of the H2A-H2B dimers and the H3-H4 tetramer from the nucleosome, respectively. (**b**) The samples, including non-heated and heated up to 45 °C, 60 °C, 65 °C, and 75 °C, and in the presence (+) or absence (−) of *Eleanor2* RNA, were analyzed by 0.2× TBE nondenaturing 6% PAGE with ethidium bromide staining. Bands corresponding to *Eleanor2* RNA, hexasome or H3-H4-DNA complexes, nucleosome, and naked DNA are indicated

7. Once the measurement has ended, convert the raw fluorescence data to normalized values as: $(F(T) - F_{26°C})/(F_{95°C} - F_{26°C})$, where $F(T)$, $F_{26°C}$, and $F_{95°C}$ indicate the fluorescence at a particular temperature, the fluorescence at 26 °C, and the fluorescence at 95 °C, respectively (Fig. 4a).

3.4 Analysis of the Disrupted Nucleosomes by Native-PAGE

The gradual steps of nucleosome disruption can be visualized on a gel, as described in the next section.

1. Prepare the reaction sample in excess (7.5 μL of sample x number of desired steps). Apply 6 μL into a 96-well plate and proceed to **step 6**.
2. Abort the analysis when the desired temperature is reached.
3. Remove the 96-well plate from the machine and recover the sample in a 1.5 mL tube. Keep at 4 °C.
4. Apply 6 μL of the reaction sample to a new 96-well plate and start the incubation once again until it reaches the next desired temperature.
5. Repeat **steps 1** and **2** for all desired temperatures.
6. For a control, keep a sample at 4 °C.
7. Add 1 μL of 30% sucrose to 3 μL of sample, and analyze the different samples representing the gradual steps of the nucleosome disruption on a 6% 0.2× TBE native gel (Fig. 4b).

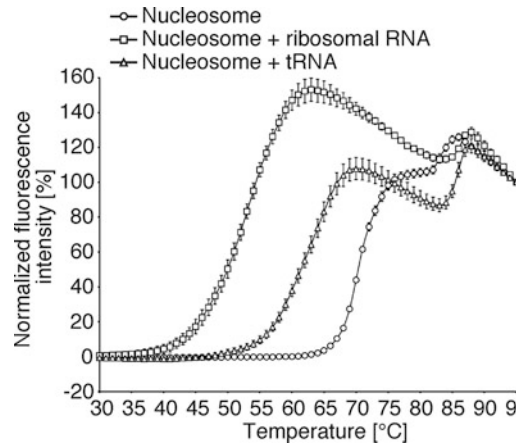


Fig. 5 Thermal stability assay of nucleosomes in the presence of other RNAs. Normalized fluorescence intensity curves of the thermal dissociation of canonical human nucleosomes in the presence of ribosomal RNA (□) or tRNA (△), or absence (▯) of RNAs

3.5 Thermal Stability Assay of the Nucleosome in the Presence of Other RNAs

The thermal stability of nucleosomes in the presence of various RNAs other than Eleanor2 was assessed. For this purpose, we used ribosomal RNA and tRNA from commercial sources.

1. Assemble the sample reaction by mixing X nucleosome and X RNA, NaCl, and SYPRO Orange to a final buffer concentration of 100 mM NaCl containing 5× SYPRO Orange.
2. After loading the sample mixture into a 96-well plate, set up the Real-Time PCR equipment and analyze the dissociation of the histones and DNA, using the same method as described in Subheading 3.3, **steps 5**–7 (Fig. 5).

4 Notes

1. Be careful to avoid collecting any of the small white precipitate of denatured protein debris formed between the aqueous and phenolic layers.
2. The tubes can be dried in a vacuum centrifuge, but should not be overdried since the RNA will become insoluble. We advise covering the tube with plastic wrap to prevent RNase contamination, since the tube is dried with the lid open.
3. Lyophilized recombinant histones expressed in *E. coli* can be obtained according to the chapter by Kujirai et al. 2018 [27], describing histone purification methods in detail.
4. The histone octamer can be used for nucleosome reconstitution, but we prefer to use the H2A-H2B dimers and H3-H4 tetramer, since this combination helps to reduce the formation

of hexasomes. For nucleosome reconstitution using the histone octamer, see the article by Kujirai et al. 2018 [27].

5. Adding more H2A-H2B dimers than H3-H4 tetramers prevents the formation of hexasomes.
6. If the titration ratios cited in **step 1** of Subheading "Reconstitution of the Nucleosome" do not work, extend the titration range since histone complexes vary from lot to lot. If no nucleosomes are formed (i.e., the naked DNA remains unincorporated with histones), increase the amount of histones relative to DNA. On the other hand, if aggregates appear and no bands are detected, decrease the amount of histones relative to DNA.
7. In order to avoid a temporary decrease of the nucleosome concentration, the nucleosomes should be added first. In parallel, NaCl should be added last, as otherwise the high salt concentration may destroy the nucleosomes.
8. Using a SYPRO Orange solution that contains precipitates will not work. Try to dissolve the precipitate by vortexing before use or prepare a new lot.
9. As SYPRO Orange tends to bind to the wall of the tube, resuspend the solution three or four times by pipetting before removing the working volume.

Acknowledgements

We are grateful to Ms. Yukari Iikura (The University of Tokyo) for her assistance. This work was supported in part by JSPS KAKENHI grants [JP18H05534 to H.K., JP19K23714 to R.F., JP20K15711 to T.K., JP20H00449 to H.K.] and by grants from the Japan Science and Technology Agency (JST) Exploratory Research for Advanced Technology (ERATO) [JPMJER1901 to H.K.] and the Platform Project for Supporting Drug Discovery and Life Science Research (BINDS) from the Japan Agency for Medical Research and Development (AMED) [JP20am0101076 to H.K.]. This work was partly supported by JST CREST grant number JPMJCR16G1 [to H.K. and T.K.].

References

1. Wolffe A (1998) Chromatin: structure and function, 3rd edn. Academic Press, San Diego
2. Luger K, Mäder AW, Richmond RK et al (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature 389: 251–260
3. Bondarenko VA, Steele LM, Ujvàri A et al (2006) Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. Mol Cell 24:469–479

4. Teves SS, Weber CM, Henikoff S (2014) Transcribing through the nucleosome. Trends Biochem Sci 39:577–586
5. Kujirai T, Kurumizaka H (2020) Transcription through the nucleosome. Curr Opin Struct Biol 61:42–49
6. Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23: 1494–1504
7. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81:145–166
8. Ding DQ et al (2012) Meiosis-specific noncoding RNA mediates robust pairing of homologous chromosomes in meiosis. Science 11:732–736
9. Batista PJ, Chang HY (2013) Long noncoding RNAs: cellular address codes in development and disease. Cell 152:1298–1307
10. Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. Cell 154: 26–46
11. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157:77–94
12. Krawczyk M, Emerson BM (2014) p50-associated COX-2 extragenic RNA (PACER) activates COX-2 gene expression by occluding repressive NF-kappaB complexes. eLife 3:e01776
13. Kopp F, Mendell JT (2018) Functional classification and experimental dissection of long noncoding RNAs. Cell 172:393–407
14. Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 489:101–108
15. Gil N, Ulitsky I (2020) Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genet 21:102–117
16. Statello L, Guo CJ, Chen LL et al (2021) Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 22: 96–118
17. Tsai MC, Manor O, Wan Y et al (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329:689–693
18. Han P, Li W, Lin CH et al (2014) A long noncoding RNA protects the heart from pathological hypertrophy. Nature 514:102–106
19. Han P, Chang CP (2015) Long non-coding RNA and chromatin remodelling. RNA Biol 12:1094–1098
20. Nozawa R, Yamamoto T, Takahashi M et al (2020) Nuclear microenvironment in cancer: control through liquid-liquid phase separation. Cancer Sci 111:3155–3163
21. Redon S, Reichenbach P, Lingner J (2010) The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. Nucleic Acids Res 38:5797–5806
22. Rossi M, Gorospe M (2020) Noncoding RNAs controlling telomere homeostasis in senescence and aging. Trends Mol Med 26:422–433
23. Tomita S, Abdalla MOA, Fujiwara S et al (2015) A cluster of noncoding RNAs activates the ESR1 locus during breast cancer adaptation. Nat Commun 29:6966–7015
24. Jeng MH, Shupnik MA, Bender TP et al (1998) Estrogen receptor expression and function in long-term estrogen-deprived human breast cancer cells. Endocrinology 139: 4164–4174
25. Dyer PN, Edayathumangalam RS, White CL et al (2004) Reconstitution of nucleosome core particles from recombinant histones and DNA. Methods Enzymol 375:23–44
26. Taguchi H, Horikoshi N, Arimura Y et al (2014) A method for evaluating nucleosome stability with a protein-binding fluorescent dye. Methods 70:119–126
27. Kujirai T, Arimura Y, Fujita R et al (2018) Methods for preparing nucleosomes containing histone variants. Methods Mol Biol 1832: 3–20

Chapter 13

Revisiting the Glass Treatment for Single-Molecule Analysis of ncRNA Function

Shuting Shen, Masahiro Naganuma, Yukihide Tomari, and Hisashi Tadakuma

Abstract

Single-molecule imaging is a powerful method for unveiling precise molecular mechanisms. Particularly, single-molecule analysis with total internal reflection fluorescence (TIRF) microscopy has been successfully applied to the characterization of molecular mechanisms in ncRNA studies. Tracing interactions at the single-molecule level have elucidated the intermediate states of the reaction, which are hidden by ensemble averaging in combinational biochemical approaches, and clarified the key steps of the interaction. However, applying a single-molecule technique to ncRNA analysis still remains a challenge, requiring laborious trial and error to identify a suitable glass surface passivation method. In this chapter, we revisit the major glass surface passivation methods using polyethylene glycol (PEG) treatment and summarize a detailed protocol for single-molecule analysis of the dicing process of Dcr-2, which may apply piRNA studies in the future.

Key words Single-molecule imaging, Non-coding RNA (ncRNA), Dicer 2, Polyethylene glycol (PEG), Surface passivation, Amino-silanization, Halo ligand, SNAP ligand, Total internal reflection fluorescence (TIRF) microscopy

1 Introduction

The interaction of the biomaterials (RNAs, proteins, etc.) is the basis of biochemical reactions. Historically, biochemical interactions have been measured by bulk experiments. Although sophisticated methods (e.g., stopped-flow) allow precise measurements with high temporal resolution, the detailed processes are hidden by averaging. This is partially because biochemical reactions are harnessed by the heterogeneous states of the molecules even using highly purified biomaterials (e.g., proteins, nucleic acids). In contrast, single-molecule imaging allows the researcher to trace the whole process. Thereby, single-molecule imaging provides

Shuting Shen and Masahiro Naganuma contributed equally with all other contributors.

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_13,

a means by which the reaction can be synchronized post-experimentally, eliminating the hindering effects of ensemble averaging. Therefore, single-molecule imaging has achieved great success in the field of non-coding RNA (ncRNA) [1–5].

In many experiments of single-molecule interaction analysis, one of the molecules (e.g., RNA) is fixed (anchored) on the glass surface. The association and dissociation of the partner molecules (e.g., protein) are then observed. The key issue with such experiments is the glass surface passivation, which improves the functionality of the fixed molecules and also prevents non-specific binding of the partner molecule to the glass surface. The role of the former is to prevent the adsorption of the fixed molecule's binding site to the glass surface (e.g., if the binding site faces to and adsorbs to the glass surface, the partner molecule will associate less frequently). The role of the latter is to distinguish the real interaction between the fixed- and partner molecules from non-specific interactions between glass surface and partner molecule.

Thus, from the beginning of single-molecule imaging, the passivation of glass is key to observing the activities of biomolecules. During the very early period of single-molecule imaging, researchers tried observing the activities of motor proteins, such as myosin, using a non-treated glass surface, where most of the proteins adhered to the glass surface and lost function. To overcome this problem, researchers used filaments of proteins (e.g., myosin filaments) to allow some of them to stay away from the glass surface, thus maintaining their activities [6–8]. Shortly after this, the glass passivation method was introduced. Researchers used silicone (e.g., Sigmacote) or high molecular weight polymers like collodion (nitrocellulose film) [9, 10]. These methods make a thin film on the glass, preventing surface adsorption of many proteins. However, the quality and condition of the thin film are highly dependent on the experimenter and/or the condition of the day (e.g., the difference of temperature, humidity). Therefore, experience is needed for reproducible and reliable results, especially when reducing the concentration of the fixed molecule.

To overcome such drawbacks, two passivation methods have been widely used: bovine serum albumin (BSA)-biotin [11, 12] and polyethylene glycol (PEG) treatment [13–16]. BSA-biotin is a convenient method in which infusion of BSA-biotin into the observation chamber is followed by fixation of biotinylated molecules through streptavidin (or Neutravidin) protein. Although convenient, non-covalently bonded BSA-biotin is sometimes replaced by the sticky proteins in the reaction solution. Furthermore, the physical size of BSA (MW 66 kDa) makes a gap on passivation thin film, causing the non-specific binding of the observation object. Therefore, most studies have used PEG treatment, where the glass surface is first aminosilanized and then reacted with N-hydroxysuccinimide (NHS) ester-modified PEG, which also

includes a small fraction of biotin-PEG-NHS ester for target molecule anchoring.

The PEG treatment is composed of three main steps. In the first step, the glass surface is cleaned. This step is important to improve the surface condition for homogeneous and high quality surface passivation, especially in the reuse process of quartz glass. The second amino-silanization step also affects the quality. In some cases, optimization of the amino-silanization reagent and solvating media are required. The third PEG treatment step is highly dependent on the observation target. Usually, linear PEG has been used. For some sticky observation targets, branched PEG (e.g., 4-arm PEG) improves the condition of the surface [16, 17]. Also, several supplements have been used to fill the gap existing in the PEG layer: a second round of PEG treatment with short PEG [15], lipid-mimic polymer [18, 19], tRNA [2], and/or detergent [20]. The diversity of methods used suggests that there is unlikely a current single best method that remains sufficiently versatile. Other approaches including lipid bilayers [21], 2D crystal of streptavidin [22, 23], and DNA nanostructure [24, 25] can provide denser passivation, but, currently, the surface area of that is uniformly passivated under these conditions is limited. Therefore, these approaches are currently limited to the imaging technologies such as atomic force microscopy (AFM) and Cryo-EM, where homogeneity only in a narrow area (approx. a few micrometers) is required, and the surface conditions can be directly evaluated. Taken together, PEG treatment is still the most reliable method for single-molecule imaging using fluorescent dye, where homogeneity of a wide area (~0.1 mm) is required.

In the ncRNA field, RNAs and proteins orchestrate sophisticated biochemical reactions often resulting in silencing of target genes [26]. In the first step of RNA silencing such as small interfering RNAs (siRNAs) or microRNAs (miRNAs), Dicer, an RNase III enzyme, cleaves long double-stranded RNAs (dsRNAs) or precursor miRNAs (pre-miRNAs). However, the details of the mode of this cleavage remain enigmatic, especially whether the Dicer cleaves the substrate dsRNA by processive- or by distributive mode. In the processive mode, the Dicer successively cleaves the dsRNA without dissociation, whereas in the distributive mode, the Dicer dissociates from the dsRNA after the first cleavage. In the biochemical assay, in theory, these two cleavage modes can be distinguished by allowing Dicer to cleave radiolabeled dsRNAs for a short time and then challenged by a vast excess of cold dsRNAs. Practically, however, this is difficult, especially if the reaction is the mixed form of these two cleavage modes. In contrast, single-molecule imaging can clearly distinguish these two modes, as it is possible to trace the whole cleavage process (i.e., association, cleavage, and dissociation). Recently, we established real-time monitoring of the fly Dicer-2 cleavage reaction. In our system, long

dsRNA (220-nt) with multiple fluorescent dye labels was used as substrate such that the fluorescent intensity of dsRNA decreased with the cleavage reaction by Dicer-2, and thus, can clearly trace the whole cleavage process at single-molecule level. Our results clearly showed that Dicer-2 cleaved the long dsRNA substrate by processive mode in majority, refining the current model of its action [27]. This single-molecule approach is a versatile method and may be applicable to the piRNA research in the future.

In this chapter, we have summarized a detailed protocol for single-molecule analysis, especially the glass passivation method as well as small tips regarding the anti-photobleaching (trolox-quinone, TQ) and protein-labeling reagent (Halo/SNAP-Cy5-biotin ligand) for the observation of dsRNA cleavage reaction by Dicer-2.

2 Materials

2.1 General Buffers, Reagents, and Cells

1. 5× lysis buffer for preparing 1× lysis buffer: 150 mM HEPES-KOH (pH 7.5), 500 mM KOAc, 10 mM $Mg(OAc)_2$.
2. 1× lysis buffer: 30 mM HEPES-KOH (pH 7.5), 100 mM KOAc, 2 mM $Mg(OAc)_2$.
3. 1 M dithiothreitol (DTT).

2.2 TIRF Microscope Setup

We constructed a prism-type custom TIRF microscope using an inverted microscope, lasers, a fluorescence image splitting system, and a back-illuminated electron-multiplying charge-coupled device camera (EMCCD) [28]. Details of the assembly are described in the literature [19]. The components of our custom microscope are listed below.

1. Inverted type microscope IX71 (Olympus).
2. Oil immersion objective lens, UAPON 150× OTIRFM, NA 1.45 (Olympus). (Other objectives such as 60× (UPLAPO60XOHR) or 100× (UPLAPO100XOHR) are also usable. The choice of objective magnification depends on the camera sensor size, pixel size, and the density of the signal and is important for precise 2D or 3D trajectory analysis [29, 30].
3. Anti-vibration table, 1200× 900 mm.
4. 532 nm optically pumped semiconductor laser (OPSL, Coherent) and 637 nm laser diode (LD, Coherent). The beam quality of gas lasers (e.g., Ar/He-Ne laser) are high, but the digital modulation capability of the OPSL/LD laser is the attractive point. The pump laser of OPSL/LD should be cleaned up by filter [FF01-532/18-25 (for 532 nm) and LD01-640/8-12.5 (for 637 nm)].

5. Back-illuminated electron-multiplying charge-coupled device (EMCCD) camera, iXon3 DU-897E-CSO-#BV, 512 × 512 pixels, (Andor Technology). With enough photons, scientific complementary metal-oxide-semiconductor (sCMOS) cameras (Andor, Hamamatsu, Teledyne Photometrics, etc.) could be used instead of EMCCD cameras.
6. Fluorescence image splitter, DualView2, (Teledyne Photometrics) with a 635 nm filter cube (Dichroic mirror, T635LPXR, (Chroma)): In the detection pathway, the DualView2 module separates spatially identical but spectrally distinct emission lights and projects onto the camera chip side by side.
7. Emission filters (Semrock), BLP01-532R-25/BSP01-633R-25 (for Cy3) and BLP01-633R-25/FF01-758/SP-25 (for Cy5) for DualView2.
8. Quarter waveplates, 532 nm (WPQ-5320-04M, OptoSigma), 633 nm (WPQ-6328-04M, OptoSigma).
9. Synthetic quartz prism, 20 × 20 × 6 mm, OPSQ-20S06-4P-3 (OptoSigma).
10. Lens, DLB-15-100PM (OptoSigma): to adjust the illumination area of TIRF.
11. Single-board microcontroller (e.g., Arduino Uno Rev3): To on/off control the laser. Shield (printed circuit expansion boards) is described in literature [31].

2.3 HaloTag and SNAP-Tag Ligand Having Both Cy5 and Biotin

Commercially available ligands have only fluorescent dye or biotin moiety for anchoring. To simplify the fluorescent labeling and glass anchoring process, we constructed dual ligands having both Cy5 and biotin [32].

1. HaloTag Amine (O4) Ligand (Promega). Dissolve the whole bottle contents (5 mg) in 1 ml DMSO to make 5 mg/ml (14.4 mM) of the stock solution and stored at −80 °C. Use fresh DMSO, or DMSO treated with molecular sieves (desiccants).
2. BG-PEG-NH2 Ligand (for SNAP-Tag protein, NEB). Dissolve the whole bottle contents (2 mg) in 570 μl DMSO to make 3.5 mg/ml (7.2 mM) of the stock solution and stored at −80 °C. Higher concentration is difficult to solve, so we choose this concentration.
3. Cy5 Bis NHS ester (Cytiva/GE). Dissolve in DMSO to make 5 mg/ml (5.1 mM) of the stock solution and stored at −80 °C. Confirm the concentration by spectrometer (e.g., Nanodrop) using the absorbance of Cy5. If measured concentration (= C_M) is lower than expected (5.1 mM), use the measured concentration (C_M).

4. Biotin-AC5-hydrazide (MedChemExpress). Dissolve the whole bottle contents (10 mg) in 2 ml DMSO to make 5 mg/ml (10.3 mM) of the stock solution and stored at −80 °C.
5. Triethylamine.
6. Dichloromethane.
7. Methanol.
8. TLC plate.
9. Thermal cycler or constant-temperature incubator.

2.4 Dicer-2 Protein Labeling with HaloTag Ligands

1. pAHisHaloW-Dcr-2 plasmid.
2. X-treamGENE HP DNA transfection reagent (Roche).
3. *Drosophila* S2 cells.
4. *Drosophila* Schneider's medium (Thermo Fisher), in some case antibiotics are omitted to make antibiotics-free medium.
5. Fetal bovine serum (FBS).
6. 1× Phosphate suffered saline.
7. Protease inhibitor cocktail (Roche).
8. Dounce homogenizer, 7 ml, "TIGHT" pestle (Wheaton).
9. HaloTag Cy5-biotin ligand.
10. Complete His-Tag Purification Resin (Sigma-Aldrich).

2.5 Preparation of Fluorescently Labeled dsRNA Target

1. Target sequence in pUC57 ([27], *see* **Note 1**).
2. Forward and reverse primer for transcription template construction, where reverse primer is modified with 2′OMe [33].
3. T7 transcription kit (CELLSCRIPT).
4. Cy3-UTP (Cytiva/GE).
5. GMP to produce 5′ monophosphorylated RNAs.
6. Water purified by a laboratory water purification system (MilliQ or equivalent).
7. 1× Lysis buffer: Dilute 5× Lysis buffer to make 1×.

2.6 Glass Cleaning and Passivation

1. Synthetic quartz glass slides, 26 × 56 × 1 mm, (Matsunami Glass).
2. Coverslips, 24 × 36 mm, (Matsunami Glass).
3. APTES ((3-Aminopropyl)triethoxysilane, Sigma).
4. 4 Arm-PEG-SC, MW 10 kDa (Laysan Bio): In the case of using linear PEG, use mPEG-SVA-5000 (Laysan Bio).
5. Biotin-PEG SVA 5K (Laysan Bio).
6. MS(PEG)4 (Thermo).

7. Sulfuric acid (H_2SO_4).
8. Hydrogen peroxide (H_2O_2).
9. Potassium hydroxide (KOH). Three concentrations of KOH solution are required: 0.1, 0.5, and 1.0 M.
10. Acetone.
11. 0.1 M Sodium carbonate/bicarbonate buffer at pH 9.5 (mix 30 ml of 0.1 M Na_2CO_3 and 70 ml of $NaHCO_3$).
12. Potassium sulfate (K_2SO_4) 0.55 M: 95.8 mg powder dissolved in 0.1 M Sodium carbonate/bicarbonate buffer (pH 9.5).
13. Ethanol.
14. Methanol.
15. Acetic acid.
16. 2 L beaker.
17. Glass staining trough, Hellendahl extended (BRAND).
18. Glass staining dish for eight slides (Matsunami Glass).
19. 96 × 160 × 73 mm deep stainless steel vat.
20. PFA Coating Tweezers 210 mm Straight (AS ONE).
21. Sonicator water bath.
22. Vacuum sealer: Compatible with both plastic bags and vacuum canisters.
23. Plasma Cleaner (Harrick).
24. Detergent (e.g., DCN90, AR BROWN).

2.7 Preparation of the Observation Chamber

1. Double-sided tape: 24 × 36 mm, window size 16 × 24 mm, thickness 25 mm (3 M).
2. Aluminum block (BIO-BIK): To make it easy to peel off the paper liner of double-sided tape (*see* below).
3. 3MM Filter paper, 10 × 100 mm (Whatman).
4. 2.5 mg/ml NeutrAvidin solution (Invitrogen): Dissolve 5 mg of NeutrAvidin in 2 ml of purified water. Make 100 μl of aliquots and flash-freeze with liquid nitrogen. Store at −30 °C.
5. Objective oil.
6. Glycerol.
7. 5% Biolipidure-203 (NOF).
8. 0.2 M TCEP: Dissolve 1 g of TCEP in 17.4 ml of purified water. Dilute further to 1 mM at use by purified water.
9. Glucose: Dissolve 4.5 g of glucose in 10 ml of purified water to make 450 mg/ml of stock solution. Make 10 μl of aliquots and flash-freeze with liquid nitrogen. Store at −80 °C.

10. Glucose–oxidase: 50,000 U (Sigma): Dissolve in purified water to make 5000 U/ml of the stock solution. Make 10 μl of aliquots and flash-freeze with liquid nitrogen. Store at −80 °C.
11. Catalase: Dissolve 100 mg (Sigma) of catalase in purified water to make 5000 U/ml of the stock solution. Make 10 μl of aliquots and flash-freeze with liquid nitrogen. Store at −80 °C.
12. 1 μM Protocatechuate 3,4-Dioxygenase (PCD, MW ~700 kDa, Sigma): Prepare and store the stocks as follows [34]:
    (a) Make stock buffer (50% glycerol, 50 mM KCl, 1 mM EDTA, 100 mM Tris-HCl pH 8).
    (b) Dissolve 8.5 mg of PCD in 12.14 ml of stock buffer (0.7 mg/ml → 1 μM).
13. 100 mM 3,4-Dihydroxybenzoic acid (protocatechuic acid, PCA, Sigma): Prepare and store the stocks as follows:
    (a) Dissolve 154 mg of PCA in 8.5 ml of purified water.
    (b) Add 1 M NaOH (−1 ml), adjust to pH 9.0 (poor solubility).
    (c) Fill up to 10 ml and make 100 μl of aliquots and flash-freeze with liquid nitrogen. Store at −30 °C.
14. (±)-6-Hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid (Trolox), (Sigma): Dissolve 10 mg of Trolox in 200 μl of ethanol to make 200 mM of the stock solution. Store at 4 °C protected from light for up to 2 weeks.
15. 1 mM TX-quinone (TQ): Prepare and store the stocks as follows (Fig. 1, [35]):
    (a) Dilute 200 mM Trolox to 1 mM Trolox using purified water.
    (b) Measure with Nanodrop or similar ($Abs_{255} = TX_0$ set to baseline).
    (c) Take 60 μl 1 mM Trolox, irradiate with UV (Y min), and measure with Nanodrop ($Abs_{255} = TX_Y$). In our hands, 30 min irradiation with "high" mode of 25 W transilluminator (UVP, 302 nm) or 8 min irradiation with xenon lamp (Asahi Spectra, MAX-350, 310 nm, at 5 mW/cm^2).
    (d) Calculate the TQ concentration by the following equation (it might be 200–250 μM at peak irradiation time):

$$TQ\ (\mu M) = (TX_Y - 0.4)/0.0112 \quad (1)$$

    (e) Add final 2 mM TX and final 50 μM TQ into the 100 μl observation buffer.

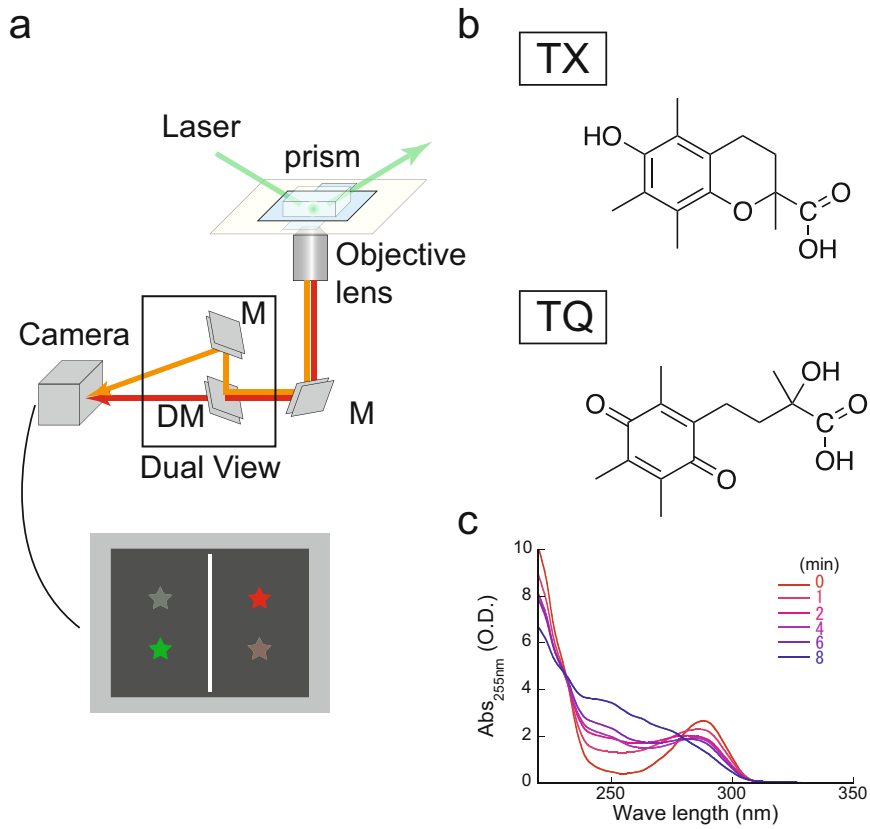


Fig. 1 Single-molecule system. (**a**) Schematic illustration of single-molecule imaging system using TIRF microscope. Evanescent field is made at the boundary of quartz glass and observation buffer using a prism. Fluorescent signal is collected by the objective lens, separated into Cy3 and Cy5 signals by a dichroic mirror (DM), and projected side by side on the camera. (**b**) Trolox (TX), a vitamin E analog, efficiently eliminates blinking related to triplet states as well as blinking occurring on longer time scales. By UV (around 300 nm) irradiation, TX can be converted to its oxidized form (TX-quinone, TQ). (**c**) Spectrum shows the TX to TQ conversion. TQ concentration can be estimated by measuring the absorption at 255 nm (A_{255}, *see* main text). The optimal ratio of TX and TQ depends on the fluorescent dye and observation solution and can be empirically determined

2.8 Software

1. Image J (https://imagej.github.io/).
2. Matlab (MathWorks).
3. vbFRET, a hidden Markov model based analysis plug-in for Matlab [36].
4. Excel (Microsoft).
5. KaleidaGraph (Synergy Software).

3 Methods

3.1 Dicer-2 Protein Labeling with HaloTag Ligands

3.1.1 HaloTag Cy5-Biotin Preparation (See **Note 2**)

1. Mix 10 μl of HaloTag Amine (O4) Ligand (14.4 mM), 90 μl of Cy5 Bis NHS ester (5.1 mM), and 0.3 μl of Triethylamine (7 M, undiluted form) in a PCR tube (molar ratio is 1:3.2:14.9. If the concentration of HaloTag Ligand and/or Cy5 dye is lower or higher, please adjust the apply amount, thereby the molar ratio is roughly similar). If necessary, confirm the completion of reaction by TLC (take small amount for frozen dry to exchange reagent to methanol, then apply TLC. The developing liquid is 3:1 (or 1:1) mix of dichloromethane and methanol).
2. Incubate for 6.5 h in thermal cycler at 50 °C. If necessary, elongate the reaction time up to 18 h.
3. Mix 10 μl of step-1 product (HaloTag Ligand-Cy5), 15 μl of Biotin-AC5-hydrazide (10.3 mM), and 0.6 μl of 10× diluted Triethylamine (0.7 M, diluted by DMSO) in another PCR tube (molar ratio is 1:3.2:10.7:29.2 = Halo:Cy5:Biotin:Triethylamine).
4. Incubate for 18 h in thermal cycler at 50 °C.
5. Confirm the completion of reaction by TLC (take small amount for frozen dry to exchange reagent to methanol, then apply TLC. The developing liquid is 10:1 mix of dichloromethane and methanol). If the separation is not clear, optimize the ratio of dichloromethane and methanol.
6. Store in −30 °C until use.

3.1.2 Halo-Dicer-2 Expression

1. Centrifuge approximately 2×10^8 S2 cells in 50-ml tube at $1000 \times g$ for 3 min at room temperature.
2. Aspirate the supernatant, resuspend the cell pellet in 200 ml of antibiotics-free medium at 1×10^6 cells/ml, and transfer into 100-mm dishes (10 ml per each).
3. Mix 200 μl of 1 μg/μl Dcr-2 plasmid vector (pAHisHaloW-Dcr-2) and 10 ml of serum-free antibiotics-free medium. Then mix 400 μl of X-treamGENE HP and incubate for 30 min at room temperature.
4. Add 510 μl of the mixture to single dishes of S2 cells and gently swirl to make sure the solution is well mixed.
5. Incubate the cells for 72 h at 27 °C.

3.1.3 Halo-Dicer-2 Labeling

1. Centrifuge the cells in the four 50-ml tubes at $1000 \times g$ for 3 min at room temperature. After centrifugation, wash the pellet in 20 ml of 1× PBS and centrifuge it again.

2. Resuspend the cells in the same volume of ice-cold 1× lysis buffer containing 1 mM DTT and 1× protease inhibitor cocktail.
3. Transfer the cells into a pre-chilled Dounce homogenizer.
4. Homogenize the cells by 20 strokes on ice.
5. Transfer the homogenized sample to 1.5-ml tubes and centrifuge them at 17,000 × *g* for 20 min at 4 °C.
6. Transfer the supernatant to new 1.5-ml tubes and centrifuge them at 17,000 × *g* for 20 min at 4 °C to clear the lysate.
7. Collect the supernatant.
8. Mix 100 μl of the supernatant and 1 μl of 100 μM HaloTag Cy5-biotin ligand. Incubate for 30 min at 25 °C and then perform SDS-PAGE analysis to check the concentration of the Dcr-2 protein.
9. Removed free ligands by purification using cOmplete His-Tag Purification Resin.
10. Labeled proteins should be supplemented with 10% glycerol, 0.2 mg/ml BSA, shock-frozen, and stored at −80 °C.

3.2 Preparation of Fluorescently Labeled dsRNA Targets

3.2.1 Transcription and Annealing

1. RNAs are in vitro transcribed in a 20 μl reaction volume using T7 transcription kit according to manufacturer's instructions with some modifications: Transcription should be performed in the presence of 15 mM GMP to produce 5′ monophosphorylated RNAs. The concentrations of other NTPs should be reduced to 5 mM. For transcription of Cy3-RNA, UTP is omitted and instead 1.5 mM Cy3-UTP should be added.
2. The reaction mixtures are incubated at 37 °C for 3 h. After phenol/chloroform treatment and ethanol precipitation of the reaction mixtures, ssRNAs are dissolved in 50 μl purified water. The yield is approximately 2 μM (100 pmol) of Cy3-RNA and 5 μM (250 pmol) of non-labeled RNA (a complementary strand).
3. To prepare dsRNAs, pairs of ssRNAs (1 μM Cy3-RNA and 1.5-fold-excess amount of complementary strand of RNA) are annealed in a 20–50 μl volume by heating at 95 °C for 5 min in lysis buffer and cooling slowly.
4. The annealed dsRNA was purified by native PAGE, resulting in 50–100 μl of 200 nM dsRNA in lysis buffer (*see* Sasaki et al. [19] for a detailed method of PAGE purification).

3.3 Preparation of the PEG-Coated Quartz Slides and Coverslips (Method #1, Fig. 2, [16])

Please *see* **Note 3** for a description of the factors to consider when choosing a cleaning method.

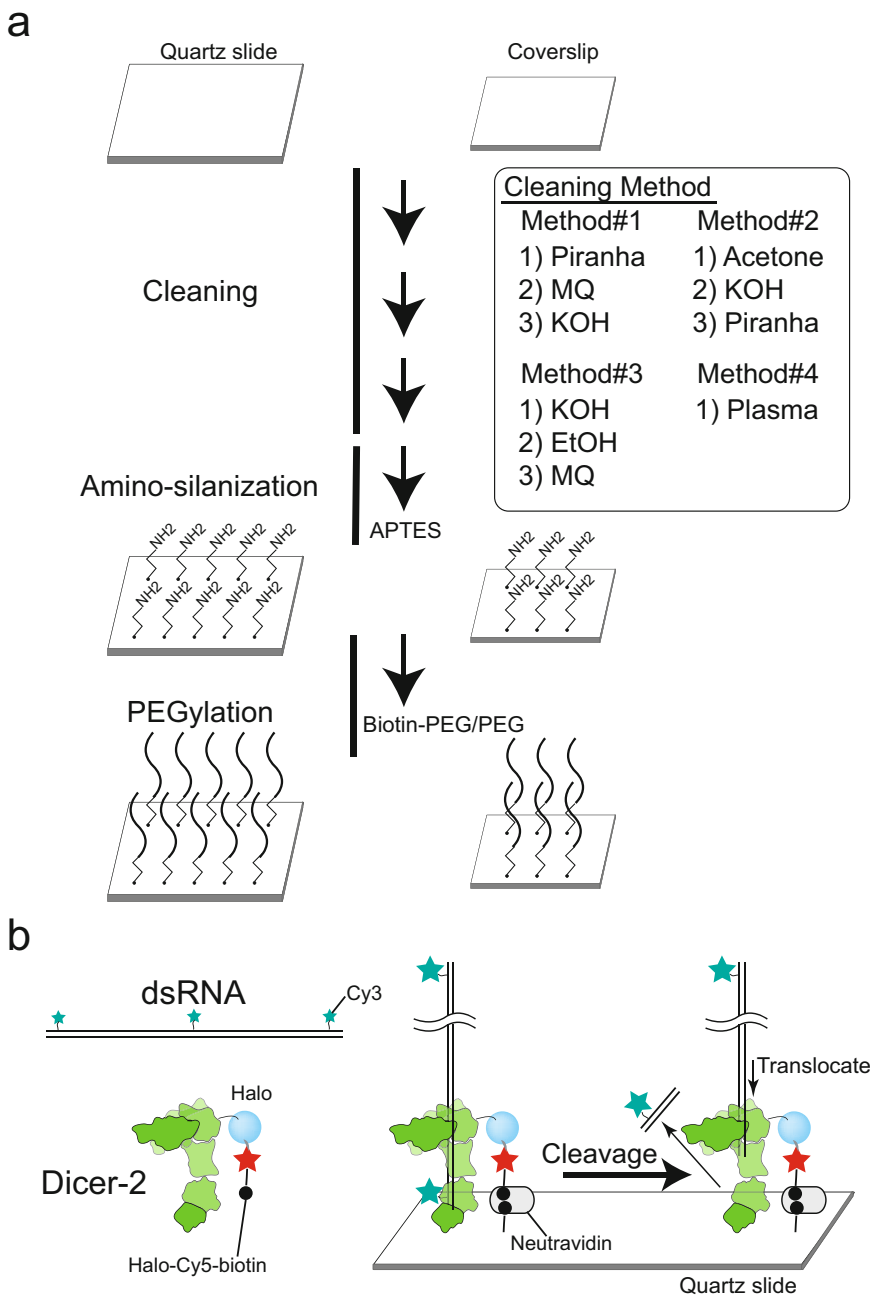


Fig. 2 Properties and rational design of the gene nano-chip activity. (**a**) Schematic illustration of glass cleaning and passivation process. *See* main text for details of the four different methods. (**b**) Schematic illustration of dicing assay. Both Dicer-2 and dsRNA are labeled by fluorescent dye, and one of the molecules (Dicer-2 in the figure) is tethered on the glass surface through biotin–avidin interaction. The partner molecule (dsRNA in the figure) freely diffuses in the reaction solution. After the association of dsRNA to Dicer-2, translocation and cleavage occur, resulting in the production of short dsRNA (21–22 nt) from long dsRNA precursor (220 nt). This single-molecule assay can trace whole the reaction at single-molecule level, thus making it possible to distinguish the cleavage mode of Dicer-2

3.3.1 Pre-cleaning of the Quartz Slides and Coverslips

1. (Optional, needed if reusing quartz glass) Clean the glass for 10 min in plasma cleaner. *See* also glass reuse method below.
2. Transfer eight slides and coverslips to a glass staining trough and place them in a beaker (2 L) that is located in a chemical hood.
3. Pour 70 ml H_2SO_4 into glass staining trough.
4. Pour 30 ml H_2O_2 into the trough. If H_2O_2 is kept in the refrigerator, the H_2O_2 solution should be warmed to room temperature in advance. Mix the 7:3 (v/v) H_2SO_4:H_2O_2 solution using a glass Komagome pipette. Upon mixing the H_2SO_4 and H_2O_2, the solution will start to boil spontaneously. Be careful not contaminate with acetone, which may cause an explosion [15].
5. Incubate for 30 min at room temperature.
6. Transfer the slides and coverslips to a glass staining trough filled with purified water, sonicate for 5 min.
7. Repeat the **step 6** once.
8. Use purified water to wash the slides and coverslips three times.
9. Place slides and coverslips in the glass staining trough filled with 0.5 M KOH, then sonicate for 20 min.
10. Place slides and coverslips in the glass staining trough filled with purified water, then sonicate for 5 min.
11. Repeat **step 10** twice.
12. Discard the Piranha solution into a designated waste once it reaches room temperature.

3.3.2 Amino-silanization of Slides and Coverslips

1. In the clean bench, mix 98 ml acetone and 2 ml amino-silane (APTES) in a large stainless steel deep vat.
2. Transfer the slides and coverslips (in purified water) to the glass staining trough containing the acetone, then discard acetone and pour the reaction mixture (APTES).
3. Incubate for 15 min at room temperature.
4. Pour the reaction mixture into a waste tank specified for acetone waste.
5. Use purified water to wash the slides and coverslips three times.
6. Sonicate for 5 min in acetone.
7. Sonicate for 5 min in purified water.
8. Refresh the purified water and repeat **step 7**.
9. Place the slides and coverslips on a clean tissue and dry them in a clean bench. A compressed air duster may also be used to blow the water away.

3.3.3 Surface Passivation Using Polymer

1. Prepare 0.55 M K_2SO_4/0.1 M carbonate-bicarbonate (pH 9.5) by dissolving 95.8 mg powder of K_2SO_4 with 0.1 M Sodium carbonate/bicarbonate buffer at pH 9.5 (mix 30 ml of 0.1 M Na_2CO_3 and 70 ml of $NaHCO_3$).
2. Dissolve 1 mg biotin-PEG in 20 μl 0.55 M K_2SO_4/0.1 M carbonate-bicarbonate (pH 9.5) to make a 50 mg/ml biotin-PEG solution.
3. Dissolve 20 mg 4 Arm-PEG in 200 μl 0.55 M K_2SO_4/0.1 M carbonate-bicarbonate (pH 9.5) to make a 100 mg/ml PEG solution. Become cloudy.
4. Mix 2 μl biotin-PEG solution and 100 μl PEG solution to make a 1% biotin-PEG/PEG solution.
5. Place 10 μl 1% biotin-PEG/PEG solution on the central part of each slide and then place a piece of parafilm to make the central area of the slide covered with the solution. This procedure should be performed in a clean bench.
6. Place 10 μl PEG solution on the central part of each coverslip and then place a piece of parafilm to make the center area of the coverslips covered with the solution. This procedure should be performed in a clean bench.
7. Incubate for 2 h at room temperature.
8. Before removing the piece of parafilm, mark the area covered with the solution to make sure where is PEG-coated. Also, write down some marks to distinguish the front and back sides.
9. Remove the piece of parafilm in fresh purified water and rinse the slides and coverslips thoroughly with 20 strokes (for ~30 s). Dry using compressed air.
10. Place the slides and coverslips on lab wipes and further dry in a clean bench for 15 min.
11. Place the coated slides and coverslips in the 50 ml plastic tubes and put the tubes in plastic bags, then seal the bag with a vacuum sealer.
12. Store them at −30 °C. The slides and coverslips may be kept for 1 month.

3.4 Preparation of the PEG-Coated Quartz Slides and Coverslips (Method #2, [15]) (See Note 4)

3.4.1 Pre-cleaning of the Quartz Slides and Coverslips

1. (Optional, needed in reusing quartz glass) Clean the glass for 10 min in plasma cleaner. *See* also glass reuse method below.
2. Place eight slides and eight coverslips in the glass staining trough filled with purified water and wash three times.
3. Pour the acetone into the glass staining trough, then sonicate for 30 min at room temperature.
4. Pour out the acetone, then use the purified water to wash them three times.

5. Place slides and coverslips in the glass staining trough filled with 1 M KOH, then sonicate for 30 min.
6. Use the purified water to wash slides and coverslips three times.
7. Transfer slides and coverslips to a glass staining trough and place them in a beaker (2 L) that is located in a chemical hood.
8. Pour 60 ml H_2SO_4 into glass staining trough.
9. Pour 20 ml H_2O_2. If H_2O_2 is kept in the refrigerator, the H_2O_2 solution should be warmed to room temperature in advance.
10. Incubate for 20 min at room temperature.
11. Transfer the slides and coverslips to the glass staining trough filled with purified water, then wash slides and coverslips three times with purified water.
12. Discard the H_2SO_4:H_2O_2 solution into a designated waste once it reaches room temperature.

3.4.2 Amino-silanization of Slides and Coverslips

1. In the clean bench, mix 120 ml methanol, 6 ml acetic acid, 3.6 ml amino-silane (APTES) in a large stainless deep vat.
2. Transfer the slides and coverslips (in purified water) to the glass staining trough containing the methanol, then discard methanol and pour the reaction mixture (APTES).
3. Incubate for 30 min at room temperature. During this process, sonicate them 1 min.
4. Discard the reaction mixture into a specific waste tank.
5. Use fresh methanol to wash the slides and coverslips.
6. Repeat **step 5** twice.
7. Place the slides and coverslips on a clean tissue and dry them in a clean bench. A compressed air duster may also be used to blow the water away.

3.4.3 Surface Passivation Using Polymer

1. Prepare 0.1 M of fresh sodium bicarbonate buffer (pH 8.5, no need of pH adjusting) by dissolving 84 mg of sodium bicarbonate in 10 ml of purified water. This solution can be frozen and stored at −20 °C.
2. Dissolve 1 mg biotin-PEG in 20 μl of 0.1 M fresh sodium bicarbonate buffer (pH 8.5) to make a 50 mg/ml biotin-PEG solution.
3. Dissolve 20 mg 4 Arm-PEG in 200 μl of 0.1 M fresh sodium bicarbonate buffer (pH 8.5) to make a 100 mg/ml PEG solution.
4. Mix 2 μl biotin-PEG solution and 100 μl PEG solution to make a 1% biotin-PEG/PEG solution.

5. Place 10 μl 1% biotin-PEG/PEG solution on the central part of each slide and then place a piece of parafilm to make the central area of the slide covered with the solution. This procedure should be performed in a clean bench.
6. Place 10 μl PEG solution on the central part of each coverslip and then place a piece of parafilm to make the center area of the coverslips covered with the solution. This procedure should be performed in a clean bench.
7. Incubate for at least 2 h at room temperature. Overnight incubation results in higher quality of surface passivation.
8. Before removing the piece of parafilm, mark the area covered with the solution to make sure where is PEG-coated. Also, write down some marks to distinguish the front and back sides.
9. Remove the piece of parafilm in fresh purified water and rinse the slides and coverslips thoroughly with 20 strokes (for ~30 s). Dry using compressed air.
10. Place the slides and coverslips on lab wipes and further dry in a clean bench for 15 min.
11. Place the coated slides and coverslips in the 50 ml plastic tubes and put the tubes in plastic bags, then seal the bag with a vacuum sealer.
12. Store them at −30 °C. The slides and coverslips may be kept for 1 month.
13. (Optional) Just before use, perform a second PEG treatment with short PEG (333 Da). 7 μl 250 mM MS(PEG)4 + 63 μl sodium bicarbonate buff. Place 10 μl short-PEG solution on the central part of each slide and then place a piece of parafilm to make the central area of the slide covered with the solution. This procedure should be performed in a clean bench. Incubate for 3 h to overnight. Then wash with purified water and dry using compressed air.

3.5 Preparation of the PEG-Coated Quartz Slides and Coverslips (Method #3, [18, 19]) (See Note 5)

3.5.1 Pre-cleaning of the Quartz Slides and Coverslips

1. (Optional, needed in reusing quartz glass) Clean the glass for 10 min in plasma cleaner. *See* also glass reuse method below.
2. Place eight slides and eight coverslips in the glass staining trough filled with 0.1 M KOH, and then sonicate for 10 min.
3. Discard KOH, use the MQ to wash the slides and coverslips three times.
4. Place slides and coverslips in the glass staining trough filled with 95% ethanol, and then sonicate for 10 min.
5. Discard the ethanol, use the purified water to wash the slides and coverslips three times.
6. Place slides and coverslips in the glass staining trough filled with purified water, and then sonicate for 10 min at room temperature.

3.5.2 Amino-silanization of Slides and Coverslips

1. In the clean bench, mix 100 ml methanol, 0.86 ml acetic acid, 2.12 ml amino-silane (APTES), and 4.25 ml purified water in a large stainless deep vat.
2. Transfer the slides and coverslips (in purified water) to the glass staining trough containing the reaction mixture (APTES).
3. Incubate them for 20 min at room temperature.
4. Discard APTES, and use purified water to wash the slides and coverslips three times.
5. Place the slides and coverslips on a clean tissue and dry them in a clean bench. A compressed air duster may also be used to blow the water away.

3.5.3 Surface Passivation Using Polymer

1. Prepare 50 mM MOPS-KOH (pH 7.5) by diluting 0.5 M MOPS-KOH (pH 7.5) (M.W. = 209.3; 5.23 g + 5 M KOH ~3 ml). This solution can be frozen and stored at −20 °C with avoiding light. If you find that the solution color turned to brown, make a new solution.
2. Dissolve 1 mg biotin-PEG in 20 μl of 50 mM MOPS-KOH (pH 7.5) to make a 50 mg/ml biotin-PEG solution.
3. Dissolve 20 mg 4 Arm-PEG in 200 μl of 50 mM MOPS-KOH (pH 7.5) to make a 100 mg/ml PEG solution.
4. Mix 2 μl biotin-PEG solution and 100 μl PEG solution to make a 1% biotin-PEG/PEG solution.
5. Place 10 μl 1% biotin-PEG/PEG solution on the central part of each slide and then place a piece of parafilm to make the central area of the slide covered with the solution. This procedure should be performed in a clean bench.
6. Place 10 μl PEG solution on the central part of each coverslip and then place a piece of parafilm to make the center area of the coverslips covered with the solution. This procedure should be performed in a clean bench.
7. Incubate for at least 2 h at room temperature. Overnight incubation results in higher quality of surface passivation.
8. Before removing the piece of parafilm, mark the area covered with the solution to make sure where is PEG-coated. Also, write down some marks to distinguish the front and back sides.
9. Remove the piece of parafilm in fresh purified water and rinse the slides and coverslips thoroughly with 20 strokes (for ~30 s). Dry using compressed air.
10. Place the slides and coverslips on lab wipes and further dry in a clean bench for 15 min.
11. Place the coated slides and coverslips in the 50 ml plastic tubes and put the tubes in plastic bags, then seal the bag with a vacuum sealer.

12. Store them at −30 °C. The slides and coverslips may be kept for 1 month.
13. (Optional) Just before use, second PEG treatment with short PEG (333 Da). 7 μl 250 mM MS(PEG)4 + 63 μl sodium bicarbonate buff. Place 10 μl short-PEG solution on the central part of each slide and then place a piece of parafilm to make the central area of the slide covered with the solution. This procedure should be performed in a clean bench. Incubate for 3 h to overnight. Then wash with purified water and dry using compressed air.

3.6 Preparation of the PEG-Coated Quartz Slides and Coverslips (Method #4) (See Note 6)

1. Place four slides into the plasma machine to treat the slides (10 min plasma cleaning).
2. Place the other four slides into the plasma machine to treat the slides.
3. Place four coverslips into the plasma machine to treat the coverslips.
4. Place the other four coverslips into the plasma machine to treat the coverslips.

3.6.1 Pre-cleaning of the Quartz Slides and Coverslips

3.6.2 Amino-silanization of Slides and Coverslips

Same as method #1.

3.6.3 Surface Passivation Using Polymer

Same as method #1.

3.7 Clean and Reuse of Quartz Slides

Quartz slides are slightly expensive, but we can clean and reuse them until the quality of the glass becomes worse (cloudy structure appears after many time of reuse). Alternatively, regeneration of PEG-slide is also available [20].

1. Put used quartz slides into glass stand (max 19 slides).
2. Wash by tap water.
3. Pour boiled water (you can use plastic beaker to bring, or microwave the whole glass stand to heat). The glue (adhesive paste) should then become white and peel off (In some case, the glass stand breaks, so take care not to pour the hot water directly to the glass stand. Please pour into the water part).
4. Add detergent (e.g., DCN90).
5. Wait half day.
6. Sonicate for 60 min (until the glue (adhesive paste) peels off).
7. Wash carefully with tap water (if possible hot water is better).
8. Move the quartz slides into new glass stand with purified water using tweezers. (Take care not to contaminate the coverslips. After you move all the quartz glass, discard coverslips.)

9. Wash with purified water (Directly from water purification machine is better).
10. Move the quartz slide into acetone-filled glass stand, then sonicate for 20 min using sonicator water bath (If you use acetone three times, discard and pour new acetone into glass stand).
11. Move the quartz slide into purified water-filled glass stand, then wash with purified water.
12. Move the quartz slide into 0.1 M KOH-filled glass stand, then sonicate for 20 min.
13. Wait overnight.
14. Move the quartz slide into purified water-filled glass stand, keep until use.
15. (Optional) After dry in clean bench, use plasma cleaner 10 min to reduce the dust.

3.8 Single-Molecule Imaging (Fig. 2) (See Note 7)

3.8.1 Preparation of Single-Molecule Observation Chamber

1. Make sure which side of the slide and coverslip is PEG-coated.
2. Place the double-sided tape on the flat bottom (back side) of an aluminum block and cool it for 15 s. This step allows the protective plastic foil to be removed easily.
3. Apply the double-sided tape on the PEG-coated side of the coverslip and remove the other protective foil.
4. Invert the coverslip to make the PEG-coated side down and place it onto the PEG-coated side of the slide. Use tweezers to gently tap the coverslip to press it down. Then invert the chamber to make the coverslip side down.
5. The sample chamber volume is ~15 μl (depend on the thickness of double-sided tape). Fluid can be flown through the chamber by placing the pipette tip in one hole and gently expelling the liquid from the pipette through the chamber. Use stripes of filter paper to suck the excess of liquid out from the other hole of the chamber.

3.8.2 Continuous Monitoring of Spot Appearance

1. Before the experiments, the excitation lights are aligned to generate evanescent field. Make sure that the incident lights are totally internally reflected at the quartz-water interface and are well overlapped. The critical angle for the interface between water and quartz is ~65.6° for visible light. In our experiments, the incident angle of excitation light at the quartz-water surface was set to be 69° to make the evanescent field thin. We continuously irradiated the laser light, but using mechanical shutter system (e.g., UNIBLITZ, VMM-D3 controller and LS2S2T1 shutter) or digital modulation of laser power (TTL trigger is provided by microcontroller (e.g., Arduino) or functional generator), alternating-laser excitation (ALEX) scheme can also be

performed, where synchronization of the laser excitation and camera capture is also achieved (*see* EMCCD manual for detail). The frame rate is set to 1 frame per second (fps). The temperature of EMCCD is set to −85 °C.

2. The observation mixture (0.03 U/ml creatine kinase, 0.1 U/ml RNasin Plus, 10 mM MgOAc. 1% Biolipidure-203 (NOF Corporation), 1 mM TCEP, 5 mM protocatechuic acid (PCA), 50 nM protocatechuate-3,4-dioxygenase (PCD), 5 mM Trolox in 1× lysis buffer) was pre-mixed at 25 °C (*see* **Note 8**).
3. Flow into the chamber 15 μl of 2.5 mg/ml NeutrAvidin solution and incubate for 2 min. Wash the chamber with 50 μl of 1× lysis buffer.
4. Flow into the chamber 20 μl of a 1:300 dilution of Cy5 biotin-labeled Dcr-2 (f. 1.7–3 nM) by 1× lysis buffer. Incubate for 2 min.
5. Washed twice with 1× lysis buffer containing 1 mM TCEP and rinsed with the observation mixture.
6. Infuse the observation mixture containing 2 nM Cy3-labeled BLT or 3′ovr dsRNA and 1 mM ATP.
7. Images were continuously taken for 600 s at a frame rate of 1 frames/s.
8. Select and image three random and non-overlapping places from one flow chamber. Save the images as FITS (Flexible Image Transport System) format. FITS files can be opened by ImageJ without additional plugins.

3.8.3 Data Processing and Image Analysis

1. Open the FITS file as a stack on ImageJ.
2. Create an average intensity projection of the first 50 frames in stack (*see* **Note 9**).
3. Automatically picked up the Cy5 fluorescent spots (Dicer-2) by using a custom-made macro on ImageJ.
4. The Cy3 spots (dsRNA) that displayed co-localization with Cy5 within the entire view field (512 × 256 pixels; 1450 μm^2) were analyzed. And the integrated intensity traces were generated to identify cleavage events of Cy3-labeled dsRNAs.
5. Referencing idealized trace generated by vbFRET, a hidden Markov model based analysis software, the duration times of cleavage events were determined by finding the events of the Cy3 signals. As vbFRET is a package on MATLAB, researcher can modify the code to define the output data format. Excel and KaleidaGraph were used to summarize the data and to draw the graph.

4 Notes

1. For Dcr-2-anchored single-molecule experiments, a 222-nt template DNA was designed based on a random sequence composed of three nucleotides (TGC), and only three adenines (A) were present at the specific positions. By transcription, three Cy3-Us were incorporated into the above specific positions and removed after first, sixth, and tenth cleavage by dicer-2. A 222-nt double-stranded DNA (the above-mentioned template and complementary DNAs) was chemically synthesized and cloned into pUC57 by GenScript.
2. In some case, fusion of SNAP-Tag protein is better than fusion of Halo Tag protein due to the difference of pI (6.1 vs. 4.9) and the size (19 vs. 34 kDa). To label SNAP-tag fused protein, please use BG-Cy5-biotin instead of Halo-Cy5-biotin.
3. Currently we mainly used this method #1 in our lab. If you cannot use H_2SO_4 and/or H_2O_2, please use method #3 below (use KOH for cleaning). Method #2 is a popular method used in single-molecule imaging. The main difference in method #3 is the cleaning order and the solution to dissolve the PEG. If you prefer the simplest cleaning, please select method #4 (instead of chemical etching of the glass, use plasma cleaner solely), although the glass passivation quality is slightly low.
4. Method #2 is a popular method used in the single-molecule imaging.
5. Method #3 is a method that we have used for long time.
6. Method #4 is the simplest glass cleaning method.
7. Please also refer our previous literature [19].
8. Glucose, glucose oxidase, catalase (GOC) system can also be used for many of the single-molecule experiments.
9. In the microscopic observation, mechanical drift in the XY plane as well as focus drift in the Z axis is a severe problem for long time monitoring. To overcome this problem, two approaches—hardware and software—can be used. For the hardware improvement, an ultrastable stage, a motorized Z-drift compensator and/or a rigid microscope body will be helpful. Also minimizing the room temperature change, which affects the metal heat extension (length) and thus causes the stage, microscope body and/or mirror drift, will improve the stability. For the post-production improvement, drift markers are useful. We used a non-specific aggregate as a drift marker for XY drift correction. Biotinylated fluorescent beads, Qdots, or gold nanoparticles could also be used as a drift marker. Also, recently developed DNA origami based ruler would be a good candidate for marker, where defined molecular layout allows to use DNA origami as a quantitative marker for localization and for fluorescent intensity.

Acknowledgements

We thank Y. Iwasaki for editing. This work was supported in part by Grants-in-Aid for Scientific Research (S) (18H05271 to Y.T.), Grant-in-Aid for Scientific Research (B) (19H03197 to H.T.), and Grant-in-Aid for Young Scientists (18K14649 to M.N.) from Japan Society for the Promotion of Science (JSPS); and by startup fund from ShanghaiTech University (to H.T.).

References

1. Kobayashi H, Tomari Y (2016) RISC assembly: coordination between small RNAs and argonaute proteins. Biochim Biophys Acta 1859(1):71–81
2. Iwasaki S, Sasaki HM, Sakaguchi Y, Suzuki T, Tadakuma H, Tomari Y (2015) Defining fundamental steps in the assembly of the drosophila RNAi enzyme complex. Nature 521(7553): 533–536
3. Yao C, Sasaki HM, Ueda T, Tomari Y, Tadakuma H (2015) Single-molecule analysis of the target cleavage reaction by the Drosophila RNAi enzyme complex. Mol Cell 59(1): 125–132
4. Tsuboyama K, Tadakuma H, Tomari Y (2018) Conformational activation of Argonaute by distinct yet coordinated actions of the Hsp70 and Hsp90 chaperone systems. Mol Cell 70(4): 722–729.e4
5. Salomon WE, Jolly SM, Moore MJ, Zamore PD, Serebrov V (2015) Single-molecule imaging reveals that argonaute reshapes the binding properties of its nucleic acid guides. Cell 162(1):84–95
6. Yanagida T, Nakase N, Nishiyama K, Oosawa F (1984) Direct observation of motion of single F-actin filaments in the presence of myosin. Nature 307(5946):58–60
7. Kron SJ, Spudich JA (1986) Fluorescent actin filaments move on myosin fixed to a glass surface. Proc Natl Acad Sci U S A 83(17): 6272–6276
8. Harada Y, Noguchi A, Kishino A, Yanagida T (1987) Sliding movement of single actin filaments on one-headed myosin filaments. Nature 326(6115):805–808
9. Harada Y, Sakurada K, Aoki T, Thomas DD, Yanagida T (1990) Mechanochemical coupling in actomyosin energy transduction studied by in vitro movement assay. J Mol Biol 216(1): 49–68
10. Toyoshima YY, Kron SJ, McNally EM, Niebling KR, Toyoshima C, Spudich JA (1987) Myosin subfragment-1 is sufficient to move actin filaments in vitro. Nature 328(6130): 536–539
11. Zhuang X, Bartley LE, Babcock HP, Russell R, Ha T, Herschlag D, Chu S (2000) A single-molecule study of RNA catalysis and folding. Science 288(5473):2048–2051
12. Taguchi H, Ueno T, Tadakuma H, Yoshida M, Funatsu T (2001) Single-molecule observation of protein-protein interactions in the chaperonin system. Nat Biotechnol 19(9):861–865
13. Ha T, Rasnik I, Cheng W, Babcock HP, Gauss GH, Lohman TM, Chu S (2002) Initiation and re-initiation of DNA unwinding by the Escherichia coli Rep helicase. Nature 419(6907): 638–641
14. Roy R, Hohng S, Ha T (2008) A practical guide to single-molecule FRET. Nat Methods 5(6):507–516
15. Chandradoss SD, Haagsma AC, Lee YK, Hwang JH, Nam JM, Joo C (2014) Surface passivation for single-molecule protein studies. J Vis Exp 86:50549
16. Park SR, Hauver J, Zhang Y, Revyakin A, Coleman RA, Tjian R, Chu S, Pertsinidis A (2020) A single-molecule surface-based platform to detect the assembly and function of the human RNA polymerase II transcription machinery. Structure 28(12):1337–1343.e4
17. Koopmans WJA, Schmidt T, Noort JV (2008) Nucleosome immobilization strategies for single-pair FRET microscopy. ChemPhysChem 9(14):2002–2009
18. Zhou ZP, Shimizu Y, Tadakuma H, Taguchi H, Ito K, Ueda T (2011) Single molecule imaging of the trans-translation entry process via anchoring of the tagged ribosome. J Biochem 149(5):609–618
19. Sasaki HM, Tadakuma H, Tomari Y (2018) Single-molecule analysis for RISC assembly and target cleavage. Methods Mol Biol 1680: 145–164
20. Paul T, Ha T, Myong S (2021) Regeneration of PEG slide for multiple rounds of single-

molecule measurements. Biophys J 120(9): 1788–1799

21. Kodera N, Yamamoto D, Ishikawa R, Ando T (2010) Video imaging of walking myosin V by high-speed atomic force microscopy. Nature 468(7320):72–76
22. Han BG, Watson Z, Kang H, Pulk A, Downing KH, Cate J, Glaeser RM (2016) Long shelf-life streptavidin support-films suitable for electron microscopy of biological macromolecules. J Struct Biol 195(2):238–244
23. Kasinath V, Beck C, Sauer P, Poepsel S, Kosmatka J, Faini M, Toso D, Aebersold R, Nogales E (2021) JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. Science 371(6527):eabc3393
24. Yan H, Park SH, Finkelstein G, Reif JH, LaBean TH (2003) DNA-templated self-assembly of protein arrays and highly conductive nanowires. Science 301(5641):1882–1884
25. Masubuchi T, Endo M, Iizuka R, Iguchi A, Yoon DH, Sekiguchi T, Qi H, Iinuma R, Miyazono Y, Shoji S, Funatsu T, Sugiyama H, Harada Y, Ueda T, Tadakuma H (2018) Construction of integrated gene logic-chip. Nat Nanotechnol 13(10):933–940
26. Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15(8): 509–524
27. Naganuma M, Tadakuma H, Tomari Y (2021) Single-molecule analysis of processive double-stranded RNA cleavage by Drosophila Dicer-2. Nat Commun 12(1):4268
28. Miyazono Y, Hayashi M, Karagiannis P, Harada Y, Tadakuma H (2010) Strain through the neck linker ensures processive runs: a DNA-kinesin hybrid nanomachine study. EMBO J 29(1):93–106
29. Yildiz A, Forkey JN, Mckinney SA, Ha T, Goldman YE, Selvin PR (2003) Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. Science 300(5628):2061–2065
30. Thompson RE, Larson DR, Webb WW (2002) Precise nanometer localization analysis for individual fluorescent probes. Biophys J 82(5): 2775–2783
31. Nicovich PR, Walsh J, Bocking T, Gaus K (2017) NicoLase-an open-source diode laser combiner, fiber launch, and sequencing controller for fluorescence microscopy. PLoS One 12(3):e0173879
32. Huang B, Jones SA, Brandenburg B, Zhuang X (2008) Whole-cell 3D STORM reveals interactions between cellular structures with nanometer-scale resolution. Nat Methods 5(12):1047–1052
33. Kao C, Zheng M, Rudisser S (1999) A simple and efficient method to reduce nontemplated nucleotide addition at the 3 terminus of RNAs transcribed by T7 RNA polymerase. RNA 5(9): 1268–1272
34. Aitken CE, Marshall RA, Puglisi JD (2008) An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments. Biophys J 94(5):1826–1835
35. Cordes T, Vogelsang J, Tinnefeld P (2009) On the mechanism of Trolox as antiblinking and antibleaching reagent. J Am Chem Soc 131(14):5018–5019
36. Bronson JE, Fei J, Hofman JM, Gonzalez RL Jr, Wiggins CH (2009) Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. Biophys J 97(12): 3196–3205

Chapter 14

Low Input Genome-Wide DNA Methylation Analysis with Minimal Library Amplification

Wan Kin Au Yeung and Hiroyuki Sasaki

Abstract

Whole genome bisulfite sequencing (WGBS) is a high-throughput DNA sequencing-based technique that is used to determine genome-wide DNA methylation patterns at base resolution. Library construction by post-bisulfite adaptor tagging (PBAT) extends the application of WGBS to several hundred cells and minimizes the required number of library amplification cycles. We herein describe a PBAT protocol to prepare WGBS libraries from 200 cells and introduce the outline of a downstream bioinformatic analysis. The prepared library can typically generate 800 million sequencing reads, which is sufficient to cover the human and mouse genomes approximately 15 times, using the Illumina NovaSeq 6000 sequencing system.

Key words DNA methylation, Whole genome bisulfite sequencing, Post-bisulfite adaptor tagging, Low input, 5-Methylcytosine, 5-Hydroxymethylcytosine, NovaSeq, Bioinformatics, Data analysis

1 Introduction

Piwi-interacting RNAs and long noncoding RNAs are often involved in the regulation of DNA methylation. Whole genome bisulfite sequencing (WGBS) is a high-throughput DNA sequencing-based technique that is used to determine genome-wide DNA methylation patterns (distribution and amount of 5-methylcytosine and 5-hydroxymethylcytosine) at base resolution. Bisulfite treatment of DNA converts unmethylated cytosine to uracil, which is subsequently converted to thymine during library preparation. In contrast, 5-methylcytosine and 5-hydroxymethylcytosine are resistant to bisulfite conversion and remain as cytosine in the WGBS library. Since 5-hydroxymethylcytosine constitutes an extremely small fraction in most cell types, cytosines in the sequencing reads are normally considered as 5-methylcytosines (*see* **Note 1**).

The original WGBS protocol (MethylC-seq) suffers from low library yield due to fragmentation of adaptor-ligated DNA during

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_14,

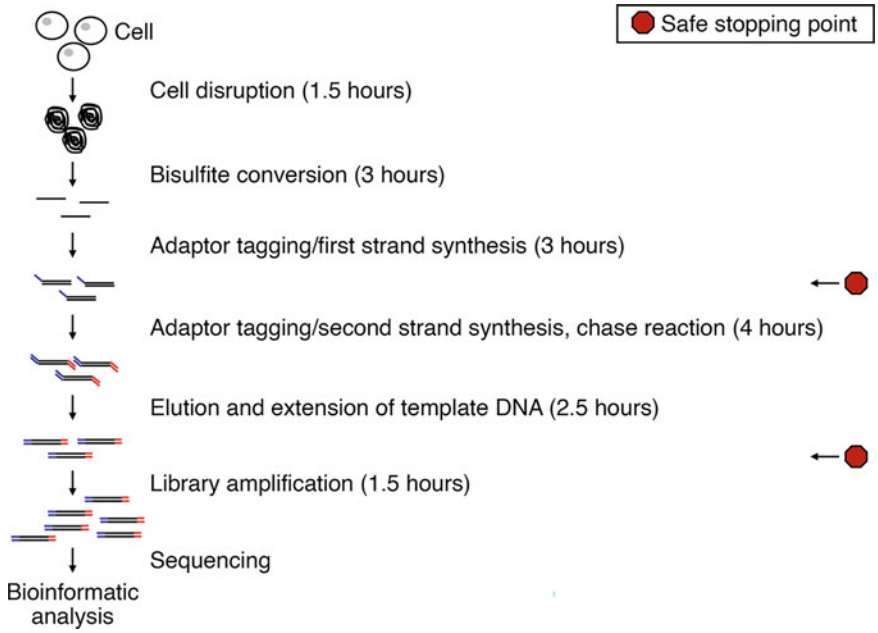


Fig. 1 Schematic workflow of low input WGBS library preparation. The workflow time for each major step and safe stopping points are indicated

bisulfite conversion, and thus is not suitable for studying small number of cells. To overcome this issue, Miura et al. developed a WGBS library construction method called post-bisulfite adaptor tagging (PBAT), which uses random priming to tag the DNA fragments with adaptors after bisulfite conversion [1]. This method minimizes the detrimental side effect of bisulfite treatment. Coupled with additional rounds of strand synthesis and multiple cycles of library amplification, the PBAT method has made it possible to generate WGBS libraries from several hundred cells to even a single cell [2–4] (*see* **Note 2**).

We herein describe a PBAT-based low input WGBS protocol with only 4 cycles of amplification for 200 cells [5, 6], or a total of 8 cycles for 30 cells (Fig. 1). The prepared library can typically generate 800 million sequencing reads, which is enough to cover the human and mouse genomes approximately 15 times, using the Illumina NovaSeq 6000 sequencing system. We also outline the bioinformatic analysis pipeline of WGBS data in the last part of this chapter.

2 Materials

Prepare all solutions using ultrapure water and analytical grade reagents. The reagents and kits used in this experiment must not be shared with other experiments (like genotyping) to avoid contamination.

2.1 Commercial Kits, Enzymes, and Consumables

1. EZ DNA Methylation-Gold Kit (Zymo Research).
2. KAPA HiFi HotStart Library Amplification Kit (Roche).
3. KAPA Library Quantification Kit (Universal for Illumina Platforms) (Roche).
4. Klenow Fragment, 3′ → 5′ exo- (50,000 units/ml) (New England Biolabs) (*see* **Note 3**).
5. *Bst* DNA Polymerase Large Fragment (8000 units/ml) (New England Biolabs).
6. Exonuclease I, *E. coli* (New England Biolabs).
7. Phusion Hot Start II DNA Polymerase (2 U/μl) (Thermo Scientific).
8. Unmethylated Lambda DNA (Promega).
9. AMPure XP (Beckman Coulter).
10. Dynabeads M-280 Streptavidin (Invitrogen).
11. 10× ExTaq Buffer (Takara Bio) or equivalent 10× PCR buffer without dye.
12. Silicone-coated flat bottom 1.5 ml centrifuge tubes.
13. 8-well 0.2 ml tube strips.
14. Low retention pipette tips (filter tips preferred).

2.2 Stock Solutions

1. Ultrapure water: Autoclave in advance.
2. 1 M Tris-acetate, pH 8.0: Autoclave in advance.
3. 1 M Tris-hydrochloride (HCl), pH 8.0: Autoclave in advance.
4. 1 M Tris-HCl, pH 7.5: Autoclave in advance.
5. 10 N sodium hydroxide (NaOH).
6. 2× BW(Li) solution: Add 0.1 ml of 500 mM ethylenediaminetetraacetic (EDTA), pH 8.0 and 0.5 ml of 1 M Tris-hydrochloride, pH 8.0 to 40 ml of ultrapure water. Add 6.3 g of anhydrous lithium chloride to the solution and wait until the exothermic dissolution is completed. Fill up to 50 ml with ultrapure water.
7. 10% sodium dodecyl sulfate (SDS).
8. 20 mg/ml proteinase K.
9. Absolute ethanol.
10. dNTP mixture (2.5 mM each of dATP, dCTP, dGTP, and dTTP).

2.3 Working Solutions

1. CT conversion reagent: The CT conversion reagent, M-Dissolving Buffer, and M-Dilution Buffer are provided in the EZ DNA Methylation-Gold Kit. Add 900 μl of ultrapure water, 50 μl of M-Dissolving Buffer, and 300 μl M-Dilution

Buffer to a vial of CT conversion reagent. Vortex the vial for at least 30 min at room temperature (*see* **Note 4**).

2. M-Wash Buffer: Take a bottle of M-Wash Buffer from the EZ DNA Methylation-Gold Kit, add 24 ml of absolute ethanol, and mix thoroughly.
3. 2% SDS: Dilute 10 μl of 10% SDS with 40 μl of ultrapure water.
4. 80% ethanol (*see* **Note 4**): Dilute 1 ml of absolute ethanol with 4 ml of ultrapure water.
5. 10 mM Tris-acetate, pH 8.0: Dilute 10 μl of 1 M Tris-acetate, pH 8.0 with 900 μl of ultrapure water.
6. 10 mM Tris-HCl, pH 8.0: Dilute 10 μl of 1 M Tris-HCl, pH 8.0 with 900 μl of ultrapure water.
7. 10 mM Tris-HCl, pH 7.5: Dilute 10 μl of 1 M Tris-HCl, pH 7.5 with 900 μl of ultrapure water.
8. 0.1 N NaOH (*see* **Note 4**): Dilute 10 μl of 10 N NaOH with 900 μl of ultrapure water.

2.4 Adaptors and Primers (100 μM, Oligonucleotide Purification Column Grade)

1. Bio-PEA2-N4: 5′-biotin-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN N-3′.
2. PE-reverse-N4: 5′-CAA GCA GAA GAC GGC ATA CGA GAT NNN N-3′.
3. Primer 3: 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3′.
4. PE-forward primer: 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC AC-3′.
5. PE-reverse primer: 5′-CAA GCA GAA GAC GGC ATA CGA GAT-3′.

2.5 Equipment

1. Heat block with a heated lid.
2. Thermocycler.
3. Real-time quantitative thermocycler.
4. Aluminum block for 0.2 ml tubes.
5. DynaMag-2 Magnet or equivalent (Invitrogen).
6. FastGene MagnaStand 0.2 (Nippon Genetics) or equivalent.
7. Vortex mixer.
8. Tube rotator.

3 Methods

Carry out all procedures at room temperature (20–25 °C) unless otherwise specified. Use only low retention pipette tips and silicone-coated 1.5 ml tubes throughout the experiments. Mixing

should be done by pipetting at least 10 times, and air bubbles should be avoided.

3.1 Cell Disruption

1. Thaw frozen cells (*see* **Note 5**) on ice.
2. Estimate the amount of genomic DNA in the sample, assuming 6 pg per mouse diploid cell.
3. Spike-in unmethylated lambda DNA equal to 1/100 or 1/200 of the amount of genomic DNA of the sample. Fill up the tube of cells to 18 μl with ultrapure water.
4. Add 1 μl each of 2% SDS and 20 mg/ml proteinase K. Mix well by pipetting. Incubate in a heat block for an hour at 37 °C and for 15 min at 98 °C.
5. Place the tube on ice and proceed to bisulfite conversion immediately.

3.2 Bisulfite Conversion

The preparation of working solutions of CT conversion reagent and M-Wash Buffer is already described in Subheading 2.3. Other buffers used in this section are included in the EZ DNA Methylation-Gold Kit.

1. Prepare a fresh vial of CT conversion reagent according to Subheading 2.3. Add 130 μl of CT conversion reagent to each sample tube (20 μl). Mix well and aliquot 50 μl each to a total of 3 wells of a 0.2 ml tube strip.
2. Place the strip in a thermocycler with a heated lid, and run the following program:

   10 min at 98 °C.

   2.5 h at 64 °C.

   Hold at 4 °C.
3. Place a Zymo-Spin IC Column in a collection tube. Add 600 μl of M-Binding Buffer to the column.
4. Combine and transfer 150 μl of bisulfite-treated DNA (product of **step 2**) to the column. Mix by pipetting and close the cap. Centrifuge at 10,000 × *g* for 1 min.
5. Reload the flow through onto the same column. Centrifuge at 10,000 × *g* for 1 min.
6. Discard the flow through.
7. Add 100 μl of M-Wash Buffer with ethanol (*see* **Note 6**) to the column. Centrifuge at 10,000 × *g* for 1 min. Discard the flow through.
8. Add 200 μl of M-Desulfonation Buffer to the column. Leave it for 15 min at room temperature.
9. Centrifuge at 10,000 × *g* for 1 min. Discard the flow through.

10. Add 200 μl of M-Wash Buffer with ethanol to the column. Centrifuge at 10,000 × *g* for 1 min. Discard the flow through.
11. Repeat **step 10** once again.
12. Place the column in a 1.5 ml tube. Add 20 μl of M-Elution Buffer directly onto the white matrix inside the column. Leave it for 2 min at room temperature.
13. Centrifuge at 10,000 × *g* for 1 min. Discard the column and place the 1.5 ml tube on ice. Proceed to first strand synthesis immediately.

3.3 Adaptor Tagging/ First Strand Synthesis

1. Prepare the first strand synthesis reaction mixture in a 0.2 ml tube strip on ice:

Ultrapure water	16 μl
10× NEB Buffer 2	5 μl
2.5 mM dNTP mixture	5 μl
Bio-PEA2-N4	4 μl
Purified bisulfite-treated DNA	20 μl

2. Place the strip in a thermocycler with a heated lid, and set the following program:

   5 min at 94 °C.

   20 min* at 4 °C.

   Increase temperature from 4 °C to 37 °C at a rate of +1 °C/min (*see* **Note 7**).

   90 min at 37 °C.

   10 min at 70 °C.

   Hold at 4 °C.
3. Begin the program. After reaching 4 °C in the second step of the program (*denoted with an asterisk), leave the tube strip in the thermocycler for 5 min, and then pause the run.
4. Place the tube strip on a pre-chilled aluminum block and add 1.5 μl of Klenow Fragment, 3′ → 5′ exo- to the reaction mixture. Mix well by pipetting.
5. Place the strip back in the thermocycler and restart the program for the remaining 15 min of the second step.
6. Proceed to DNA purification, or store the strip at 4 °C for at most 1 day (*see* **Note 8**).

3.4 DNA Purification

1. Transfer the first strand synthesis reaction mixture (51.5 μl) to a 1.5 ml tube.
2. Add 50 μl of AMPure XP beads (*see* **Note 9**) to the mixture. Mix well by pipetting until the mixture becomes homogenous. Leave the tube for 10 min at room temperature.

3. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
4. Discard the supernatant using a micropipette without disturbing the magnetic beads.
5. Add 200 μl of 80% ethanol to the tube. Leave it at room temperature for 1 min.
6. Discard the supernatant using a micropipette. Dry the magnetic beads (roughly 3 min at room temperature) (*see* **Note 10**).
7. Remove the tube from the magnetic stand. Add 45 μl of 10 mM Tris-acetate, pH 8.0 to the tube. Mix well by pipetting and collect the solution at the bottom by spinning the tube briefly. Leave the tube for 2 min at room temperature.
8. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
9. Transfer the supernatant (45 μl) to a new 1.5 ml tube. Add 5 μl of 10× ExTaq Buffer. Mix well by pipetting and spin down briefly.
10. Add 50 μl of AMPure XP beads to the 50 μl mixture. Mix well by pipetting until the mixture becomes homogenous. Leave the tube for 10 min at room temperature.
11. Repeat **steps 3–6**.
12. Remove the tube from the magnetic stand. Add 50 μl of 10 mM Tris-acetate, pH 8.0 to the tube. Mix well by pipetting and spin down briefly. Leave the tube for 2 min at room temperature.
13. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
14. Transfer the supernatant (containing purified DNA) to a new 1.5 ml tube. Place the tube on ice.

3.5 Complexation of Biotinylated DNA with Streptavidin Beads

1. Vortex Dynabeads M-280 Streptavidin stock solution vigorously.
2. Transfer 20 μl of Dynabeads M-280 Streptavidin to a new 1.5 ml tube. Add 200 μl of 2× BW(Li) solution. Mix well by pipetting and collect the solution at the bottom by spinning the tube briefly. Place the tube on the magnetic stand and leave it for 2 min at room temperature.
3. Discard the supernatant using a micropipette without disturbing the beads. Resuspend the beads with 50 μl of 2× BW (Li) solution.
4. Add the purified DNA (from Subheading 3.4) to the beads. Mix well by pipetting.

5. Leave the tube for 30 min at room temperature with gentle rotation (*see* **Note 11**).
6. Spin down briefly. Place the tube on the magnetic stand and leave it for 2 min at room temperature.
7. Discard the supernatant using a micropipette. Add 180 μl of 2× BW(Li) solution to the tube. Leave it at room temperature for 2 min.
8. Discard the supernatant using a micropipette. Add 180 μl of 0.1 N NaOH to the tube. Leave it at room temperature for exactly 2 min.
9. Discard the supernatant using a micropipette. Add 180 μl of 0.1 N NaOH to the tube. Leave it at room temperature for 30 s.
10. Discard the supernatant using a micropipette.
11. Remove the tube from the magnetic stand. Resuspend the beads with 180 μl of 2× BW(Li) solution by pipetting. Spin down briefly. Place the tube on the magnetic stand and leave it for 2 min at room temperature.
12. Discard the supernatant using a micropipette.
13. Remove the tube from the magnetic stand. Resuspend the beads with 180 μl of 10 mM Tris-HCl, pH 7.5 by pipetting. Spin down briefly. Keep the tube at room temperature.

3.6 Adaptor Tagging/ Second Strand Synthesis

1. Prepare the second strand synthesis reaction mixture in a 1.5 ml tube on ice:

Ultrapure water	36 μl
10× NEB buffer 2	5 μl
2.5 mM dNTP mixture	5 μl
PE-reverse-N4	4 μl

2. Place the tube (beads resuspended with 10 mM Tris-HCl, pH 7.5) on the magnetic stand and leave it for 2 min at room temperature.
3. Discard the supernatant using a micropipette.
4. Remove the tube from the magnetic stand. Resuspend the beads with 50 μl of second strand synthesis reaction mixture by pipetting. Collect the solution at the bottom by spinning the tube briefly. Transfer all suspension to a 0.2 ml tube strip.
5. Place the strip in a thermocycler with a heated lid, and set the following program:

   5 min at 94 °C.

   20 min* at 4 °C.

Increase temperature from 4 °C to 37 °C at a rate of +1 °C/min (*see* **Note 7**).

30 min at 37 °C.

10 min at 70 °C.

Hold at 4 °C.

6. Begin the program. After reaching 4 °C in the second step of the program (*denoted with an asterisk), leave the tube strip in the thermocycler for 5 min, and then pause the run.
7. Place the 0.2 ml tube strip on a pre-chilled aluminum block and add 1.5 μl of Klenow Fragment, 3′ → 5′ exo- to the reaction mixture. Mix well by pipetting.
8. Place the strip back in the thermocycler and restart the program for the remaining 15 min of the second step.

3.7 Chase Reaction

1. Prepare the chase reaction mixture in a 1.5 ml tube on ice:

Ultrapure water	40 μl
10× ThermoPol buffer	5 μl
2.5 mM dNTP mixture	5 μl
Bst DNA polymerase large fragment	1 μl

2. Place the 0.2 ml tube strip on the magnetic stand and leave it for 2 min at room temperature.
3. Discard the supernatant using a micropipette without disturbing the beads.
4. Remove the strip from the magnetic stand. Resuspend the beads with 51 μl of chase reaction mixture by pipetting.
5. Place the strip in a thermocycler with a heated lid. Incubate for 30 min at 65 °C.

3.8 Elution and Extension of Template DNA

1. Prepare the elution and extension reaction mixture in a 1.5 ml tube on ice:

Ultrapure water	35 μl
5× Phusion HF buffer	10 μl
2.5 mM dNTP mixture	5 μl
Primer 3	0.4 μl
Phusion Hot Start II DNA Polymerase	1 μl

2. Place the 0.2 ml tube strip on the magnetic stand and leave it for 2 min at room temperature.
3. Discard the supernatant using a micropipette without disturbing the beads.

4. Remove the strip from the magnetic stand. Resuspend the beads with 51 μl of elution and extension reaction mixture by pipetting.
5. Place the strip in a thermocycler with a heated lid, and run the following program:

   5 min at 94 °C.

   15 min at 55 °C.

   30 min at 68 °C.

   Hold at 4 °C.
6. Collect the solution at the bottom by spinning the tube strip briefly. Place the tube strip on the magnetic stand and leave it for 2 min at room temperature.
7. Transfer all supernatant to a 1.5 ml tube. Add 1 μl of Exonuclease I and mix well by pipetting. Incubate in a heat block for 30 min at 37 °C and for 10 min at 70 °C.
8. Spin down briefly. Place the tube on ice.

3.9 DNA Purification

1. Add 50 μl of AMPure XP beads to the mixture. Mix well by pipetting until the mixture becomes homogenous. Leave the tube for 10 min at room temperature.
2. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
3. Discard the supernatant using a micropipette without disturbing the beads.
4. Add 200 μl of 80% ethanol to the tube. Leave it at room temperature for 1 min.
5. Discard the supernatant using a micropipette. Dry the magnetic beads (roughly 3 min at room temperature).
6. Remove the tube from the magnetic stand. Add 25 μl of 10 mM Tris-acetate, pH 8.0 to the tube. Mix well by pipetting and spin down briefly. Leave the tube for 2 min at room temperature.
7. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
8. Transfer the supernatant (containing purified library DNA) to a new 1.5 ml tube. Store the tube at −80 °C.

3.10 Library Amplification

1. Prepare the amplification mixture in a 1.5 ml tube on ice:

2× KAPA HiFi HotStart ReadyMix	25 μl
PE-forward primer	0.2 μl
PE-reverse primer	0.2 μl
Purified library DNA	24.6 μl

2. Place the strip in a thermocycler with a heated lid, and run the following program:

1 min at 98 °C	
1 min at 98 °C 1 min at 55 °C 1 min at 72 °C	4–8 cycles (*see* **Note 12**)
Hold at 4 °C	

3. Transfer all supernatant to a 1.5 ml tube and place the tube on ice.

3.11 DNA Purification

1. Add 50 μl of AMPure XP beads (*see* **Note 9**) to the amplification mixture. Mix well by pipetting until the mixture becomes homogenous. Leave the tube for 10 min at room temperature.
2. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
3. Discard the supernatant using a micropipette without disturbing the beads.
4. Add 200 μl of 80% ethanol to the tube. Leave it at room temperature for 1 min.
5. Discard the supernatant using a micropipette. Dry the magnetic beads (roughly 3 min at room temperature).
6. Remove the tube from the magnetic stand. Add 45 μl of 10 mM Tris-acetate, pH 8.0 to the tube. Mix well by pipetting and collect the solution at the bottom by spinning the tube briefly. Leave it for 2 min at room temperature.
7. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
8. Transfer the supernatant (45 μl) to a new 1.5 ml tube. Add 5 μl of 10× ExTaq Buffer. Mix well by pipetting and spin down briefly.
9. Add 50 μl of AMPure XP beads to the 50 μl mixture. Mix well by pipetting until the mixture becomes homogenous. Leave the tube for 10 min at room temperature.
10. Repeat **steps 2–5**.
11. Remove the tube from the magnetic stand. Add 20 μl of 10 mM Tris-acetate, pH 8.0 to the tube. Mix well by pipetting and spin down briefly. Leave the tube for 2 min at room temperature.
12. Place the tube on the magnetic stand and wait until the solution becomes clear (roughly 2 min).
13. Transfer the supernatant (containing purified library DNA) to a new 1.5 ml tube. Store the tube at −80 °C.

3.12 Library Quantification

1. Prepare the quantification mixture in a 0.2 ml tube strip on ice:

Ultrapure water	4 µl
2× KAPA SYBR FAST qPCR Master Mix	10 µl
Primer Premix (10×)	2 µl
10,000-fold diluted library (*see* **Note 13**) or quantification standards	4 µl

2. Place the strip in a real-time quantitative thermocycler with a heated lid, and run the following program:

1 min at 95 °C	
30 s at 95 °C 45 s at 65 °C (data acquisition)	35 cycles
Melt curve analysis	

3. Examine the melt curve of the library sample and confirm that there is a peak at around 78 °C (Fig. 2) (*see* **Note 14**).
4. Calculate the library concentration using the quantification standards provided in the KAPA Library Quantification Kit (*see* **Note 15**).

3.13 Library Dilution for NovaSeq Sequencing (SP and S1 Flow Cell, Xp Workflow)

1. For single-lane sequencing, 14.4 µl of 250 pM library solution is prepared (*see* **Note 16**). To do this, take an aliquot of the library in a new 1.5 ml tube, and dilute it to a final concentration of 250 pM by adding 10 mM Tris-acetate, pH 8.0 so that the final volume of the library solution will be 14.4 µl.

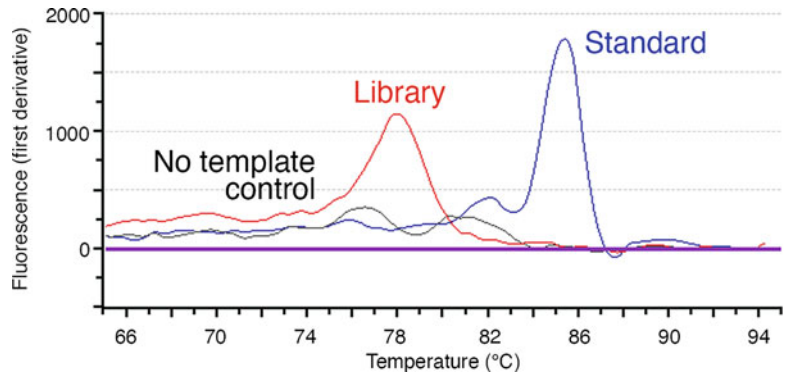


Fig. 2 A typical melt curve of a WGBS library prepared using this protocol. The library sample should have a peak at around 78 °C

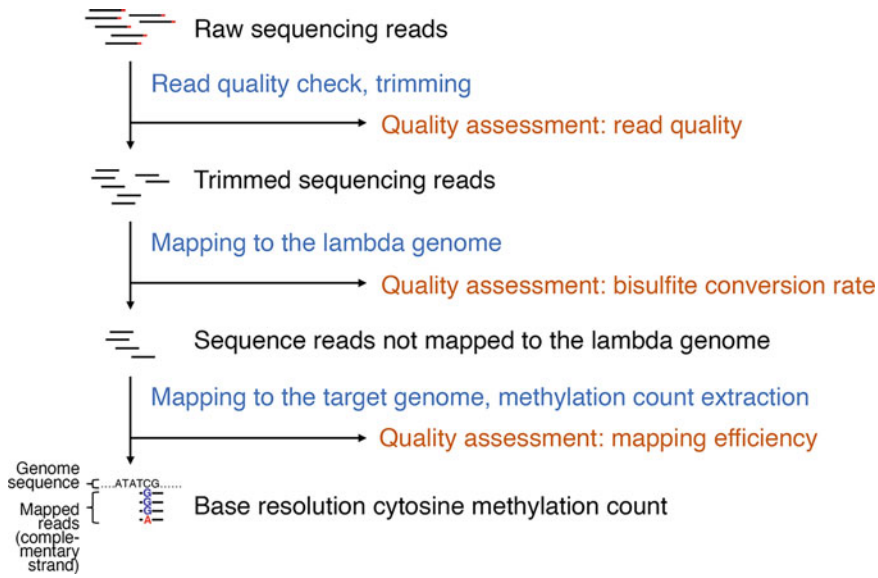


Fig. 3 Schematic overview of the bioinformatics analysis of WGBS data. Relevant data (black), major processing steps (blue), and quality assessment (orange) are indicated

2. Dilute 10 nM PhiX Sequencing Control V3 to 250 pM by adding 10 mM Tris-HCl, pH 8.0. Add 3.6 μl of the diluted PhiX Sequencing Control to 14.4 μl of the diluted library solution. Mix well by pipetting. Store the tube at −80 °C until the day of sequencing.
3. Perform sequencing according to manufacturer's instructions.

3.14 Bioinformatic Analysis of WGBS Data

This section introduces the bioinformatic analysis of WGBS data, describing how to calculate genome-wide DNA methylation levels based on the sequencing reads (Fig. 3). This workflow assumes the usage of the command line on a local linux platform but can be adapted to interactive cloud platforms as well (*see* **Note 17**).

1. The WGBS sequencing read files are gzipped files in the fastq format (*see* **Note 18**). The reads are derived from the genomes of lambda phage (*Escherichia virus Lambda*) and your target species (e.g., mouse, human, etc.), but cytosines appear as either cytosines (methylated) or thymines (unmethylated). They also appear as guanines (methylated) or adenines (unmethylated) on the other strand.
2. Check the quality of the sequencing reads using FastQC [7]. The sequence reads contain low-quality bases, especially in the regions near the 5′ and 3′ ends, and adaptor sequences at the 3′ end. Remove them using Trim Galore [8].
3. Map (= Find the genomic positions of the reads) the trimmed reads to the lambda genome sequence (unmethylated control)

(GenBank J02459.1) using Bismark [9]. This program aligns the reads to the in silico bisulfite converted genome sequences, as a vast majority of cytosines in the genome is unmethylated and converted to thymines in any species, referring to the original genome sequence to identify original cytosine positions. Make sure to include two options: --pbat and --un. While the original WGBS protocol (MethylC-seq) encodes the methylation status of each cytosine in the same strand, the PBAT protocol encodes it in the complementary strand. Therefore, the --pbat option instructs the program to determine the methylation status using the information of the complementary strand. Thus, if the base of the complementary strand appears as guanine, the cytosine is methylated; if that base appears as adenine, the cytosine is unmethylated. The --un option generates a new fastq file from the remaining reads that are not mapped to the lambda genome sequence. Assess the bisulfite conversion rate using the reads mapped to the lambda genome sequence (*see* **Note 19**).

4. Map the remaining sequencing reads to the target genome sequence using Bismark, in the same way as for the lambda genome. Make sure to include the --pbat option. Assess the mapping efficiency (*see* **Note 20**).
5. For each cytosine of the target genome, count the number of times being sequenced as methylated and unmethylated using Bismark methylation extractor. The output genome-wide cytosine report file (Fig. 4a) is used for the downstream analysis. In eukaryotes, cytosine methylation most frequently occurs at cytosine-guanine (CG) dinucleotides. However, it can also occur at other dinucleotide contexts (CH, where H is either adenine, cytosine, or thymine). The CG and CH methylation levels along a portion of the target genome can be visualized as a genome browser shot using Integrative Genomics Viewer [10] (Fig. 4b) (*see* **Note 21**).

4 Notes

1. Genome-wide detection of 5-hydroxymethylcytosine at base resolution can be done using Tet-assisted bisulfite sequencing (TAB-seq) [11]. Since the level of 5-hydroxymethylcytosine is generally low in most cell types, this direct detection method requires a much higher sequencing depth (usually 10 times) in comparison to regular WGBS methods. An alternative cost-effective approach is to detect both 5-methylcytosine + 5-hydroxymethylcytosine by a regular WGBS method and only 5-methylcytosine by oxidative bisulfite sequencing (oxBS-seq) [12] with comparable sequencing depths. Then, genome-wide

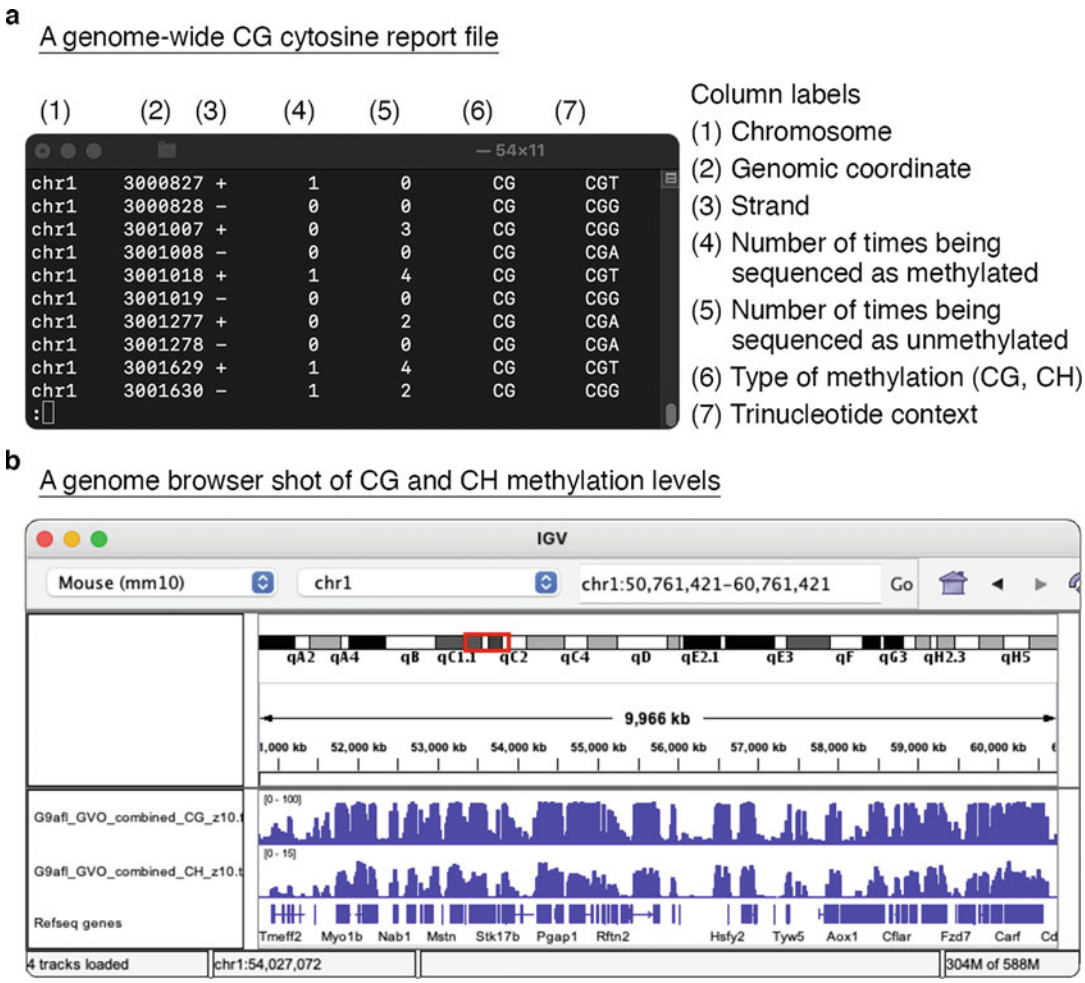


Fig. 4 Example of an output file and a genome browser shot. The control mouse fully grown oocyte data [5] is reprocessed as described above. (**a**) The first ten rows of the chromosome 1 genome-wide CG cytosine report file. Each row represents the methylation information of a single cytosine of CpG dinucleotides. The columns are tab-delimited. (**b**) A genome browser shot of the CG and CH methylation levels along chromosome 1. Refseq genes are included inside IGV by default

distribution of 5-hydroxymethylcytosine is estimated by subtracting the level of 5-methylcytosine (by oxBS-Seq) from that of 5-methylcytosine + 5-hydroxymethylcytosine at each CpG site (by a regular WGBS method). A commercial kit of oxBS-Seq is available from NuGEN Technologies (Tecan Group).

2. Recently, Miura et al. developed an improved PBAT protocol, which replaces the random priming step during adaptor tagging/second strand synthesis (*see* Fig. 1 for the outline of conventional PBAT) with terminal deoxyribonucleotidyl transferase-assisted adenylate connector-mediated single-stranded DNA (TACS) ligation [13]. TACS ligation attaches

the adaptor directly to the 3′-end of single-stranded DNA before the second strand synthesis. This TACS ligation-mediated protocol reduces the occurrence of unmappable chimeric reads when compared to conventional PBAT-based methods. We expect that this improved protocol will be incorporated into the existing PBAT-based methods and commercial kits.

3. Klenow Fragment (3′ → 5′ exo-) is offered in two versions with different concentrations of the enzyme. Only use the high concentration version (50,000 units/ml) in this protocol.
4. Prepare freshly on the day of experiment.
5. The cells (<1000 cells) should be collected in a single silicone-coated 1.5 ml centrifuge tube and stored at −80 °C. The buffer volume should be kept minimal (<5 μl). The following buffers have been validated for storage: M2 medium, KSOM medium and phosphate buffered saline.
6. Make sure you have added ethanol to reconstitute the M-Wash Buffer.
7. Instead of increasing the temperature at a rate of +1 °C/min, an increase of +0.5 °C per 30 s is also possible.
8. At this stage the sample DNA is double-stranded and stable. Excess primers will protect the sample DNA.
9. Bring AMPure XP to room temperature for 30 min prior to use.
10. The magnetic beads appear shiny soon after ethanol has been removed. At room temperature (20–25 °C), the magnetic beads will become dull and dry within 5 min. If the surface starts to crack, the magnetic beads are already over-dried. Proceed to the next step immediately.
11. Rotation minimizes the aggregation of the beads. Due to the viscosity of the 2× BW(Li) solution, the solution will always stay at the bottom of the tube.
12. To construct a library from 200 mouse diploid cells, a 4-cycle amplification yields a library suitable for single-lane sequencing on NovaSeq 6000 SP and S1 flow cell (i.e., 3600 amol). A minimal number of amplification cycles is preferred.
13. Perform serial 1:100 dilutions of the library. Since DNA readily binds to tube walls, leading to underestimation of the concentration, dilute the library only after the quantification mixture is ready. Do not re-use the diluted libraries.
14. A library prepared according to this protocol should have a unimodal distribution of fragment lengths with the mode at around 300 bp. This corresponds to a peak at around 78 °C in the melt curve. If this is not the case, the library should be discarded: make a new library again.

15. Since the quantification standard provided in the KAPA Library Quantification Kit has a single fragment size of 452 bp, accurate quantification of the library requires a correction based on the average fragment size. In our experience, however, sequencing works well without performing this correction.
16. We tested several final concentrations of PBAT-based WGBS libraries for NovaSeq sequencing. Based on our results, 250 pM is optimal, which is much lower than the loading concentrations suggested by Illumina for RNA-seq and ChIP-seq libraries (500 pM to 2.5 μM).
17. All tools described here are developed to run on a linux platform (CentOS, Ubuntu, etc.). The minimal system configuration that we have tested consists of a quad-core central processing unit (CPU), 8 GB random access memory (RAM), and 3 TB hard disk drive. The analysis typically requires several days to complete, depending on the number of sequencing reads and system configuration. It is possible to run the analysis on a macOS system: however, it is not recommended. For users who prefer a non-command-line option, interactive cloud platforms such as Galaxy and Illumina BaseSpace allow you to upload your data to a cloud server and perform an analysis through a web browser.
18. Descriptions of the fastq format can be found at the NCBI homepage (https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/#fastq-files).
19. Unmethylated lambda DNA is added as spike-in before bisulfite conversion (*see* Subheading 3.1). Due to imperfect conversion, a portion of cytosines of lambda DNA is not converted to thymine. The bisulfite conversion rate is used to monitor the efficiency of this reaction (calculated from methylation summary file bismark_SE_report.txt). A bisulfite conversion rate of around or above 99.5% is required for accurate estimation of methylation levels.
20. A typical PBAT-based WGBS sequencing data should have a mapping efficiency of 50–60%.
21. Integrative Genomics Viewer (IGV) is the software program used to visualize genome-wide data with reference to the genome sequences. The accompanying tool "igvtools" can generate IGV-compatible files (tdf format). Manually convert the output genome-wide cytosine report file to a bedGraph file. Then, use igvtools to convert the bedGraph file to the final tdf file.

Acknowledgments

This work was supported by a JSPS KAKENHI grant to H.S. (JP18H05214). We would like to thank Dr. Fumihito Miura and Dr. Kenjiro Shirane (Kyushu University) for their helpful discussion.

References

1. Miura F, Enomoto Y, Dairiki R, Ito T (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res 40:e136
2. Miura F, Ito T (2014) Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging. DNA Res 22:13–18
3. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods 11: 817–820
4. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, Bock C (2015) Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep 10:1386–1397
5. Au Yeung WK, Brind'Amour J, Hatano Y, Yamagata K, Feil R, Lorincz MC, Tachibana M, Shinkai Y, Sasaki H (2019) Histone H3K9 methyltransferase g9a in oocytes is essential for preimplantation development but dispensable for CG methylation protection. Cell Rep 27:282–293
6. Richard Albert J, Au Yeung WK, Toriyama K, Kobayashi H, Hirasawa R, Brind'Amour J, Bogutz A, Sasaki H, Lorincz M (2020) Maternal DNMT3A-dependent de novo methylation of the paternal genome inhibits gene expression in the early embryo. Nat Commun 11: 5417
7. Andrews S (2020) FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 7 Apr 2021
8. Krugger F (2019) Trim Galore. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 7 Apr 2021
9. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27(11):1571–1572
10. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26
11. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min JH, Jin P, Ren B, He C (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149(6):1368–1380
12. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science 336(6083):934–937
13. Miura F, Shibata Y, Miura M, Sangatsuda Y, Hisano O, Araki H, Ito T (2019) Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res 47(15):e85

Chapter 15

Solid-Support Directional (SSD) RNA-Seq as a Companion Method to CLIP-Seq

Abd-El Monsif Shawky, Mahmoud Dondeti, Zissimos Mourelatos, and Anastasios Vourekas

Abstract

CLIP-Seq (Deep Sequencing after in vivo Crosslinking and Immunoprecipitation, HITS-CLIP) has emerged as a key method for the study of RNA-binding proteins (RBPs), as it can scrutinize the RNAs bound by an RBP in vivo, with minimum manipulation of biological samples. CLIP-Seq is best used to reveal changes of the RNA cargo of an RBP and differences on binding patterns of the bound RNAs in living cells in different genetic backgrounds or after experimental treatment, rather than simply identifying RNA species. It is therefore crucial that a reference of the steady state levels of the RNAs present in the samples used for the CLIP-Seq experiment is included in the bioinformatic analysis. A simple directional RNA-Seq method was developed that uses the same oligonucleotides and the same PCR amplification steps as our CLIP-Seq method, which therefore can be analyzed using the same bioinformatic pipeline as the CLIP-Seq data. This greatly simplifies and streamlines the analysis process, and at the same time reduces the chances of protocol-specific artifacts and biases interfering with data interpretation. Some considerations on ways to integrate CLIP-Seq and RNA-Seq analyses are also provided herein.

Key words RNA-Seq, Strand-specific, Directional, Stranded, Transcriptomic analysis, Posttranscriptional RNA processing, HITS-CLIP, CLIP-Seq, RNA-IP, Next generation sequencing, Illumina, cDNA, RNA-binding protein, Ribonucleoprotein complexes, Argonaute, Piwi

1 Introduction

High Throughput Sequencing brought a revolution in the field of genomic and transcriptomic analyses. Methods to analyze the entire or a selected fraction of the transcriptome of a particular tissue or cell sample have revealed intricate and complex phenomena by which gene expression is regulated at the level of transcription and post-transcriptionally [1]. Proteins that bind RNA and/or act on it are responsible for mRNA regulation, and High Throughput Sequencing after Crosslinking and Immunoprecipitation (HITS-CLIP, CLIP-Seq) is now routinely employed to reveal mechanistic

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_15,

details and idiosyncratic features of how these protein factors contact RNA and what are the consequences of this interaction [2].

Routinely, after the alignment of the CLIP-Seq library of reads on the genome, a list with the entire collection of the mRNAs annotated on said genome is generated, which contains the number of reads mapped on every transcript, based on genomic coordinates. mRNAs with the large number of mapped reads represent highly bound mRNAs, which can be also interpreted as the mRNAs the protein factor we are studying spends more time binding (this can be quantified by other methods as dwell time). Therefore by analyzing these mRNAs we can extract meaningful insight on the mechanistic role of the protein factor. However, it is often observed that not all protein–mRNA interactions uncovered by CLIP-Seq are consequential (i.e., upon interaction there is a consequence for the mRNA). Such mRNAs are often very abundant, and the context of the interaction (passive contact, "scanning" of a potential but not true RNA target, "sponge" RNA) or nature of it (in vivo or artifact) is unclear.

A common strategy to reveal which interactions are consequential is to further perform CLIP-Seq after experimental treatment or after genetic manipulation, combined with other functional assays (e.g., ribosome profiling) [3]. This approach allows for the functional verification of the molecular and biological output of the protein–RNA interaction. Such experimental manipulations often lead to significant changes in RNA abundance, and therefore RNA-Seq should be performed for all conditions to provide a reference.

It becomes apparent from the above that knowing the steady state abundance levels of the RNAs engaged with the protein factor (and also the non-interacting ones) is essential for understanding how the protein factor is distributed on its RNA targets, and therefore through which RNAs its cellular function is delivered. It stands to reason that the most highly bound RNAs are the prime targets of the protein factor, and therefore we start the analysis of the CLIP-Seq library by ranking the bound RNAs by the number of mapped reads or a normalized measure of binding (more often mapped reads per million of mapped reads in the CLIP library aka RPM, less often RPM per kilobase of mRNA, aka RPKM). Parallel analysis of RNA-Seq libraries from the same biological samples informs on the abundance of the highly bound RNAs, and depending on the function of the protein factor, there is often some correlation between RNA-Seq abundance and CLIP-Seq binding. For example, it is expected that CLIP-Seq abundance will be strongly correlated to RNA-Seq abundance if the protein interrogated by CLIP contributes to RNA stability [4]; this is less observed in CLIP-Seq libraries of proteins with distinct tissue or subcellular localization patterns, where they may interact with a select fraction of the transcriptome [5].

Most importantly, the parallel analysis of CLIP-Seq and RNA-Seq can reveal whether there is enriched protein binding of certain RNAs relative to their amount in the RNA-Seq library. This is an important point, which needs to be further elaborated. It is reasonable to assume that a highly bound RNA is an important RNA target for the protein factor, but if its abundance in the RNA-Seq library is relatively higher than its abundance in the CLIP-Seq library this means that this RNA is not particularly preferred by the protein for binding. Considering the above, one could think that we should always normalize the CLIP-Seq abundance with the RNA-Seq abundance, to study the RNAs that are enriched for protein binding (normalization can be simply performed by dividing the CLIP-Seq abundance with RNA-Seq abundance). This indeed holds some value, but one should be very cautious for the following reasons: (1) a highly bound RNA is *by definition* a candidate important RNA target even if it is less abundant in the CLIP-Seq library than the RNA-Seq library, because even if it is not enriched, it is still bound by a large number of molecules of the protein factor we study, and the regulation of this RNA will likely have a considerable impact for the cell, because of its abundance; (2) RNAs of lower abundance tend to have higher abundance variability between samples (in both CLIP-Seq and RNA-Seq libraries), and as a result, dividing with a low value that has high variance can artificially bring such RNAs at the top of a ranked list. Moreover, abundance changes of such RNAs are less confidently determined. To avoid this problem, the abundance cut-off for considering an RNA present in a library can be set to a higher, more stringent threshold. Finally, normalizing for RNA-Seq abundance will "blind" a researcher with regard to the abundance of this RNA in the cell; the regulation of a low abundance RNA can still be important for the cell, but ignoring the dimension of abundance can lead to unfortunate biases. In our work we routinely study both the CLIP-Seq only mRNA lists and the RNA-Seq normalized ones, to avoid inherent biases of the two approaches and benefit from their advantages.

We reasoned that different library preparation protocols can introduce specific sequence biases in the cDNA library and that such biases can interfere in unknown ways with the parallel analysis of the two types of data. Having all the above in mind, this RNA-Seq method was developed following two early decisions: the protocol should unambiguously determine the DNA strand from which the RNA came (i.e., it should be directional), and if possible it should follow similar protocol steps with CLIP-Seq, using the same oligonucleotide adapters and primers for the library preparation. We developed this custom RNA-Seq protocol by using total RNA depleted from ribosomal RNA, as the poly-adenylation status of many non-coding RNAs is unknown. The outline of the protocol is illustrated in Fig. 1. Since CLIP-Seq requires by definition a

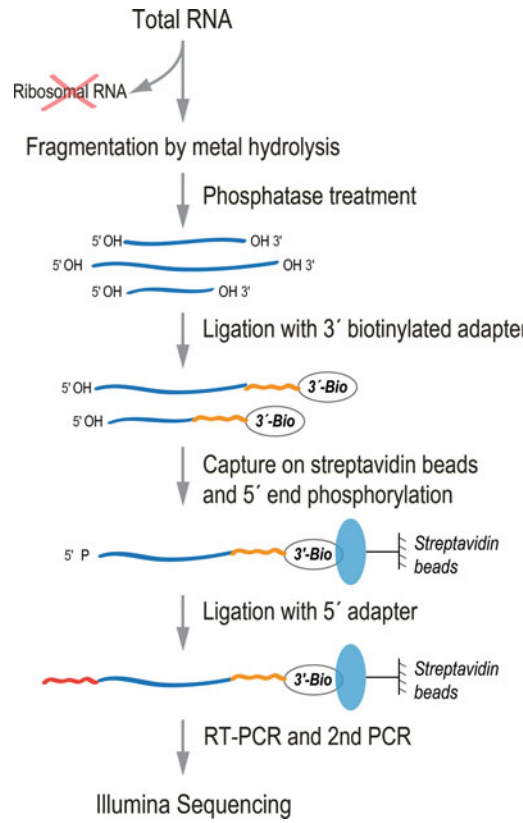


Fig. 1 SSD RNA-Seq schematic

short read sequencing approach, the total ribosome-depleted RNA is fragmented, by use of metal hydrolysis which shows minimum sequence bias at cleavage sites, if any. Similar to how the protein-bound RNA fragments in the CLIP protocol are immobilized on magnetic beads by way of their interaction and crosslinking on the protein, which is bound on the beads by an antibody, the RNA fragments in the RNA-Seq protocol are immobilized by ligating them first with a biotinylated adapter, and then binding them on streptavidin beads (hence, Solid-Support Directional RNA-Seq). Bead immobilization facilitates the subsequent enzymatic treatments of the RNA fragments in the RNA-Seq protocol up to the point of PCR amplification. The presence of the same adapter sequences at the 5′ and 3′ ends of the RNA fragments results in a cDNA library that bears the same trimmable sequences as the CLIP-Seq library, enabling the use of the CLIP-Seq pipeline (which includes CLIP-Seq tools [6]) for the analysis of both CLIP and RNA-Seq libraries, significantly simplifying their parallel analysis.

2 Materials and Equipment

2.1 Isolation of Total RNA from a Tissue or Cell Sample Using TRIzol

1. Freshly harvested tissue or trypsinized cells, or frozen biological samples.
2. TRIzol reagent (Invitrogen).
3. Chloroform.
4. Isopropanol.
5. 100% ethanol, ice cold.
6. 75% ethanol in water, ice cold.
7. RQ1 DNase (Promega).
8. rRNasin (Promega).
9. 25:24:1 Phenol/$CHCl_3$/Isoamyl alcohol.
10. 24:1 Chloroform:Isoamyl alcohol.
11. 3 M Sodium Acetate, pH 5.2.
12. 5 mg/mL Glycogen.
13. Nuclease-free water (Ambion).
14. Nanodrop UV spectrophotometer (or equivalent).

2.2 Ribosomal RNA Depletion Using RiboMinus Eukaryote Kit for RNA-Seq

1. RiboMinus™ Eukaryote Kit for RNA-Seq or RiboMinus™ Eukaryote System v2 (Invitrogen).
2. Magnetic stand.
3. Thermomixer capable of 1000 rpm.

2.3 Fragmentation of rRNA-Depleted Total RNA

1. RNA Fragmentation Reagents (Invitrogen).
2. G25 Sephadex microspin column (GE healthcare).
3. Antarctic Phosphatase (New England Biolabs, NEB).
4. 10× Antarctic Phosphatase buffer (NEB).

2.4 Biotinylated Adapter Ligation and Capture of Ligated RNAs Using M280 Streptavidin Dynabeads

1. 3′ RNA adapter:
   RL3-Bio: /5Phos/rGrUrGrUrCrArGrUrCrArCrUrUrCrCrArGrCrGrG/3BioTEG/ (*see* **Note 1**).
2. T4 RNA ligase (Thermo).
3. 50% PEG 8000 solution in water.
4. M280 Streptavidin Dynabeads (Invitrogen).
5. Solution A: 0.1 M NaOH, 0.05 M NaCl.
6. Solution B: 0.1 M NaCl.
7. B + W buffer: 10 mM Tris HCl pH 7.5, 1 mM EDTA, 2 M NaCl, 0.05% Tween-20.
8. PNW buffer: 50 mM Tris HCl pH 7.4, 5 mM $MgCl_2$, 0.1% Nonidet P-40.

2.5 On-Beads Ligation with the 5′ Adapter

1. T4 polynucleotide kinase (PNK) (NEB).
2. 10× PNK buffer (NEB; contained in T4 PNK kit).
3. 10 mM ATP (NEB; contained in T4 PNK kit).
4. 5′ RNA Adapter:
   RL5D: /5′InvddT/rArGrGrGrArGrGrArCrGrArUrGrCrGrGrNrNrNrNrNG (*see* **Note 2**).

2.6 Reverse Transcription and PCR Amplification of Adapter-Ligated RNA

1. Reverse transcription and first PCR primers:
   P3-*C*A*C primer: 5′-CCGCTGGAAGTGACTGA*C*A*C-3′ (*phosphothioate linkage).
   DP5 primer: 5′-AGGGAGGACGATGCGG-3′
2. Titan One Tube RT-PCR System (Sigma).
3. Deoxynucleotides (dNTPs, Invitrogen).
4. 100 mM Dithiotreitol (DTT, Thermo).
5. MetaPhor Agarose (Lonza).
6. Ethidium Bromide (Invitrogen).
7. Accuprime Pfx Supermix (Invitrogen).
8. GeneJET Gel Extraction Kit (Thermo).
9. Second PCR primers.
   The forward primers for the second PCR incorporate the barcode for multiplexing samples. These are based on the CLEAR-CLIP protocol developed by Robert Darnell laboratory [7] (*see* **Note 3**).
   Forward index primers:
   Index 1-**CTAG**
   5′-AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCTCTA-GAGGGAGGACGATGCGG-3′
   Index 2-**GATC**
   5′-AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCTGAT-CAGGGAGGACGATGCGG-3′
   Index 9-**TCAC**
   5′-AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCTTCA-CAGGGAGGACGATGCGG-3′
   Index 10-**AGTG**
   5′-AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCTAGT-GAGGGAGGACGATGCGG-3′

Index 11-**TACG**
5′-AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCTTAC-GAGGGAGGACGATGCGG-3′

Reverse primer: MSFP3:
5′- CAAGCAGAAGACGGCATACGAGATCCGCTGGAAG TGACTGACAC- 3′

Sequencing primer: Read 1, Universal TruSeq adapter:
5′- ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

3 Methods

Use nuclease-free consumables and reagents, and overall technique. Tubes containing reagents, beads, RNA, and reaction mixtures should be kept on ice in between incubations, unless otherwise noted.

3.1 Isolation of Total RNA from a Tissue or Cell Sample Using TRIzol

Harvest the cultured cells or tissue samples according to standard procedures. The pelleted cells or the tissue sample should be kept on ice and mixed as soon as possible after harvesting, with TRIzol (*see* **Note 4**). If you use frozen biological samples do not thaw them, just mix with the TRIzol reagent directly.

1. Mix 1 mL of TRIzol reagent per 50 mg of cell or tissue sample (the volume of the sample should not exceed 10% of the volume of TRIzol). Resuspend the cells by pipetting with a 1000 μL tip.
2. (Optional) Centrifuge at max speed in a tabletop microcentrifuge (*see* **Note 5**) for 10 min at 4 °C, and keep the clear supernatant (avoid top layer containing fat, if present).
3. Incubate at room temperature for 5 min, add 0.2 mL chloroform per 1 mL TRIzol, shake vigorously by hand for 30 s and leave at room temperature for 2 min.
4. Centrifuge at max speed for 15 min at 4 °C, keep upper (colorless) aqueous phase containing RNA.
5. Add 0.5 mL isopropanol, incubate at room temperature for 10 min, centrifuge at max speed for 10 min at 4 °C. The RNA precipitate forms a gel-like pellet on the side and bottom of the tube.
6. Discard the supernatant, wash pellet with 75% (v/v) ethanol, centrifuge at 7500 × *g* for 5 min at 4 °C.
7. Aspirate, air dry, resuspend in 64 μL of nuclease-free water.
8. To eliminate traces of DNA, add the following to the resuspended RNA:

(a) 8 μL 10× RQ1 DNase I buffer.

(b) 4 μL rRNasin.

(c) 4 μL RQ1 DNase I.

9. Incubate at 37 °C for 20 min.
10. Add 320 μL H_2O. Proceed with phenol extraction and ethanol precipitation as follows: add 400 μL Acid Phenol/Chloroform/Isoamyl alcohol 25:24:1, and vortex vigorously for 30 s. Centrifuge at max speed for 5 min. Take upper (water) phase and transfer to clean tube (*see* **Note 6**).
11. Add 400 μL chloroform/isoamyl alcohol 24:1, and vortex vigorously for 30 s. Centrifuge at max speed for 5 min. Take upper (water) phase and transfer to clean tube.
12. To the aqueous phase, add 40 μL (1/10 volume) 3 M sodium acetate, pH 5.2, 1 μL glycogen (5 μg/μL), 1 mL (2.5 volumes) ethanol 100%. Precipitate at −80 °C for 30 min and centrifuge at max speed and at 4 °C for 30 min.
13. Discard the supernatant, wash the RNA pellet with 75% ethanol, and centrifuge at max speed for 5 min. Discard the supernatant and air dry the RNA pellet, then resuspend it in 25 μL nuclease-free water (*see* **Note 7**).

3.2 Ribosomal RNA Depletion Using RiboMinus Eukaryote Kit for RNA-Seq

In this step we deplete ribosomal RNA from the total RNA preparation, using biotinylated oligonucleotides which hybridize with rRNA, and then remove the rRNA-oligo duplexes by binding on streptavidin beads (*see* **Note 8**).

1. Mix 3–10 μg of total RNA (*see* **Note 9**) with the RiboMinus probe in the hybridization buffer as follows:
   (a) 10 μL total RNA (3–10 μg).
   (b) 10 μL RiboMinus Probe (15 pmol/μL).
   (c) 100 μL Hybridization buffer.
2. Incubate at 70 °C for 5 min to denature the RNA and begin hybridization of the probe to the rRNA. Allow the sample to gradually cool down to 37 °C. This can be achieved by detaching the heating block from the apparatus, and leaving at room temperature to cool down. This should take approximately 25 min.
3. While waiting, proceed with magnetic beads preparation. Fully resuspend the beads suspension in the stock vial (*see* **Note 10**) and transfer 750 μL of beads suspension per sample into a clean tube.
4. Place the tube on a magnetic stand, and leave for 30–60 s for the beads to fully attach to the wall. The buffer should appear clear. Remove the supernatant and add 1 mL of nuclease-free

water. Fully resuspend the beads by first rotating the tube on the magnet and then taking it off the magnet and tilting it end over end until you don't see a bead pellet on the wall of the tube. Briefly centrifuge to remove any solution from the lid of the tube, and then place again on the magnet and remove the supernatant once the beads are attached to the wall (*see* **Note 10**). Repeat this step one more time.

5. After the wash, resuspend the beads in 750 μL of hybridization buffer. Transfer 250 μL to a new tube, and mark the tube as "step 2" and place it at 37 °C for use at a later step.
6. Place the tube containing the remaining 500 μL of beads suspension on the magnetic stand, remove the buffer, and resuspend beads using 200 μL Hybridization buffer. Mark this "step 1" tube. Keep at 37 °C.
7. Transfer the hybridized sample (120 μL) from **step 2** into tube marked "step 1" and fully mix.
8. Incubate at 37 °C for 15 min on a Thermomixer (1000 rpm) or on a standard heating block, mixing occasionally.
9. Place "step 1" and "step 2" tubes on magnet, remove supernatant from *"step 2" tube*, and transfer supernatant, which contains the rRNA-depleted RNA, from *"step 1" tube* (~320 μL) to "step 2" tube. Mix well and incubate at 37 °C as before. Discard tube 1.
10. Put "step 2" tube on the magnetic stand and collect the supernatant, which contains the rRNA-depleted RNA.
11. Precipitate RNA using ethanol. Add the following to the supernatant from the previous step:
    (a) 2 μL 5 μg/μL glycogen.
    (b) 32 μL 3 M sodium acetate, pH 5.2.
    (c) 800 μL 100% ethanol.
    (d) mix well and incubate at −80 °C for 30 min.
12. Centrifuge at max speed on a tabletop centrifuge at 4 °C for 30 min. Aspirate and wash pellet with 75% ice-cold ethanol. Centrifuge again for 5 min to pellet the RNA, aspirate, air dry and resuspend in 10 μL of nuclease-free water (*see* **Note 11**).
13. Use 1 μL of the rRNA-depleted RNA for measuring concentration. In the case of mouse testis, rRNA depletion of 10 μg total RNA yields approximately 10 μL of 170 ng/μL rRNA-depleted RNA (~1.7 μg total).

3.3 Fragmentation of rRNA-Depleted Total RNA and Dephosphorylation

The rRNA-depleted total RNA is fragmented by metal hydrolysis, and then dephosphorylated in preparation for the first ligation step.

1. Add 1 μL of RNA fragmentation buffer to the 9 μL rRNA-depleted RNA and incubate at 70 °C for 10 min. (*see* **Note 12**).
2. Stop by adding 1 μL of 10× STOP solution.
3. The RNA solution is desalted using G25 Sephadex microspin column. Resuspend the resin in the column by repeated quick manual inversion, until the resin is entirely resuspended. Remove the bottom plug and unscrew the lid. Pre-spin the column for 1 min at 735 × *g* (*see* **Note 13**), discard the bottom tube, and transfer the packed column into a clean tube.
4. To the solution from **step 2** add 40 μL of nuclease-free water and mix. Apply the mix onto the packed resin dropwise without disturbing its surface, and spin the column for 2 min at 735 × *g*. Collect the eluate; its volume is usually 55–60 μL.
5. To the eluate add:
   (a) 7 μL 10× buffer.
   (b) 3 μL Antarctic Phosphatase (total volume should be ~70 μL).
6. Incubate at 37 °C for 30 min. Add 330 μL of nuclease-free water and proceed with phenol extraction and ethanol precipitation as described in **steps 10** to **13**, Subheading 3.1.

3.4 Biotinylated Adapter Ligation and Capture of Ligated RNAs Using M280 Streptavidin Dynabeads

The fragmented RNA is ligated at its 3′ end with a biotinylated adapter.

1. Calculate the approximate molar concentration of fragmented rRNA-depleted RNA (average size of fragmented RNAs ~100 nts, MW_{100nts} = 31KDa) from last step. Ligate with the RL3-Bio adapter by following a 1:1 to 2:1 molar ratio of RNA: RL3-biotin adapter (*see* **Note 14**).
2. All RNA from previous step can be used for one ligation reaction. Mix ~60 pmoles of fragmented, phosphatase treated RNA with the following:
   (a) 8 μL 10× T4 RNA ligase buffer.
   (b) 40 μL of 50% solution of PEG 8000 in water.
   (c) 2 μL RNasin.
   (d) 2 μL T4 RNA ligase.
   (e) 0.8 μL of 50 pmoles/μL RL3-Bio.
   (f) Nuclease-free water up to 80 μL total.
3. Incubate overnight at 16 °C.
4. Pipette 100 μL of fully resuspended M280 Dynabeads into a clean tube (*see* **Note 15**). Wash the beads once with 1 mL of

Solution A, once with 1 mL of Solution B, and three times with 1 mL of buffer B + W. Remove the last wash buffer and resuspend beads in 200 μL of B + W buffer (twice the initial volume of the beads slurry).

5. To the ligation reaction mixture from **step 3**, add 120 μL of nuclease-free water, and transfer all 200 μL to the beads. Incubate for 15 min at room temperature with rotation (*see* **Note 16**).
6. Wash the beads twice with B + W buffer, and twice with PNW buffer. Keep the beads in last wash.

3.5 On-Beads Ligation with the 5′ Adapter

1. Prepare the polynucleotide kinase treatment mixture:
   (a) 8 μL 10× PNK buffer.
   (b) 1 μL 10 mM ATP.
   (c) 4 μL T4 PNK.
   (d) 2 μL rRNasin.
   (e) 65 μL nuclease-free water.
2. Remove last wash buffer from beads and mix with the prepared PNK treatment mixture. Incubate at 37 °C for 20 min, on a thermomixer at 1000 rpm (*see* **Note 17**).
3. Wash twice with B + W buffer, and twice with PNW buffer.
4. Remove last wash, and perform an on-beads ligation of the 5′ adapter RL5D by mixing the beads on which the ligated RNA is captured, with the following mixture:
   (a) 8 μL 10× T4 RNA ligase buffer.
   (b) 40 μL of 50% solution of PEG 8000.
   (c) 2 μL RNasin.
   (d) 2 μL T4 RNA ligase.
   (e) 1 μL of 50 pmoles/μL RL5D.
   (f) Nuclease-free water up to 80 μL.
5. Incubate at 16 °C for 6 h (or o/n) under shaking (1000 rpm).
6. Wash twice with B + W buffer.
7. Wash beads twice with nuclease-free water, and keep on ice until the next step.

3.6 Reverse Transcription and PCR Amplification of Adapter-Ligated RNA

For each RT-PCR use up to 1/2 of the beads (*see* **Note 18**)

1. After washing steps remove last wash and resuspend beads in 32.5 μL of H_2O.
2. To the beads suspension add:

(a) 1 μL of 10 mM DNTPs.

(b) 1 μL of 20 pmol/μL P3-*C*A*C primer (final concentration is 0.4 μM).

(c) 1 μL of 20 pmol/μL DP5 (final concentration is 0.4 μM).

3. Incubate at 70 °C on the thermomixer (1000 rpm) for 5 min, set the temperature to 37 °C and continue the incubating under shaking (1000 rpm) until the temperature reaches the set value. This should take approximately 20 min.
4. To the reaction mixture add:
   (a) 2.5 μL 100 mM DTT.
   (b) 1 μL rRNasin.
   (c) 10 μL 5× buffer Titan One Tube RT-PCR System.
   (d) 1 μL Titan One Tube enzyme mix.
5. While still on the thermomixer, proceed with the reverse transcription program: incubate at 50 °C for 45 min, and 55 °C for 15 min.
6. Transfer the entire mix (beads included) in a PCR tube and proceed with amplification:

PCR step 1: 94 °C for 2 min	
PCR step 2: 94 °C for 20 s PCR step 3: 58 °C for 30 s PCR step 4: 68 °C for 40 s	Repeat **steps 2** through **4**, 17 times
PCR step 5: 68 °C for 5 min	
PCR step 6: 4 °C for ever	

7. Collect reaction mixture, and discard the beads. Analyze reaction products on 3% MetaPhor gel, 1×TAE. 5′-3′ adapter only ligation products (with no insert) will give a band at 36 base pairs (Fig. 2). Anything above that size (~70–150 bp) is amplified ligated fragmented rRNA-depleted RNA.
8. Gel extract the ~70–150 bp products using GeneJET gel extraction kit. Weigh the gel fragment, and use the same w/v (mg/μL) volume of solubilizing solution. Incubate at 55 °C (preferably on a thermomixer, shaking at 1000 rpm) until the gel is dissolved. Add equal volume of isopropanol to enhance recovery of the small DNA molecules. Load onto the spin column and follow the standard procedure. Elute from the column with minimum elution volume (25 μL).
9. Use 2 μL of the RT-PCR gel extracted DNA and set up the 2nd PCR reaction as follows:

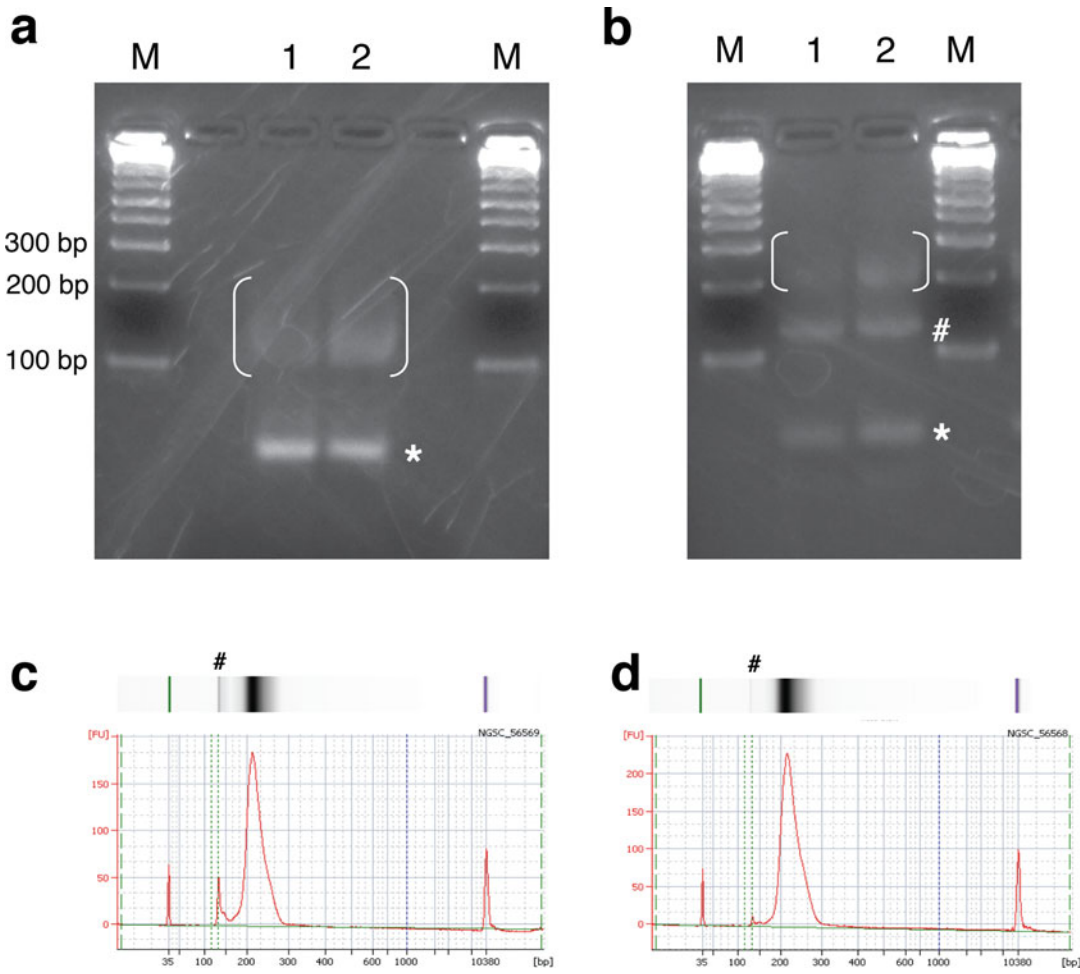


Fig. 2 Experimental images of RNA-Seq library preparation. (**a**) RT-PCR products analyzed on 3% MetaPhor gel. M: DNA markers. 1, 2: two replicate libraries. White brackets mark the ligated, fragmented and reverse transcribed rRNA-depleted RNA that needs to be extracted from the gel. A white asterisk marks the adapter dimer, which contains no insert (36 bp). (**b**) Second PCR products: analyzed on 3% MetaPhor gel. M: DNA markers. 1, 2: two replicate libraries. White brackets mark the RNA-Seq library DNA that needs to be extracted from the gel. The hash symbol marks the adaptor dimers (128 bp), and the asterisk marks the free primers. (**c**, **d**) Bioanalyzer results of the analysis of the two libraries extracted from the gel shown in (**b**). The hash marks the peak of the adapter dimers. The size of this peak in the two samples is within acceptable limits and will not substantially reduce the output of library sequencing

(a) 27 μL Accuprime Pfx Supermix (Invitrogen).
(b) 0.5 μL 20 pmol/μL forward index primer.
(c) 0.5 μL 20 pmol/μL MSFP3 primer.
(d) 2 μL of the RT reaction gel extract.

10. Run the following program on the thermocycler:

PCR step 1: 94 °C for 2 min	
PCR step 2: 94 °C for 20 s PCR step 3: 58 °C for 30 s PCR step 4: 68 °C for 40 s	Repeat **steps 2** through **4**, 11 times
PCR step 5: 68 °C for 5 min	
PCR step 6: 4 °C for ever	

11. Run the PCR products on a 3% MetaPhor gel, 1×TAE, and cut the gel piece that contains DNA molecules that are larger than 128 base pairs (~170–300 base pair, Fig. 2b, c). Extract the DNA using GeneJet as in the previous step. As a routine procedure before Illumina Sequencing, the cDNA library is analyzed on a Bioanalyzer chip to evaluate size range of the cDNA library (Fig. 2c), by Qubit fluorometry for accurate DNA concentration measuring, and by KAPA Library Quantification (Roche) to measure with higher accuracy the amplifiable part of the cDNA library. The libraries can be pooled (*see* **Note 3**), but because the oligo scheme is custom, the pool cannot be de-multiplexed automatically by the Illumina post-sequencing pipeline. The libraries are routinely subjected to Illumina single-end sequencing (**Note 19**).

3.7 Pre-processing of Sequencing Data and Mapping

After acquiring the raw sequencing data we perform de-multiplexing using *cutadapt* [8] with the standard parameters allowing one mismatch during multiplexing. Then, we use *cutadapt* to remove the 3′ adapter (GTGTCAGTCACTCT) and the 5′ adapter (AGGGAGGACGATGCGG) with the standard parameters for trimming with only one mismatch allowed. Then, we collapse the reads and remove the UMI (NNNNNG) by UMI-tools [9]. At the final step we perform the mapping using STAR [10] with the standard parameters with 10% mismatches allowed. For basic processing of the mapped reads, such as mRNA expression quantification we use CLIP-Seq tools [6].

4 Notes

1. Oligos are usually ordered at 100 nmole synthesis scale, and purified by RNase-free HPLC or PAGE (apart from the degenerate oligo RL5D). The biotin group at the 3′ end of RL3-Bio serves the dual purpose of preventing concatamerization and ligation at that end, in addition to enabling the capture of ligated fragments by the streptavidin magnetic beads.
2. The inverted ddT moiety serves the purpose of blocking 5′ from ligating with itself or an adapter. N stands for any nucleotide. This stretch of five random nucleotides serves as a unique

molecular identifier (UMI). If more than one read with both the same insert sequence and the same UMI is observed, it can be "collapsed" and counted as a single independent occurrence, as these reads are most likely PCR amplification artifacts.

3. Due to these index sequences being only four nucleotides, we recommend that the indexes you combine do not have more than one nucleotide in common, otherwise de-multiplexing can be problematic. Even for HiSeq libraries we rarely pool more than five libraries together. A combination in which none of the indexes shares more than one nucleotide with any of the other four indexes is 1, 2, 9, 10, 11.
4. You can store the biological sample after mixing with TRIzol, for long-term storage at −80 °C. You would resume total RNA extraction at this step.
5. The maximum speed of the tabletop microcentrifuges we commonly use is in the range of 15–17,000 × *g*. We recommend using a centrifuge capable of at least above 12,000 × *g* for this protocol.
6. When retrieving the upper (aqueous) phase, leave a small amount of it behind, to avoid contaminating the aqueous phase with organic solvent.
7. Nanodrop the extracted total RNA. Routinely, the concentration of the RNA in the 25 μL solution is more than 1 μg/μL (2–4 μg/μL) per 50 mg of tissue, and more than that for cultured cells. Note the $OD_{260/280}$ ratio and the form of the absorption curve; if either is suboptimal it is often because of residual phenol in the acquired water phase. You may re-extract the water phase by bringing the volume up to 400 μL with nuclease-free water and adding chloroform/isoamyl alcohol, and precipitating with ethanol as described herein. This additional step usually eliminates any contamination with organic solvent.
8. This protocol was first developed using RiboMinus™ Eukaryote Kit for RNA-Seq, which covers several eukaryotic species, including human, mouse, rat, drosophila, and others. Having one kit for multiple species is convenient for a lab that works with multiple species. This kit works very well, and the method description refers to that kit. This kit is still available as of this writing; however, it is anticipated to be discontinued. A newer kit, RiboMinus™ Eukaryote System v2, works on the same principle (biotinylated oligo capture of rRNA, and streptavidin beads depletion of the oligo captured rRNA) with some increased capability for ribosomal RNA depletion (most likely additional and/or better oligos). The protocol steps for the v2 kit will be essentially the same with minor changes.
9. It is important not to use more than 10 μg as this will result in incomplete removal of ribosomal RNA from your samples. As

low as 500 ng of total RNA can be used in this step, to retrieve enough ribosomal RNA-depleted total RNA to generate an RNA-Seq library, but success with such low amount is not guaranteed.

10. We prefer not to resuspend the beads by vortexing, instead by swirling the vial until there is no bead pellet at the bottom. During the following steps, do not aspirate the supernatant of the beads suspension with vacuum, even when the beads are immobilized by placing the tube on the magnet. Use only pipette (beads are in large amounts and are easily dislodged from the tube walls). Washes of magnetic beads in this and following sections are performed using 1000 μL of solution.
11. Residual magnetic beads may be present in the RNA pellet (especially if the pellet appears brownish). To get rid of them put the tubes with the resuspended RNA on a magnetic stand, let it sit for 1 min, and transfer the RNA solution into a new clean tube.
12. It is recommended that a time course of total RNA fragmentation is performed before treating your precious samples. We chose 10 min which is shorter than the recommended duration because we occasionally observed over-fragmentation of the RNA. Choose a time point where the RNA forms a smear around ~70–150 nucleotides.
13. 750 × *g* or 700 × *g* also works fine.
14. This molar ratio will ensure that there is no excess RL3-Bio oligo in the ligation reaction. The reason we want to secure this is that in the next step the ligation reaction is mixed with streptavidin beads without size fractionation (not necessary if the molar ratio is followed), and therefore both the ligated and the unligated adapter is present in the reaction and captured by the beads, and carried to the next step. If the 1:1 molar ratio is not followed, it is possible that an excess of the unligated biotin oligo will dominate the subsequent reactions leading to a failure of the cDNA library preparation.
15. M280 Streptavidin beads have a nominal binding capacity of 200 pmoles of biotin- single stranded oligos per 1 mg of beads. The beads suspension is 10 mg per mL. Use enough beads to capture all RL3 added in ligation reaction.
16. The purpose of this step is to capture RNA fragments ligated to the RL3-Bio adapter, and to remove unligated RNA fragments (which could cause ligation artifacts in subsequent steps) from the reaction mixture. Unligated RL3-bio adapter will also be captured on the beads, but since the ligation reaction of previous step had limiting amounts of adapter, it is expected that ligated RNAs will be in larger amounts than unligated adapter on the beads, and therefore can be efficiently removed by

agarose gel electrophoresis and extraction after amplification at subsequent steps.

17. This step restores a 5′ phosphate end to the RNAs ligated to the 3′ adapter, so that they can be ligated with the 5′ adapter.
18. The M280 manual states that more than 50 mg of beads in the reaction can inhibit RT-PCR. You may perform two reactions separately and combine. Usually just one half reaction produces enough cDNA to proceed to PCR.
19. Libraries prepared using this protocol are compatible with Illumina sequencing, but the technicians of the sequencing facility should be contacted before sample submission, and notified of the sequence structure of the DNA fragments in these libraries.

Acknowledgments

Anastasios Vourekas is grateful for the mentoring and guidance he received from Zissimos Mourelatos during his post-doc in the Mourelatos laboratory. Anastasios Vourekas is grateful to members of the Mourelatos laboratory, especially Panagiotis Alexiou and Manolis Maragkakis for discussions and brainstorming during the analysis of CLIP-Seq and RNA-Seq data.
We thank all the members of the Vourekas laboratory for discussions. Part of the high-performance computational needs of the Vourekas laboratory are provided by the Louisiana Optical Network Infrastructure (http://www.loni.org). Research in the Vourekas laboratory is currently supported by Department of Biological Sciences and College of Science start-up package.

Authors Contribution M.D. and A.E.M.S. have established in the Vourekas laboratory the bioinformatic pipeline with which CLIP-Seq and RNA-Seq data are analyzed. Z.M. and A.V. designed, and A.V. developed the method and wrote the paper.

References

1. Van den Berge K, Hembach KM, Soneson C et al (2019) RNA sequencing data: Hitchhiker's guide to expression analysis. Annu Rev Biomed Data Sci 2:139–173. https://doi.org/10.1146/annurev-biodatasci-072018-021255
2. Lee FCY, Ule J (2018) Advances in CLIP technologies for studies of protein-RNA interactions. Mol Cell 69:354–369. https://doi.org/10.1016/j.molcel.2018.01.005
3. Hafner M, Katsantoni M, Köster T et al (2021) CLIP and complementary methods. Nat Rev Methods Prim 1:20. https://doi.org/10.1038/s43586-021-00018-1
4. Kini HK, Silverman IM, Ji X et al (2015) Cytoplasmic poly(A) binding protein-1 binds to genomically encoded sequences within

mammalian mRNAs. RNA 22:61–74. https://doi.org/10.1261/rna.053447.115.4

5. Vourekas A, Alexiou P, Vrettos N et al (2016) Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm. Nature 531:390–394. https://doi.org/10.1038/nature17150
6. Maragkakis M, Alexiou P, Nakaya T, Mourelatos Z (2016) CLIPSeqTools—a novel bioinformatics CLIP-seq analysis suite. RNA 22:1–9. https://doi.org/10.1261/rna.052167.115
7. Moore MJ, Scheel TKH, Luna JM et al (2015) miRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. Nat Commun 6:8864. https://doi.org/10.1038/ncomms9864
8. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10. https://doi.org/10.14806/ej.17.1.200
9. Smith T, Heger A, Sudbery I (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 27:491–499. https://doi.org/10.1101/gr.209601.116
10. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635

Chapter 16

UPA-Seq-Based Search Method for Functional lncRNA Candidates

Saori Yokoi and Shinichi Nakagawa

Abstract

Long noncoding RNAs (lncRNAs) constitute a large fraction of the transcriptome in mammals, and recent studies have revealed important functions of lncRNAs in a variety of biological processes. However, the fraction of lncRNAs that have been functionally validated is small, and only sequence and expression information are available for most lncRNAs. Here, we describe the procedures for UV-phenol aqueous-phase RNA sequencing (UPA-seq), a method for searching for functional lncRNA candidates among whole genomes based on the assumption that functional lncRNAs exert their functions through associations with proteins.

Key words lncRNAs, UV crosslinking, RNA-Seq, Phenol-chloroform

1 Introduction

Long noncoding RNAs (lncRNAs) are defined as RNA transcripts longer than 200 nucleotides and constitute a large class of RNAs. Recently, it has become clear that lncRNAs are involved in a variety of biological processes, including the regulation of epigenetic gene expression, the formation of nonmembranous organelles, and the sequestration of associating molecules [1]. However, the functions of more than 98% of lncRNAs have not yet been characterized in humans [2]. Current models suggest that the transcription of certain regions of DNA by RNA polymerase II, but not the transcribed products, plays important functional roles. Thus, some lncRNAs are generated as transcriptional byproducts without a function. In addition, it is possible that stochastic interactions of RNA polymerase II with open regions of chromatin may result in the production of nonfunctional RNA transcripts or "transcriptional noise," which may constitute a significant portion of less abundant lncRNAs. It is therefore important to predict whether lncRNAs of interest are likely to be functional before starting

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_16,

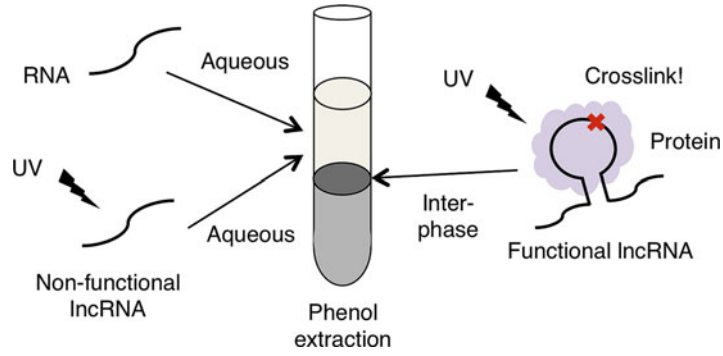


Fig. 1 Schematic illustration of the UPA-seq principle. Functional lncRNAs are crosslinked with proteins by UV irradiation and transferred to the interphase after phenol-chloroform extraction, while nonfunctional lncRNAs are transferred to the aqueous phase, even after UV irradiation

in-depth functional experiments. Although conserved amino acid sequences and motifs provide important clues for predicting protein functions, the lack of such clues in lncRNAs makes it difficult to discriminate functional lncRNAs from nonfunctional lncRNAs.

LncRNAs so far reported to be functional most often mediate their function as a component of ribonucleoprotein complexes. Herein, we describe the procedures for UV-phenol aqueous-phase RNA sequencing (UPA-seq), which is an efficient method to search for likely functional lncRNAs [3]. This method takes advantage of the fact that UV irradiation induces covalent bond formation between RNAs and proteins. In conventional acid guanidinium phenol-chloroform extraction, RNAs are extracted from the aqueous phase, and nonfunctional lncRNAs tend to be transferred to the aqueous phase even after UV irradiation. On the other hand, functional lncRNAs, which are assumed to associate with proteins, are crosslinked with proteins by UV irradiation and transferred to the interphase. Thus, the RNAs whose content in the phenol-chloroform extract is reduced by UV irradiation are more likely to be functional lncRNAs (Fig. 1). Furthermore, in UPA-seq analysis, as all extracted RNAs are sequenced, it is possible to comprehensively search for lncRNAs in the whole genome.

Most of the previous functional analyses of lncRNAs have been performed using cultured cells. However, the phenotypes of some lncRNAs predicted based on cellular studies have not been observed in mutant animals [4]. Therefore, there is increasing consensus that it is important to analyze the molecular interactions of lncRNAs and physiological functions not only at the cellular level but also at the animal/organism level. UPA-seq methods can be applied not only to cultured cells but also to non-cultured tissues and cells from animals. In this chapter, we describe the protocols for using both cultured cells and tissues, which will help characterize the functions of lncRNAs at the cellular and organism levels.

2 Materials

2.1 Equipment for Sample Preparation

2.1.1 Equipment for Sample Preparation (Common)

1. UV cross-linker (We used a Funa UV linker (#FS-800, Funakoshi, Japan). Any UV cross-linker is suitable as long as it allows irradiation at 254 nm UV with 120 mJ/cm^2 total energy.).
2. Tray.
3. Aluminum foil.
4. Heat block, preheated to 55 °C.
5. Tube rotator.
6. Refrigerated microcentrifuge.
7. Spectrophotometer.

2.1.2 Equipment for Sample Preparation (Cultured Cell Sample)

1. 100 mm tissue culture dish.

2.1.3 Equipment for Sample Preparation (Tissue Sample)

1. Fine forceps.
2. Micro-ophthalmic scissors.
3. 35 mm dish.
4. Cell strainer (40 μm pore size).
5. 1 mL syringe plunger.
6. Silicon sealant (A white silicon sealant is easier to see than a transparent one.).
7. Cotton-tipped swab/applicator.

2.2 Stock Solutions

Store these solutions at room temperature. Use ultrapure deionized water (DW) for all of the stock solutions. The use of diethylpyrocarbonate (DEPC)-treated water is not recommended for safety reasons.

2.2.1 Stock Solutions (Common)

1. 2 M Tris-HCl (pH 8.0): 121.1 g Tris, ~40 mL concentrated HCl; adjust the pH to 8.0 and the total volume to 500 mL with DW and then autoclave.
2. 0.5 M EDTA (pH 8.0): 93.1 g EDTA-2Na, ~10 g NaOH pellets; adjust the pH to 8.0 and the total volume to 500 mL with DW and then autoclave.

2.2.2 Stock Solution (Cultured Cell Sample)

1. 10× HCMF: 80 g NaCl, 4 g KCl, 1.2 g $Na_2HPO_4(12H_2O)$, 24 g HEPES, 10 g glucose, 1.92 g NaOH; adjust the pH to 7.4 and the total volume to 1 L with DW and then autoclave.

2.3 Working Solutions

2.3.1 Working Solution (Common)

1. TE: Add 0.5 mL of 2 M Tris-HCl (pH 8.0) and 0.2 mL of 0.5 M EDTA (pH 8.0) to 90 mL of sterilized DW. Adjust the total volume to 100 mL with DW and then autoclave.

2.3.2 Working Solutions (Cultured Cell Sample)

1. HBSS: Add 50 mL of 10× HCMF to 400 mL of sterilized DW, adjust the volume to 500 mL, and autoclave. Add 0.5 mL of 1 M $CaCl_2$ and 0.5 mL of 1 M $MgCl_2$. Do not autoclave.

2.4 Other Reagents

1. TRIzol (for cultured cell samples) or TRIzol-LS (for tissue samples) (Invitrogen).
2. Ribo-Zero Gold rRNA removal Kit (H/M/R) (Illumina).
3. TruSeq Standard mRNA Library Prep kit (Illumina).
4. TruSeq RNA Single Indexes SetA and SetB (Illumina).

3 Methods

Wear gloves for all of the steps. Perform all procedures at room temperature unless otherwise indicated.

3.1 Preparation of Samples

3.1.1 Preparation of Samples from Cultured Cells and UV Irradiation

1. Grow cells to 80–90% confluence in 100 mm cell culture dishes.
2. Replace culture medium with 1 mL of ice-cold HBSS.
3. Float the dish on a tray filled with ice water. Remove the lid of the dish and irradiate cells with UV (120 mJ/cm^2 at 254 nm) in a UV cross-linker (*see* **Note 1**). To prepare nonirradiated samples, cover the dish with aluminum foil to avoid UV irradiation of the cells.
4. Remove HBSS and add 1 mL of TRIzol Reagent (Invitrogen) (*see* **Note 2**). Lyse and homogenize the sample by pipetting up and down. Collect the sample in a 1.5 mL tube (*see* **Note 3**). Samples can be used for the following steps immediately or stored at −80 °C. The sample is stable for at least 1 month at −80 °C.

3.1.2 Preparation of Samples from Tissues and UV Irradiation

The irradiation of dissected organs or tissues with UV light may result in insufficient crosslinking of functional lncRNAs with proteins due to a lack of UV light penetration deep inside organs. Thus, a cell suspension is prepared by forcing tissues through a cell strainer in a saline solution (*see* **Note 4**).

1. Dissect organs or tissues using forceps and micro-ophthalmic scissors.

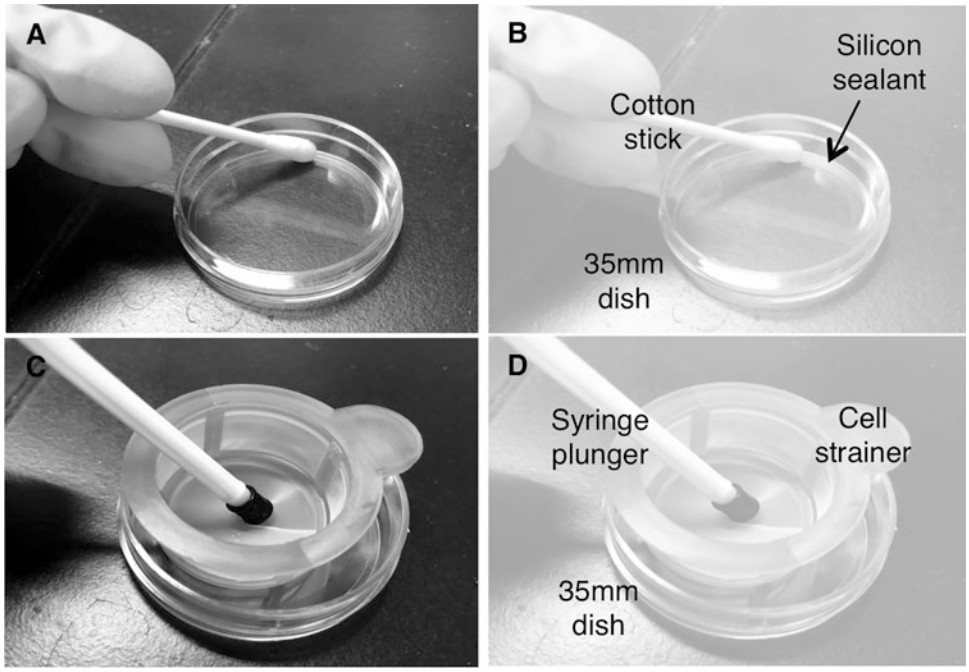


Fig. 2 Pretreatment procedure for organ tissue cells before UV irradiation. (**a**, **b**) Silicon sealant is applied to the head of a cotton stick and spread along the inside edge of the bottom of the 3.5 cm dish. (**c**, **d**) The dissected organ is ground with a syringe plunger in a cell strainer, and the filtered cells are suspended in saline

2. Spread silicon sealant along the inside edge of the bottom of a 35 mm dish using a cotton-tipped applicator to prevent cells from accumulating along the sidewalls of the dish where they will be shielded from UV light (Fig. 2a). Place the dish on ice.
3. Place a cell strainer on a 35 mm siliconized dish (Fig. 2b). Grind the dissected organ with a 1 mL syringe plunger in the cell strainer and suspend the filtered cells with 250 μL of ice-cold saline (*see* **Note 5**). Add 1 μL (40 U) of RNase inhibitor to the suspension (*see* **Note 6**). Spread the suspension thinly, avoiding contact with the silicon sealant (Fig. 3, *see* **Note 7**).
4. Float the dish on a tray filled with ice water. Remove the lid of the dish and irradiate cells with UV (120 mJ/cm^2 at 254 nm) in a UV cross-linker (*see* **Note 1**). To prepare the nonirradiated samples, cover the dish with aluminum foil to avoid UV irradiation of the cells.
5. Add 750 μL of TRIzol-LS Reagent (Invitrogen) (*see* **Note 2**). Lyse and homogenize the sample by pipetting up and down. Collect the sample in a 1.5 mL tube (*see* **Note 3**). Samples can be used for the following steps immediately or stored at −80 °C. The sample is stable for 1 month at −80 °C.

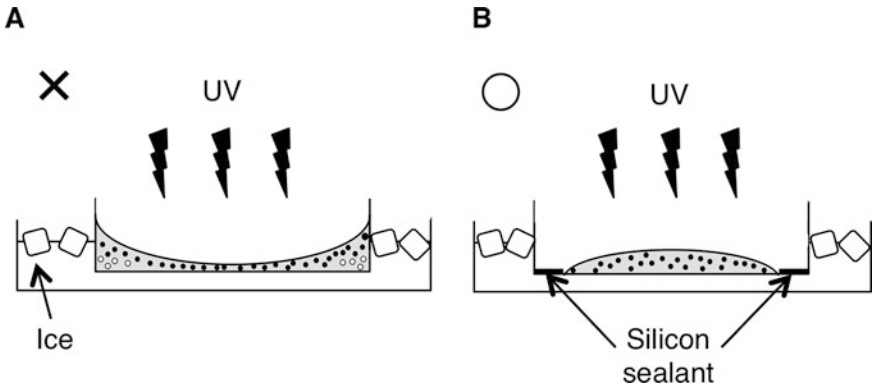


Fig. 3 Bad and good examples of how to spread sample suspension. Side views of a cell suspension in a 35 mm dish floated on a tray filled with ice water. Black and white dots indicate UV-irradiated cells and non-UV-irradiated cells, respectively. (**a**) A bad example. Without the use of a silicon sealant, the cell suspension tends to climb the wall of the dishes due to surface tension, preventing UV light from reaching the cells at the bottom near the sidewalls. (**b**) A good example. Silicon sealant repels the cell suspension, and the suspension is spread evenly, resulting in all cells in the suspension being irradiated by UV light

3.2 RNA Extraction

1. If samples are frozen, thaw them at room temperature. After thawing, incubate the samples at 55 °C for 10 min (*see* **Note 8**).
2. Allow the samples to cool to room temperature.
3. Add 200 μL of chloroform (*see* **Note 9**) and shake by hand vigorously for 30 s.
4. Rotate the tube using a tube rotator for 5 min to mix the sample well.
5. Centrifuge at 15,000 × *g* for 5 min at 4 °C.
6. Collect as much of the upper aqueous phase as possible and transfer it to a fresh tube (*see* **Note 10**).
7. Centrifuge at 15,000 × *g* for 5 min at 4 °C.
8. Carefully collect the upper aqueous phase without disrupting the intermediate and lower phases, transferring it to a fresh tube (*see* **Note 11**).
9. Add 500 μL of isopropanol and mix well (*see* **Note 12**).
10. Incubate 4 °C for 30 min.
11. Centrifuge at 15,000 × *g* for 10 min at 4 °C.
12. Carefully remove the supernatant without disturbing the pellet (*see* **Note 13**).
13. Add 1 mL of 80% ethanol and vortex briefly.
14. Centrifuge at 15,000 × *g* for 1 min at 4 °C.

15. Carefully remove the supernatant without disturbing the pellet.
16. Air dry the pellet and dissolve it in TE (*see* **Note 14**).
17. Determine the concentration of the RNA sample with a spectrophotometer (*see* **Note 15**).

3.3 RNA Sequencing (RNA-seq)

Generate RNA-seq libraries for high-throughput RNA-seq using a Ribo-Zero Gold rRNA removal Kit (H/M/R) (Illumina), TruSeq Standard mRNA Library Prep kit (Illumina), and TruSeq RNA Single Indexes SetA and SetB (Illumina) according to the manufacturer's protocols. Sequence the libraries on the Illumina HiSeq 4000 platform.

3.4 Analysis of RNA Sequence Data

Perform adaptor trimming and quality filtering of sequence data using Trimmomatic [5]. Map sequence reads to the genome using HISAT2 [6]. Count the numbers of sequence reads uniquely mapped to each exon using featureCounts [7] with the –O option (allowMultiOverlap) (*see* **Notes 16** and **17**).

Analyze the count data using DESeq2 [8] with the Wald test, and determine the genes whose normalized count levels are significantly decreased by UV treatment (FDR < 0.05) (Fig. 4). These genes represent the lncRNAs depleted from the aqueous phase (i.e., those enriched in the discarded interphase due to their association with proteins). These lncRNAs are presumably components of ribonucleoprotein complexes, a useful filter for likely functional lncRNAs.

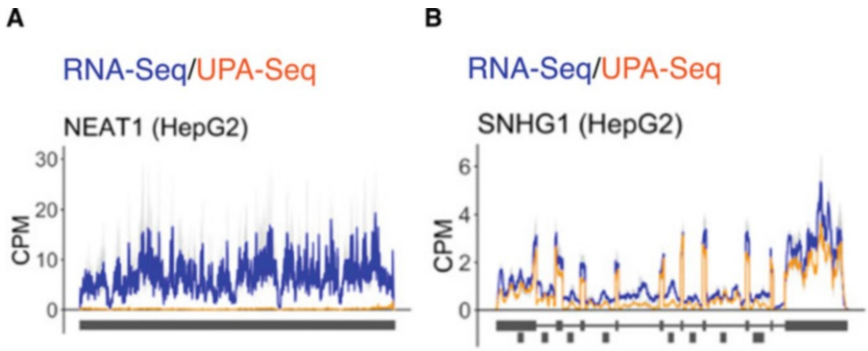


Fig. 4 Examples of UPA-seq results using HepG2 cells. Read distribution of UPA-seq and RNA-seq reads of genes. CPM represents counts per millions of reads. Blue and orange lines indicate the distribution of reads obtained via UPA-seq and RNA-seq, respectively. Data are represented as the means ± SD (gray lines) of three biological replicates. (**a**) The height of the sequencing signal of NEAT1, whose transcript is thought to be a functional lncRNA, is dramatically reduced by UV irradiation. (**b**) The height of the sequencing signal of SNHG1, whose transcript is thought to be a nonfunctional lncRNA, is not changed by UV irradiation. Figure adapted from [3]

4 Notes

1. UV irradiation generates heat, which may alter the structure of ribonucleoprotein complexes. Floating the sample dish on a tray filled with ice water will keep the sample temperature low and prevent such alterations.
2. TRIzol Reagent is recommended for use with solid samples (cultured cell samples) and TRIzol-LS Reagent is recommended for use with liquid samples (tissue suspension samples).
3. The viscosity of genomic DNA may prevent pipetting up and down. In this case, cut the pipet tip to widen the tip diameter. Mix the sample by pipetting until it is no longer viscous, as otherwise the RNA yield will be reduced if mixing is insufficient.
4. Do not sonicate the tissue. Sonication disrupts cells and sensitive ribonucleoprotein complexes may be broken up before UV irradiation.
5. A cell suspension concentration that is too high may prevent UV light from reaching all the cells. Before generating an RNA-seq library, qRT-PCR-based confirmation that the extracted amount of known functional lncRNAs is reduced by UV irradiation is recommended. If their content is not reduced, increase the amount of saline added to the cell suspension to dilute it.
6. The use of RNAlater to inhibit RNA degradation is not recommended because RNAlater contains highly concentrated ammonium sulfate, which may alter the compositions of ribonucleoprotein complexes.
7. If a silicon sealant is not used or the dish is tilted too much even when a silicon sealant is used, the cell suspension will adhere to the dish wall, preventing UV light from reaching the cells at the bottom near the sidewalls (Fig. 3). To irradiate all the cells in the cell suspension with UV light, spread the cell suspension evenly inside the area of the silicon sealant.
8. Without heat or needle shearing treatment, many architectural lncRNAs are difficult to extract via acid guanidinium phenol-chloroform extraction because they tend to be trapped in the protein phase [9]. As more than 100 rounds of needle shearing are required to improve lncRNA extraction, heat treatment is recommended.
9. Add chloroform in a flow cabinet for safety.
10. After this step, there is one more chance to collect the aqueous phase (**step 8**, Subheading 3.2). Therefore, it does not matter

if a small amount of the interphase or organic phase is collected in this step.

11. Interphase and organic phase contamination cannot be removed in the following procedures. Thus, care should be taken to collect only the colorless and transparent aqueous phase and not to touch other phases with the pipet tip.
12. For small samples (1–10 mg tissue or 10^2 to 10^4 cells), add 5 μg of glycogen along with isopropanol. Glycogen acts as a nucleic acid carrier and increases the RNA yield. Up to 4 mg/mL glycogen does not inhibit the following reactions.
13. Completely remove the supernatant from the tube. If the supernatant is insufficiently removed, guanidine thiocyanate contamination remains in the purified RNA solution and inhibits enzymatic reactions involved in generating the RNA-seq library. Guanidine thiocyanate contamination can be checked by calculating the A260/A280 ratios of purified RNA samples measured with a spectrophotometer (*see* **Note 15**).
14. Dissolve the RNA pellet in TE before it is completely dried because the solubility of the completely dried RNA pellet is very low. If the RNA pellet is inadvertently completely dried, 10 min of incubation at 55 °C after adding TE will make it soluble again.
15. Check the A260/A280 and A260/A230 ratios of the purified RNA samples determined with a spectrophotometer. If the A260/A280 ratio is less than 1.6, RNA may not be dissolved in TE. If the A260/A230 ratio is less than 2.0, contamination with guanidine thiocyanate or other salts may remain in the RNA samples. Further purification is recommended using a kit such as the Direct-zol RNA MicroPrep kit (Zymo Research) if the A260/A280 ratio is less than 1.7 and/or the A260/A230 ratio is less than 1.8. Alternatively, repeated ethanol precipitation can also remove the salt contamination.
16. Transcript annotations for read counts can be obtained not only from the existing gtf file downloaded from a reference database (e.g., GENCODE, Ensembl) but also from the gtf file generated from read alignment data output by HISAT2. To generate the gtf file, merge the alignment data, and output the transcript annotations with StringTie [10]. At this point, it is possible to use the existing gtf file as the reference annotation to guide the assembly process (if the existing gtf file is available), producing annotations that include reference transcripts and novel assembled transcripts as the output.
17. Without the -O option, a read that overlaps more than one feature will not be counted, and the expression levels of alternatively spliced genes that share common exons will be estimated to be lower than they actually are.

References

1. Marchese FP, Raimondi I, Huarte M (2017) The multidimensional mechanisms of long noncoding RNA function. Genome Biol 18: 206. https://doi.org/10.1186/s13059-017-1348-2
2. de Hoon M, Shin JW, Carninci P (2015) Paradigm shifts in genomics through the FANTOM projects. Mamm Genome 26:391–402. https://doi.org/10.1007/s00335-015-9593-8
3. Komatsu T, Yokoi S, Fujii K et al (2018) UPA-seq: prediction of functional lncRNAs using differential sensitivity to UV crosslinking. RNA 24:1785–1802. https://doi.org/10.1261/rna.067611.118
4. Nakagawa S (2016) Lessons from reverse-genetic studies of lncRNAs. Biochim Biophys Acta 1859:177–183. https://doi.org/10.1016/j.bbagrm.2015.06.011
5. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
6. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10.1038/nmeth.3317
7. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656
8. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8
9. Chujo T, Yamazaki T, Kawaguchi T et al (2017) Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncoding RNA s. EMBO J 36:1447–1462. https://doi.org/10.15252/embj.201695848
10. Pertea M, Pertea GM, Antonescu CM et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295. https://doi.org/10.1038/nbt.3122

Chapter 17

Large-Scale Analysis of RNA–Protein Interactions for Functional RNA Motif Discovery Using FOREST

Emi Miyashita, Kaoru R. Komatsu, and Hirohide Saito

Abstract

RNA transcripts can form a variety of higher-order structures. We developed a large-scale affinity analysis system, FOREST (Folded RNA Element Profiling with Structure Library), to investigate the function of these RNA structures on transcriptome-wide scale. Here we describe a protocol to analyze RNA–protein interactions using FOREST. Users of the protocol prepare an RNA structure library comprised of diverse species of transcripts and perform high-throughput characterization of the RNA–protein interactions to obtain quantitative and comprehensive information on the binding affinities and specificities. Moreover, we demonstrate how FOREST can be used to analyze a non-canonical structure, the RNA G-quadruplex, without sequencing bias, because the quantification is performed directly on a microarray without sequence amplification. FOREST will contribute to the discovery of RNA structure motifs that determine RNA–-protein interactions.

Key words RNA structure, RNA motif, RNA-binding proteins, Microarray, Large-scale affinity analysis, RNA–protein interactions

1 Introduction

RNA can form various higher-order structures and play a role in gene expression and cell-fate regulation.

The discovery of functional RNA structures is important for understanding their regulatory mechanisms. Aptamers, riboswitches, and ribozymes are typical examples of functional RNA structures. They have been discovered by homology-based searches and structure probing with prediction software [1–4]. These strategies provide a means of investigating the regulatory networks governed by functional RNA structures. However, the roles of these RNA structures remain mostly unknown.

Combining RNA structure dataset analysis with massively parallel sequencing data can elucidate the biological phenomena influenced by RNA structures [5, 6]. However, transcriptome-wide analysis techniques, such as CLIP-seq [7], reveal only the footprint

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_17,

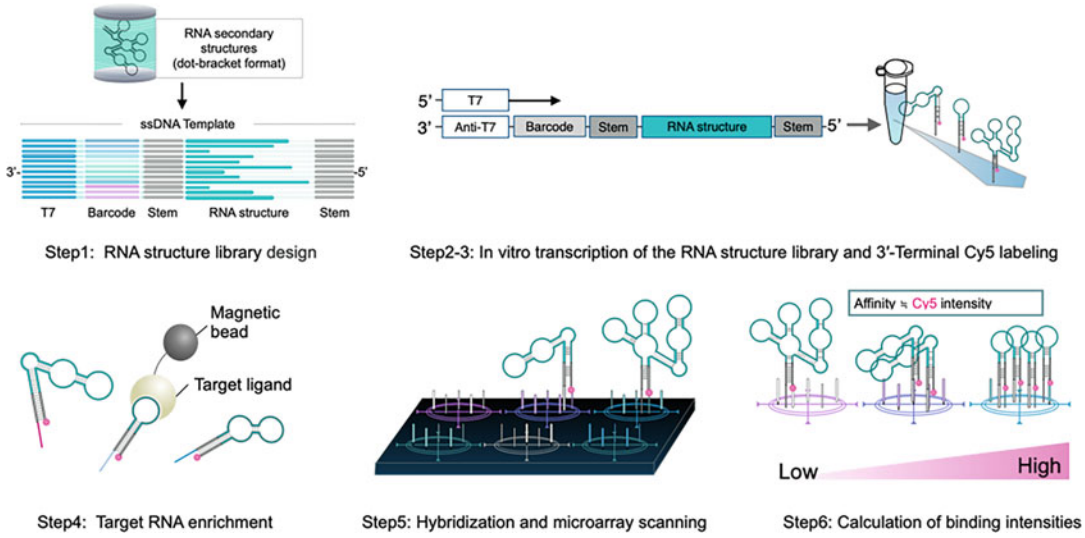


Fig. 1 Summary of steps for the analysis of RNP interactions using FOREST. FOREST consists of the following three major steps. (1) [Step 1] In silico extraction and design of the RNA structure library. Each RNA has a barcode for identification and a common stem sequence for structural stabilization. (2) [Steps 2–4] Transcription of the RNA library and enrichment of the target RNA by a biochemical assay (e.g., pull-down assay). (3) [Steps 5 and 6] Fluorescence quantification of the target RNAs by barcode microarrays. Binding intensities can be quantified from the fluorescence intensity of Cy5 added to each RNA

of the target RNA sequence, and additional validation is required to obtain information of both RNA structures and sequences that are essential for the target binding. In addition, many RNA structures may be missed in the analysis due to variable efficiency of reverse transcription of these structures during the library preparation, which impairs quantitative evaluation. Furthermore, current in vitro experiments mostly use RNA libraries containing random sequences [8], which cannot reproduce RNA structures encoded by natural genomes.

To overcome these problems, we have developed FOREST (Folded RNA Element Profiling with Structure Library) for efficient discovery of functional RNA structural motifs through ligand interactions (Fig. 1) [9]. FOREST provides quantitative and efficient identification of functional RNA structures using an RNA structure library and is composed of three integrated technologies.

The first technology is an RNA structural unit extraction program (Fig. 2). This program recognizes RNA structures from a database that includes specified RNA structural units and constructs a large list of RNA structures for the assay. For this protocol, we extract RNA structural units whose 5′-3′ ends form a double-stranded stem from highly conserved regions or whose double-stranded regions are known to form in cells.

Second, based on the list of RNA structures, a large amount of RNA is transcribed in vitro to synthesize the RNA structure library.

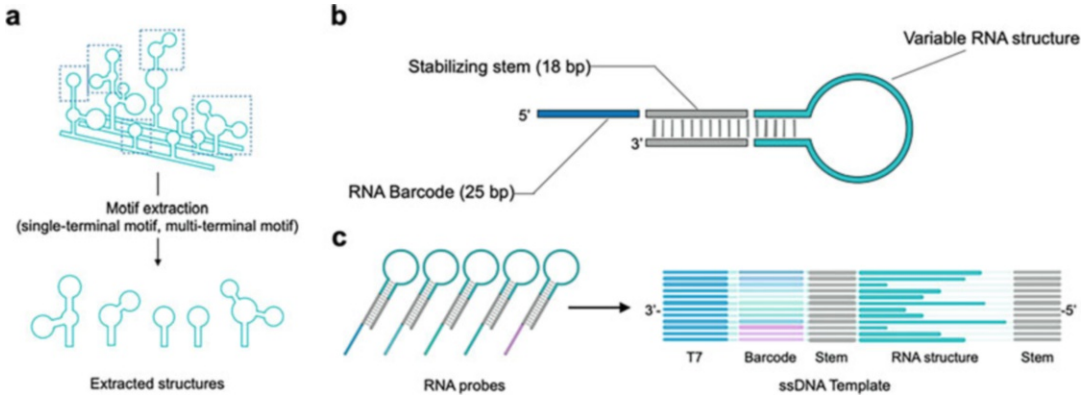


Fig. 2 RNA structure library design in silico. (**a**) Various terminal RNA motifs are extracted from RNA secondary structures in dot-bracket format. (**b**) All RNA probes have stabilizing stem sequences and an RNA barcode. Multiple barcodes are assigned to each RNA structure. (**c**) The sequences of the designed RNA probes are converted to ssDNA templates for in vitro transcription

This approach allows for the inclusion of up to a million different RNA structures in the library in a single tube. The library can be further expanded according to the purpose of the experiment (e.g., viral, bacterial, or mammalian RNA research).

The final technology is the RNA-DNA barcode microarray, which is specialized for large-scale detection of the RNA structure library. In general, the sequencing of RNA structures requires reverse transcription and PCR, which may introduce amplification bias depending on the RNA structure. In order to avoid this bias, quantitative detection by RNA-DNA hybridization using barcode sequences is adopted for direct and large-scale quantification of the target-acting RNA structures. This approach enables FOREST to quantify highly structured RNAs.

By integrating these elemental technologies with arbitrary assays, the ligand-binding tendency of RNA structures can be investigated on a large scale. This chapter provides a detailed protocol that applies FOREST to RNA–protein (RNP) interactions by the RNA pull-down enrichment of binding groups. This protocol is based on our previous analysis of two RNA-binding proteins (RBPs), U1A [8, 10] and LIN28A [11, 12]. Using FOREST, we quantified the RNP affinity and specificity and evaluated the preferred structures of U1A or L IN28A-binding RNAs. We also succeeded in quantifying the RNA guanine quadruplex (rG4) structure, which is a non-canonical RNA structure, and clarified the binding specificity and affinity of rG4-binding proteins (Fig. 3) [13, 14]. In addition, by analyzing the obtained binding data, we identified rG4 structures from human microRNA precursors. In summary, FOREST can analyze arbitrary RNA structure motifs with biochemical assays for RNA enrichment, leading to a comprehensive understanding of RNP interactions on a large scale.

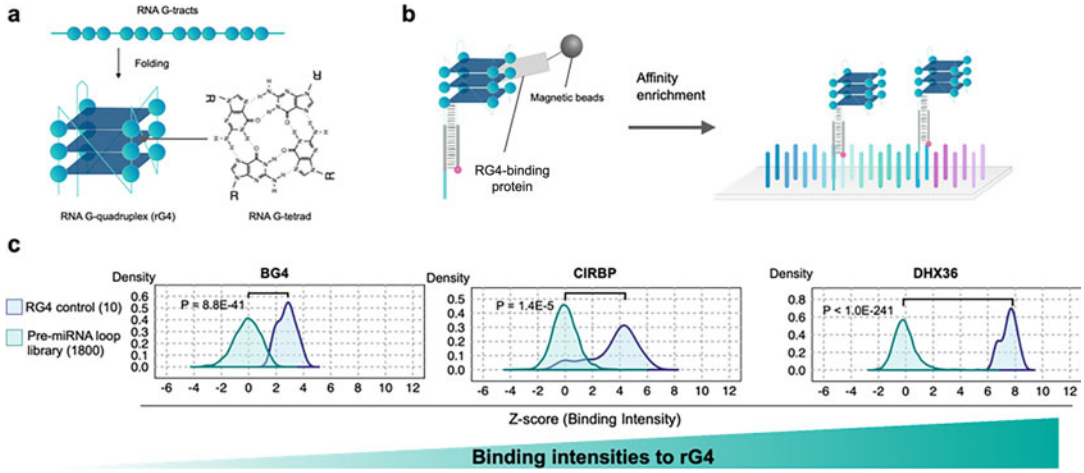


Fig. 3 (**a**) Schematic of the rG4 structures. (**b**) Schematic of the FOREST-based characterization of rG4-binding proteins. (**c**) Histogram of binding intensities comparing rG4 controls (10 different sequences) and human pre-miRNAs (1800 different sequences); *p* values were determined by the two-tailed Brunner–Munzel test ($n = 2$). From this data, we interpret that the rG4 binding intensities of the tested proteins are DHX36 > CIRBP > BG4. The data are taken from [9]

2 Materials

2.1 RNA Structure Library Design

1. Published codes on Github (https://github.com/KRK13/FOREST2020/).

2.2 In Vitro Transcription of RNA Probes and the RNA Structure Libraries

1. MEGAshortscript T7 Transcription Kit (Thermo Fisher Scientific).
2. TURBO DNase (Thermo Fisher Scientific).
3. Zymo RNA Clean and Concentrator (Zymo Research).

2.3 3′-Terminal Cy5 Labeling

1. 1× T4 Ligase Buffer (Thermo Fisher Scientific).
2. 100 μM pCp-Cy5 (Jena Bioscience).
3. 0.5 U/μL T4 RNA Ligase (Thermo Fisher Scientific).
4. Zymo RNA Clean and Concentrator (Zymo Research).

2.4 Target RNA Enrichment

1. 1 M HEPES KOH, pH 7.5.
2. 1 M KCl.
3. 5 M NaCl.
4. Glycerol.
5. UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific).
6. 20 mg/mL Bovine Serum Albumin (BSA).
7. 500 mM DTT.

8. 25:24:1 (v/v/v) Phenol:Chloroform:Isoamyl Alcohol, pH 7.9 (Nacalai tesque).
9. Chloroform.
10. Ethanol.
11. Ethathinmate (Nippongene).
12. Magnetic beads.
    RNA pull-down with His-tagged proteins.
    (a) TALON Magnetic beads (Takara).
    RNA immunoprecipitation with Flag-tagged proteins.
    (a) The Dynabeads Protein Immunoprecipitation Kit (Thermo Fisher Scientific).
    (b) M2 Anti-FLAG antibody (Sigma).
13. 10% SDS.
14. UltraPure Tris-HCl pH 7.5.
15. 0.5 M EDTA, pH 8.0.

2.5 Hybridization and Microarray Scanning

1. Custom CGH DNA microarray (Agilent Technologies).
2. Gene Expression Hybridization Kit (Agilent Technologies).
3. Gene Expression Wash Buffer 1 (Agilent Technologies).
4. Gene Expression Wash Buffer 2 (Agilent Technologies).

3 Methods

The outline of the procedures is shown in Fig. 1. Use RNase-free tubes in all procedures.

3.1 RNA Structure Library Design

RNA probes in the RNA structure library are generated from single-stranded (ss)DNA templates by in vitro transcription (explained later). The templates include five components in the following order from the 3′ end.

1. CC + T7 promoter sequence: A sequence for in vitro transcription using template DNA. The sequences of the following components are transcribed using T7 polymerase. The first two bases are included to increase the Tm (melting temperature) value. More bases can be attached to the 5′ end to increase the RNA yield.
2. RNA barcode sequence (25 nt): This portion of molecule immobilizes each RNA probe on the designated microarray spots, hybridizing the barcode sequences orthogonally with the complementary strands of the barcode sequences placed on the DNA barcode microarray. Assign multiple barcode sequences to each RNA structure in the RNA structure library.

3. Stabilizing stem structure sequence forward (18 nt): A unique sequence that forms a stem structure to stabilize the RNA structure and insulate the RNA structure and RNA barcode sequence.
4. RNA structure: Variable RNA structures extracted from RNA structure datasets. The 5′ and 3′ portions should form a double-stranded structure via complementary base-pairing.
5. Stabilizing stem structure sequence reverse (18 nt): A stem structure constructed by base-pairing with the stabilizing stem structure sequence forward.

Design the ssDNA templates for each RNA probe of the RNA library. You can use the Github page with the instruction (https://github.com/KRK13/FOREST2020/) when performing **steps 2–4**.

1. Prepare RNA sequences and the secondary structures in dot-bracket format.
2. Extract the terminal RNA motifs from the prepared secondary structures (Fig. 2a).
   (a) All hairpin loop structures are collected from the datasets of the RNA secondary structure.
   (b) The stem structure adjacent to each hairpin loop structure is extracted by recognizing the base pair and the structural boundary, and the combination of the two is defined as a single-terminal motif and selected. During this procedure, the stem structure can include one or more bulge and/or loop structures.
   (c) A multi-terminal motif is defined as a multi-branched stem-loop structure containing more than two single-terminal motif structures.
   (d) If a target RNA motif is longer than the RNA length limitation, the motif will repeatedly shorten to the maximum length or to the nearest base-paired end. The selected terminal motif is adopted as an output when the length of the selected RNA is less than the synthetic length limit.
3. Add the stabilizing stem structure sequences and the barcode sequence to each extracted motif (Fig. 2b).
4. Generate the ssDNA templates for in vitro transcription (Fig. 2c) (*see* **Note 1**).

3.2 In Vitro Transcription of the RNA Structure Library

To produce the RNA structure library, use the MEGAshortscript T7 Transcription Kit (Thermo Fisher Scientific) following the manufacturer's instructions.

1. Mix the reaction solution gently. The total volume is 20 μL. Each solution contains ssDNA templates and the ssDNA coding T7 promoter sequence.
2. Mix the content thoroughly by gently flicking the tube. Then microfuge the tube briefly to collect the reaction mixture at the bottom of the tube.
3. Incubate the reaction with ProFlex PCR System (Thermo Fisher Scientific) at 37 °C for 20 h with a lid heated to 105 °C for preventing evaporation.
4. Add 2 μL TURBO DNase (Thermo Fisher Scientific) to the reaction solution, mix by pipetting, and incubate at 37 °C for 15 min.
5. Purify the RNA products using Zymo RNA Clean and Concentrator (Zymo Research).

3.3 3′-Terminal Cy5 Labeling

For detection and quantification on the microarray, RNA probes in the library are labeled with fluorescent dye at the 3′ end.

1. Mix 1× T4 Ligase Buffer (Thermo Fisher Scientific), 100 μM pCp-Cy5 (Jena Bioscience), 10 μM RNA structure library, and 0.5 U/μL T4 RNA Ligase (Thermo Fisher Scientific) in a 100 μL reaction mixture.
2. Incubate the mixture at 16 °C for 48 h under light-shielded conditions.
3. Purify the labeled RNA probes in the library using Zymo RNA Clean and Concentrator (Zymo Research).
4. Store the labeled RNA structure library at −28 °C (*see* **Note 2**).

3.4 Target RNA Enrichment

Enrich target-binding RNA with biochemical assays appropriate for your interested protein. In this chapter, we introduce the step-by-step protocol for the pull-down using His-tagged protein and immunoprecipitation using FLAG-tagged proteins as an example. In all steps, be careful not to leave the beads in a dry state for a long time to prevent them from aggregating.

3.4.1 RNA Pull-Down with His-Tagged Proteins

1. Prepare TALON magnetic beads (Clontech).
   (a) Completely resuspend the magnetic beads by rotating for 5 min.
   (b) Mix the medium slurry thoroughly by vortexing.
   (c) Dispense 20 μL of the homogenous medium slurry into each tube and place the tubes in a magnetic rack.
   (d) Keep for 1 min and then remove the storage solution.

(e) Add 500 μL protein-binding buffer (20 mM HEPES pH 7.5, 80 mM KCl, 20 mM NaCl, 10% glycerol, 2 mM DTT, and 0.1 μg/μL BSA) (*see* **Note 3**) and resuspend the medium to wash the beads.

(f) Centrifuge the samples and remove the liquid.

2. Mix the optimized amount of His-tagged protein (*see* **Note 4**), 20 μL of TALON magnetic beads, and 1 μg of the RNA structure library in 1 mL of protein-binding buffer. Prepare a mixture containing no protein as a control.
3. Incubate the mixture on a rotator at 4 °C for 30 min.
4. Wash the samples with 1× protein-binding buffer three times.
   (a) Add 1× protein-binding buffer and resuspend the medium. Keep for 2 min.
   (b) Mix completely and spin down carefully.
   (c) Remove the liquid.
5. Add 200 μL elution buffer (1% SDS, 10 mM Tris-HCl, 2 mM EDTA) to the magnetic beads and heat the mixture at 95 °C for 3 min.
6. Collect the RNA from the supernatant by removing the magnetic beads.
7. Extract the RNA structure library in the mixture with phenol and chloroform.
8. Purify the RNA structure library with ethanol precipitation.
   (a) Add 1 μL of Etathinmate and 3.3 μL of 3 M NaOAc to 100 μL of RNA solutions and vortex the tubes.
   (b) Add 200 μL of ethanol and vortex the tubes.
   (c) Chill the mixture for at least 15 min at −80 °C.
   (d) Centrifuge the tubes at 20,400 × *g* for 15 min at 4 °C.
   (e) Remove supernatant from the tubes.
   (f) Add 500 μL of 80% EtOH and centrifuge at 20,400 × *g* for 15 min at RT.
   (g) Remove supernatant from the tubes (*see* **Note 5**).
   (h) Incubate the tubes for 3 min at 80 °C with opening the lid (*see* **Note 6**).
   (i) Resuspend the pellet by 20 μL of UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific) in DNA LoBind tube.
   (j) Incubate in a water bath or heat block set at 55–60 °C for 10–15 min.
   (k) Transfer RNA diluted in UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific) to new DNA LoBind tube.

3.4.2 RNA Immunoprecipitation with Flag-Tagged Proteins

The Dynabeads Protein Immunoprecipitation Kit (Thermo Fisher Scientific) is used following the manufacturer's instructions with modifications.

1. Prepare the Dynabeads.
   (a) Vortex the Dynabeads for more than 30 s to resuspend.
   (b) Transfer 20 μL Dynabeads to each tube and place the tubes in a magnetic rack.
   (c) Remove the storage solution from the tube on a magnetic rack.
2. Mix 200 μL Ab Binding Buffer (Thermo Fisher Scientific) and 10 μL M2 mouse monoclonal antibodies with FLAG-tag (Sigma). Prepare a mixture that does not contain the target antibody as a control.
3. Incubate the mixture at room temperature for 13 min with a tube rotator.
4. Place the tubes on a magnetic rack and remove the supernatant.
5. Wash the magnetic beads-Ab-Ag complex with 200 μL Ab Binding Buffer.
6. Add 200 μL cell lysate (1 μg/μL) including Flag-tagged protein to each tube and incubate for 1 h at 4 °C.
7. After the reaction, place the tube on a magnetic rack and remove the supernatant.
8. Wash the magnetic bead-Ab-Ag-RNP complex with 1 mL of 1× protein-binding buffer.
   (a) Add 1× protein-binding buffer and resuspend the medium. Keep for 2 min.
   (b) Mix completely and then spin down carefully.
   (c) Remove the liquid.
9. Mix 500 μL of 1× protein-binding buffer and 500 ng of the Cy5-labeled RNA structure library.
10. Incubate the mixture on a tube rotator at 4 °C for 1 h.
11. Wash the samples with 500 μL of protein-binding buffer three times.
    (a) Add 1× protein-binding buffer and resuspend the medium. Keep for 2 min.
    (b) Mix completely and then spin down carefully.
    (c) Remove the liquid.
12. Add 200 μL elution buffer (1% SDS, 10 mM Tris-HCl, 2 mM EDTA) to the magnetic beads and heat the mixture at 95 °C for 3 min.
13. Collect the RNA from the supernatant by removing the magnetic beads.

14. Extract the RNA structure library in the mixture with phenol and chloroform.
15. Purify the RNA structure library with ethanol precipitation.
    (a) Add 1 μL of Ethachinmate and 3.3 μL of 3 M NaOAc to 100 μL of RNA solutions and vortex the tubes.
    (b) Add 200 μL of ethanol and vortex the tubes.
    (c) Chill the mixture for at least 15 min at −80 °C.
    (d) Centrifuge the tubes at 20,400 × *g* for 15 min at 4 °C.
    (e) Remove supernatant from the tubes.
    (f) Add 500 μL of 80% EtOH and centrifuge at 20,400 × *g* for 15 min at RT.
    (g) Remove supernatant from the tubes (*see* **Note 5**).
    (h) Incubate the tubes for 3 min at 80 °C with opening the lid (*see* **Note 6**).
    (i) Resuspend the pellet by 20 μL of UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific) in DNA LoBind tube.
    (j) Incubate in a water bath or heat block set at 55–60 °C for 10–15 min.
    (k) Transfer RNA diluted in UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific) to new DNA LoBind tube.

3.5 Hybridization and Microarray Scanning

1. Mix 18 μL of enriched RNA (*see* **Note 7**) with 4.5 μL of 10× Blocking Agent (Agilent Technologies) and 22.5 μL Hi-RPM Hybridization Buffer (Agilent Technologies) (*see* **Note 8**).
2. Mix the samples by vortexing and centrifuge to collect the reaction mixture at the bottom of the tube.
3. Incubate the samples for 5 min in a heat block set at 104 °C.
4. Immerse the samples in ice water and incubate for 5 min.
5. Apply the samples to an 8× 60 K Agilent microarray gasket slide (Agilent Technologies) (*see* **Note 9**).
6. Assemble the gasket slide and CGH custom array 8× 60 K (Agilent Technologies) with SureHyb.
7. Perform hybridization in a hybridization oven (Robbins Scientific) for 20 h at 55.5 °C while rotating at 20 rpm.
8. After hybridization, wash the microarray slide for 5 min with Gene Expression Wash Buffer 1 (Agilent Technologies) in a glass container at room temperature.
9. Transfer the microarray slide to a glass container containing Gene Expression Wash Buffer 2 (Agilent Technologies) immersed in a thermostat bath at 37 °C and wash for 5 min (*see* **Note 10**).

10. Obtain fluorescence image data of the microarray by SureScan (Agilent Technologies) (*see* **Note 11**).
11. Convert the captured images of the microarray slide to numeric fluorescence intensities of each spot by Feature Extraction (Agilent Technologies) and GeneSpringGX (Agilent Technologies).

3.6 Calculation of Binding Intensities

1. Calculate the protein-binding intensities of each RNA probe by subtracting the fluorescence intensities of the negative control samples (no protein or no antibody samples) from those of the enriched protein samples.
2. To alleviate the effect of undesired interactions with the barcode region, calculate the average fluorescence intensity of each RNA structure from the intensities of the RNA probes that had the same RNA structure but different RNA barcodes.

4 Notes

1. The template DNA pools used in this study were synthesized using OLIGONUCLEOTIDE LIBRARY SYNTHESIS (OLS, Agilent Technologies). The listed DNA strands were submitted to the SureDesign server as CGH custom array design services (Agilent Technologies). A custom CGH DNA microarray was purchased in 8× 60K array format.
2. Shield the Cy5-labeled RNA from light with aluminum foil at all times to prevent discoloring of the fluorescence dye.
3. The protein-binding buffer should be prepared on the day of the experiment and stored at 4 °C.
4. In order to optimize the protein concentration, it is necessary to prepare samples containing different amounts of protein in your first experiment. In our study with His-tagged recombinant proteins (human U1A and LIN28A), we first examined three protein concentrations (5, 20, and 100 μM) and finally determined 20 μM as appropriate for LIN28A and 5 μM for U1A protein. The optimum concentration was set as the minimum concentration in which the fluorescence intensity was within dynamic range.
5. Remove ethanol as much as possible so that the organic solvent can be easily dried in the next incubation, and if pellets are floating, centrifuge the tubes before removing ethanol.
6. The maximum time is 3 min. When the pellets are translucent and the organic solvent is completely dry, stop incubation and remove the tubes.

7. The concentration of RNA should be quantified in advance. If the concentration is too high, dilute it to 0.4 ng/μL with nuclease-free water. Apply 2.5 μL of diluted RNA and 15.5 μL nuclease-free water.
8. Apply Hi-RPM Hybridization Buffer with a low retention tip, because it is highly viscous.
9. Warm Gene Expression Wash Buffer 2 and the glass container to 37 °C before placing them in the water bath.
10. If there is dust on the slide, remove it with nitrogen gas before use.
11. If there are water droplets on the microarray slide, put the slide back into the glass container containing Wash Buffer 2, pull the slide out of the container slowly so that no water droplets remain, and then set the slide on the scanner.

Acknowledgments

We would like to thank P. Karagiannis for critical reading of the paper. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI No. 20H05626, 15H05722, to H.S.

References

1. Yao Z, Weinberg Z, Ruzzo WL (2006) CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 22:445–452
2. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935
3. Siegfried N, Busan S, Rice G et al (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods 11: 959–965
4. Lu Z et al (2016) RNA duplex map in living cells reveals higher-order transcriptome structure. Cell 165:1267–1279
5. Mustoe AM et al (2018) Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. Cell 173:181–195
6. Zhang Z, Xing Y (2017) CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res 45:9260–9271
7. Hafner M, Katsantoni M, Köster T et al (2021) CLIP and complementary methods. Nat Rev Methods Primers 1:20
8. Dominguez D et al (2018) Sequence, structure, and context preferences of human RNA binding proteins. Mol Cell 70:854–867
9. Komatsu KR, Taya T, Matsumoto S et al (2020) RNA structure-wide discovery of functional interactions with multiplexed RNA motif library. Nat Commun 11:6275
10. Ray D et al (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol 27: 667–670
11. Triboulet R, Pirouz M, Gregory RI (2015) A single Let-7 microRNA bypasses LIN28-mediated repression. Cell Rep 13:260–266
12. Ustianenko D et al (2018) LIN28 selectively modulates a subclass of Let-7 microRNAs. Mol Cell 71:271–283
13. Huang Z-L et al (2018) Identification of G-quadruplex-binding protein from the exploration of RGG motif/G-quadruplex interactions. J Am Chem Soc 140:17945–17955
14. Tippana R, Chen MC, Demeshkina NA, Ferré-D'Amaré AR, Myong S (2019) RNA G-quadruplex is resolved by repetitive and ATP-dependent mechanism of DHX36. Nat Commun 10:1855

Part IV

Bioinformatic and Biophysical Methods to Study Non-coding RNAs

Chapter 18

Computational Methods for the Discovery and Annotation of Viral Integrations

Umberto Palatini, Elisa Pischedda, and Mariangela Bonizzoni

Abstract

The transfer of genetic material between viruses and eukaryotic cells is pervasive. Somatic integrations of DNA viruses and retroviruses have been linked to persistent viral infection and genotoxic effects. Integrations into germline cells, referred to as Endogenous Viral Elements (EVEs), can be co-opted for host functions. Besides DNA viruses and retroviruses, EVEs can also derive from nonretroviral RNA viruses, which have often been observed in piRNA clusters. Here, we describe a bioinformatic framework to annotate EVEs in a genome assembly, study their widespread occurrence and polymorphism and identify sample-specific viral integrations using whole genome sequencing data.

Key words Viral integrations, Lateral gene transfer, Bioinformatics, Next generation sequencing, Genome variability

1 Introduction

Lateral gene transfer (LT) from viruses to eukaryotic cells is a well-recognized phenomenon [1]. Somatic integrations of DNA viruses and retroviruses have been linked to persistent viral infection and genotoxic effects, including various types of cancer [2, 3]. Viral sequences that integrate into germline cells can be transmitted vertically and co-opted for host functions. These Endogenous Viral Elements (EVEs) have long been known, with studies focusing mainly on EVEs from retroviruses in mammalian genomes [4].

In the last decade, innovative Next Generation Sequencing (NGS) technologies paved the way to produce high quality genome assemblies for non-model eukaryotes [5]. This genomic leap led to the discovery of EVEs not only from DNA viruses and retroviruses, but also from nonretroviral RNA viruses in organisms of different evolutionary lineages, including arthropods (i.e., ticks, mosquitoes, bees, crustaceans), fish, snakes, birds, vertebrates (i.e., primates, mouse, rat, opossum), and plants [6–9]. These nonretroviral

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_18,

(nr) EVEs are highly variable in number and many occur proximal to transposable element (TE) sequences in piRNA clusters [7, 10]. Experimental evidence is increasing on the role of nrEVEs in different biological processes such as antiviral immunity, tolerance to cognate viral infection, and regulation of host gene expression [6]. However, the enrichment of nrEVEs in repetitive piRNA clusters complicates their genome annotation and hinders the ability to detect viral integrations through alignment to a reference genome of whole genome sequencing (WGS) data from natural samples or samples collected under hypothesis-driven conditions. The commands and pipelines we describe here were designed to overcome these issues and allow the precise annotation of nrEVEs in a genome assembly and to study their sequence polymorphisms and distribution. Because the genomic situation of nrEVEs is the most complex, our computational framework can be adopted to discover and annotate any viral integration, providing an ad hoc viral database to interrogate.

Our bioinformatic protocol accomplishes three different tasks:

1. Find and annotate EVEs in a genome assembly.
2. Test EVE frequency and their sequence polymorphism using WGS data.
3. Identify, from WGS data, viral integrations that are absent in the reference genome assembly.

The execution of task 1 depends on BLASTx, a popular algorithm for the comparison of proteins to nucleic acids [11]. BLASTx is used to identify sequences in a genome assembly showing similarity to a user-provided set of viral proteins. The resulting BLASTx hits are further filtered by custom scripts (*see* Subheading 3.1) to reduce instances of false positive hits. Tasks 2 and 3 are implemented with WGS short reads that are mapped to a genome assembly. Task 3 is executed by running Vy-PER [12], a pipeline originally designed to identify viral integrations into the human genome, followed by ViR [13], a pipeline designed to improve predictions of new integration sites by accounting for dispersion of reads in repetitive DNA sequences, such as piRNA clusters. The identification of viral integrations using alignment of WGS data to a genome assembly is based on the identification of chimeric reads (a read pair in which one read maps to the host genome, hereafter referred to as host read, and the other read to a virus, hereafter referred to as viral read) and/or unmapped reads, which are extracted from the WGS dataset. The three tasks described above can be executed with any genome assembly as long as a fasta file is available, without the need for gene models.

2 Materials

The framework we describe here is divided into three procedures that can be used to answer different questions about the presence of EVEs in an organism's genome. Each procedure can be run independently depending on the user's needs. The annotation of existing EVEs in a reference assembly (obtained with procedure 1) is strictly required for procedure 2 and is advised before running procedure 3. Performing each of the three parts of this protocol requires different starting data, programs, and computational power. Familiarity with the Unix environment and command line and the ability to understand and run simple scripts is advised. Users should also be familiar with NGS and have a clear idea of the theory behind short reads sequencing technologies, as well as experience with BLAST and BWA [11, 14].

2.1 Required Programs and Scripts

Our protocol can be run entirely with open-source tools and scripts. A Unix machine or cluster is required to install and run all the programs (*see* Subheading 2.2). The programs required for each step of the protocol are listed below. For the installation process and additional dependencies, refer to the guides of each tool. Some of these tools/pipelines necessitate both Python 2 and 3 (https://www.python.org/download/releases/) to be run. It is convenient to install a program for the manipulation of fasta sequences with a graphical user interface and the Integrative Genome Browser (IGV) [15], or a similar software, to visualize alignments of sequencing reads.

2.1.1 Procedure 1, Annotation of EVEs in a Genome Assembly

1. BLAST+ [16] (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) or DIAMOND [17] (https://github.com/bbuchfink/diamond).
2. EVE_finder [18] (https://data.mendeley.com/datasets/d6zf6fvzwn/1).
3. Refine EVEs annotation pipeline (https://github.com/BonizzoniLab/Refine_EVEs_annotation).
4. BEDtools [19] (https://bedtools.readthedocs.io/en/latest/).
5. Virus-Host Classifier [20] (https://github.com/Kzra/VHost-Classifier).
6. Taxonkit [21] (https://bioinf.shenwei.me/taxonkit/).

2.1.2 Procedure 2, Assessment of EVEs Polymorphism

1. GATK [22] (https://gatk.broadinstitute.org/hc/en-us).
2. Platypus [23] (https://www.well.ox.ac.uk/research/research-groups/lunter-group/lunter-group/platypus-a-haplotype-based-variant-caller-for-next-generation-sequence-data).
3. Freebayes [24] (https://github.com/freebayes/freebayes).

4. VarDict [25] (https://github.com/AstraZeneca-NGS/VarDict).
5. BCFtools [26] (http://www.htslib.org/download/).
6. SVD pipeline [27] (https://github.com/BonizzoniLab/SVD).

2.1.3 Procedure 3, Identification of Novel EVEs

1. Vy-PER [12] (https://www.ikmb.uni-kiel.de/vy-per/).
2. ViR [13] (https://github.com/epischedda/ViR).
3. BLAST+ [16] (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/).
4. BEDtools [19] (https://bedtools.readthedocs.io/en/latest/).
5. BWA [14] (http://bio-bwa.sourceforge.net/bwa.shtml).
6. BLAT [28] (http://hgdownload.soe.ucsc.edu/admin/exe/).
7. Phobos (Christoph Mayer, http://www.rub.de/spezzoo/cm/cm_phobos.htm).
8. SAMtools [29] (http://www.htslib.org/download/).
9. Trinity [30] (https://github.com/trinityrnaseq/trinityrnaseq/).

2.2 Computational Power

The most important factors influencing the computational power, running time, and disk occupancy required by the procedures described here are (1) the size of the reference genome assembly, (2) the size and number of the raw sequencing reads (fastq) files and the consequent aligned reads (sam/bam) files, and (3) the size of the databases (of viral proteins and/or genomes) used to search for EVEs. A high-performance computing (HPC) cluster or a dedicated server with at least 32 computing threads and 64GB of RAM are recommended, especially to run Vy-PER as this pipeline can run multiple alignment processes in parallel via a scheduling manager (*See* **Note 1**) [12, 13]. The advantage of running multiple alignments in parallel depends on the behavior of BWA, the most computationally intensive step in the protocol. BWA run time begins to stabilize at about 10–12 threads when aligning a single sample [31]. This means that increasing further the number of threads does not significantly increase the "wall clock" time required to execute the program. Splitting the WGS data (fastq files) into multiple parts and running them separately is the best solution to maximize the outcome of having multiples of eight threads available.

2.3 Notes Before Beginning

1. Carefully read the manuals of the tools described in this protocol. Detailed information on the execution of the programs and their options are available online. Additionally, Vy-PER and ViR include step-by-step descriptions and example files that

can be used to test the configuration before using actual data [12, 13].

2. Commands to execute the three tasks described here are not concatenated to make it easier for users to understand each passage. However, they can be concatenated together by preparing a script or integrating them into a workflow management tool. Commands/scripts can be run in the background using the nohup command or a HPC cluster workload manager.
3. The design of the database of viral sequences for the annotation of EVEs in a reference genome (procedure 1) and for the search of novel EVEs (procedure 3) should be carefully considered, as it varies depending on the research question, the desired sensitivity, and the available computational power.

3 Methods

3.1 Annotation of EVEs in a Genome Assembly

Among the different strategies that have been proposed for EVEs annotation [7, 10, 18, 32–35], we consider the one described below as the best compromise between sensitivity, precision, and computing time. Any reference genome assembly can be screened for viral integrations using an approach based on the BLASTx algorithm included in the BLAST+ suite [16]. BLASTx automatically translates query nucleotide sequences and compares them against a target database of proteins using a heuristic approach which initially finds short matches between two sequences (seeding). After BLASTx-based identification of potential EVEs [11, 18], automatic filtering steps are implemented to reduce false positive results. Filtering relies not only on a "reverse" BLASTx search of the BLASTx hits against the entire NCBI protein nr database but also on a taxonomic analysis of each putative EVE to eliminate results of non-viral origin.

3.1.1 Preliminary Operations

The following items are required before launching any program:

- A DNA fasta file of the genome assembly to be screened. The names of the DNA sequences in the fasta file should not contain spaces or special characters and should be concise to reduce the risk of errors in downstream applications.
- A database of viral proteins in fasta format. Be careful not to include duplicate sequences.
- A large reference protein database to eliminate false positive results. These databases are large (up to 200GB) so ensure to have enough space on the working folder.

The size of the target DNA file and the protein databases will influence both the time required for the BLASTx run and the size of the output file. A regular update of the taxonomic databases used by Taxonkit and VHost-Classifier is required to classify predicted EVEs according to the most recent taxonomy [20, 21]. The update can be done following instructions included in the readme file of the EVEs annotation refinement pipeline (https://github.com/BonizzoniLab/Refine_EVEs_annotation). Of note, Taxonkit requires the taxon database to be in the folder "home/user/.taxonkit" by default. Refer to the tool manual (https://bioinf.shenwei.me/taxonkit/) for more information.

3.1.2 Running the Pipeline

1. If an indexed database is not already available, create one using the command included in the BLAST+ package:

```
$ makeblastdb -in ProteinDatabase.fasta -dbtype prot
```

2. After building the database files, run BLASTx using as query the target assembly:

```
$ blastx -query ReferenceGenome.fasta -db ProteinDatabase.
fasta -evalue 1e-6 -num_threads 8 -outfmt '6 qseqid qstart
qend salltitles evalue qframe pident qcovs sstart send slen'
-out EVEs_target.blastx
```

   The "-evalue" option defines a threshold for the hits on the basis of the Expect Value. The closer to zero the E-Value threshold is set, the more significant the retained hits will be. The "-num_threads" option sets the number of computing threads and should be adjusted based on your system and data. For genome assemblies with a size larger than 1Gbp, we suggest using eight threads, as increasing further the number of threads does not increase the computing speed. Threads can be reduced for smaller assemblies or protein databases or if extended computing time is not an issue.

3. Run the "Blast_to_Bed3.py" script in the eve_finder toolbox to convert the tabular BLASTx output to Browser Extensible Data (BED) format. The output will have the "EVEs.bed" suffix:

```
$ python eve_finder/Blast_to_Bed3.py EVEs_target.blastx
```

   If the script produces an error, try adding an underscore ("_") in the BLASTx output file name.

4. Sort the resulting BED file for position relative to the target assembly and merge overlapping hits, if present:

```
$ bedtools sort -i EVEs_target.EVEs.bed > EVEs.sorted.bed
```

```
$ bedtools merge -i EVEs.sorted.bed -c 4,5,6,7,8,9,10,11 -o collapse,collapse,distinct,collapse,collapse,collapse,collapse,collapse > EVEs.merged.bed
```

5. Use the "Top_score_BED2.py" script in the "eve_finder" toolbox to select the best hit from each cluster of overlapping hits. The output is a bed file containing the position of the uniquely selected EVEs in the reference assembly.

```
$ python EVE_finder/Top_score_BED2.py EVEs.merged.bed EVEs_top.bed
```

6. Get the fasta sequence of the EVEs from the genome assembly. Each EVE will be named with a combination of the best hit viral species and starting genomic position:

```
$ bedtools getfasta -s -name -fi ReferenceGenome.fasta -bed EVEs_top.bed -fo EVEs_top.fasta
```

7. Use the final collection of EVE sequences extracted from the assembly as a query for a BLASTx against a large database of proteins. We suggest using DIAMOND [17] for this search because it is faster than the native BLASTx algorithm in the BLAST+ [16] package when used to align queries on a large database. This reverse BLASTx can be done against any database of choice but, to our knowledge, the most comprehensive and up to date is the NCBI protein nr database (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/), which includes protein sequences from NCBI RefSeq, GenPept, Swissprot, PDB, and other databases.

   When using DIAMOND [17], remember to build your database from the nr database fasta file (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) with the "diamond makedb" command first and then launch the alignment with the "diamond blastx" command:

```
$ diamond makedb --in nr_database.fasta -d nr_diamond.dmnd
$ diamond blastx -d nr_diamond.dmnd -e 1e-06 --threads 8 -f 6 qseqid qstart qend salltitles evalue qframe pident qcovhsp sstart send slen staxids -q TopHits.fasta -o EVEsTop_nr.blastx
```

   When using BLAST+ [16], download and extract the preformatted nr database (https://ftp.ncbi.nlm.nih.gov/blast/db/) and directly use the "blastx" command:

```
$ blastx -query EVEs.top.fasta -db nr/nr_protein -evalue 1e-6
-num_threads 8 -outfmt '6 qseqid qstart qend salltitles evalue
qframe pident qcovs sstart send slen staxids' -out EVEsTop_nr.
blastx
```

8. Run the custom script to select unique NCBI taxonomical IDs and extract hits with similarity to viruses from the BLASTx table. This step is important to remove false positives that do not have similarity to any known viral sequences, have a strong similarity to eukaryotic proteins and/or are low-complexity sequences (e.g., tandem repeats). In addition, this script assigns EVEs to a viral species.

```
$ bash Refine_EVE_Annotation.sh \
 -pipeline_directory Refine_EVE_annotation_folder/ \
 -tool diamond \
-VHC_directory VHost-Classifier/ \
 -file_blastx EVEsTop_nr.blastx \
 -file_bed_tophit EVEsTop.bed \
 -output_directory Output_directory \
 -taxonkit_exe Taxonkit_0.6/taxonkit
```

Change the "-tool" option to "diamond" or "blastx" depending on the chosen tool.

3.1.3 Pipeline Output

Two folders are created in the path defined by the user. The "VHC/" folder contains the files created by VirusHost-Classifier while the "Output/" folder contains the tab separated value (tsv) tables produced by the pipeline. The "CompleteTable_classified" file in the "Output/" folder incorporates comprehensive data about the genomic position of each EVE, the best viral and non-viral BLASTx/DIAMOND hits, and the assignment of each EVE to the viral taxonomy. For a detailed description of the fields in the table, refer to the pipeline readme (https://github.com/BonizzoniLab/Refine_EVEs_annotation).

3.2 Assessment of EVE Polymorphism

We developed the Structural Variants Definition (SVD) pipeline (Fig. 1) to test the occurrence and sequence polymorphism of EVEs in samples collected from the field or under hypothesis-driven experimental conditions (https://github.com/BonizzoniLab/SVD). The SVD pipeline can be applied to WGS data from one or multiple samples, the latter called a population. The SVD pipeline relies on four Variant Callers (Freebayes [24], VarDict [25], GATK [22], and Platypus [23]) for the detection of Single Nucleotide Polymorphisms (SNPs) and Insertions and/or Deletions (INDELs) in the samples and it implements four subsequent steps (Fig. 1): (1) coverage evaluation; (2) variant

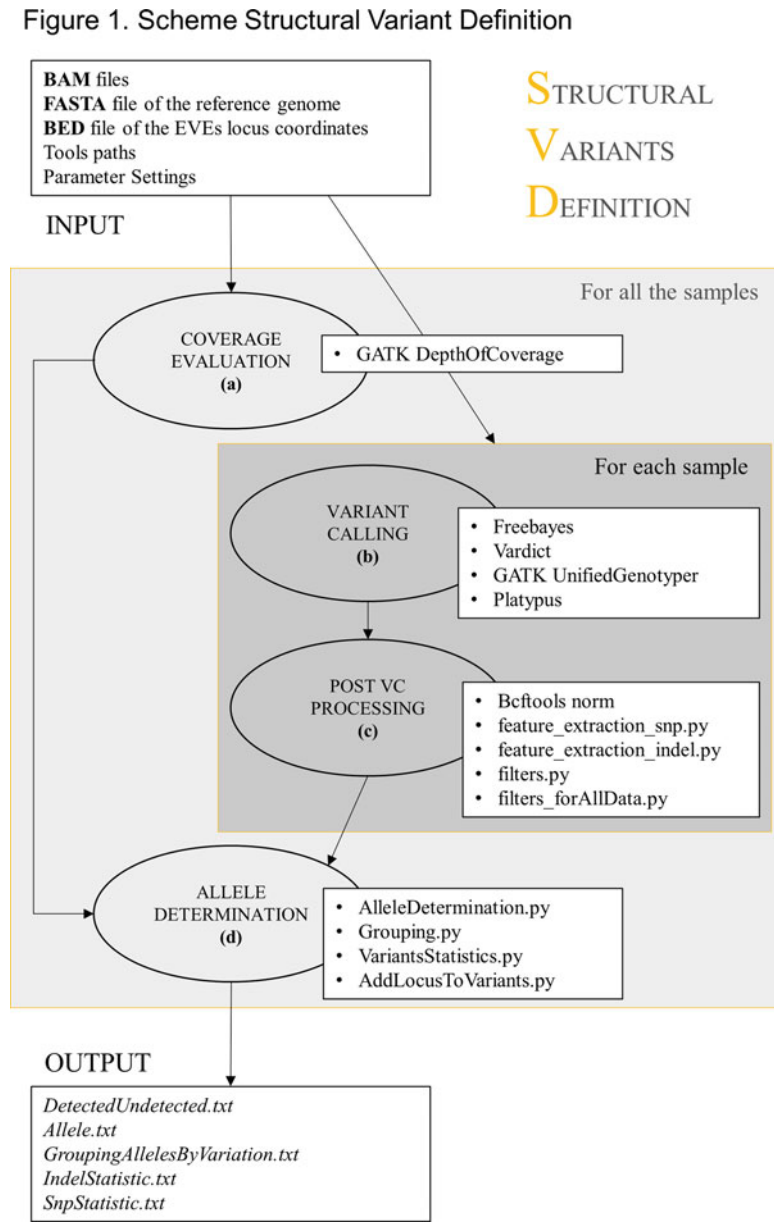


Fig. 1 Graphical overview of the structural variant definition pipeline

calling; (3) post variant calling processing, and (4) final alleles determination.

3.2.1 *Preliminary Operations*

The input files required by the SVD pipeline are:

- A reference genome assembly in fasta format.
- A bed file with coordinates of EVEs which have been annotated in the genome assembly. In addition, Platypus [23] requires a file

where each feature is annotated with the format: "scaffold:start-end". Refer to the documentation of the program for additional information.

- For each sample, a bam file containing the WGS reads aligned to the reference genome assembly. A list containing the full paths of all the bam files must be prepared.

3.2.2 Running the SVD Pipeline

To perform this procedure, WGS data must be aligned on a reference genome using BWA [14] or a similar short reads aligner to produce bam files. Here we will show the alignment of paired-end reads from an individual using bwa mem.

1. Align the paired-end reads with BWA and sort them with SAMTools. Here we concatenate the commands and use default parameters.

```
$ bwa mem -t 16 -R RGID reference_genome_assembly.fasta R1.
fastq R1.fastq | samtools view -@ 8 -m 8G -b Sample.bam
```

2. Sort the bam file.

```
$ samtools sort -@ 8 -m 8G Sample.bam -o Sample.sorted.bam
```

3. Run the SVD pipeline through a single command line where all the input files and programs can be set. The user can define generalized parameters to filter variant calling results. The pipeline will use these parameters to homogenize the identified variants in terms of features common to the different variant callers [27]. The following command runs the pipeline with default parameters. Refer to the program manual for detailed information about the options (https://github.com/BonizzoniLab/SVD).

```
$ bash StructuralVariantDefinition.sh \
 -c samples_BAM_files.list \
 -b reference_EVEs.bed \
 -b_pl platypus_EVEs.txt \
 -f reference_genome_assembly.fasta \
 -i StructuralVariantsDefinition_09_11_18/ \
 -o output_folder/ \
 -fbpath freebayes/bin/freebayes \
 -gkpath gatk.jar \
 -vdpath VarDict/vardict
 --fileR VarDict/teststrandbias.R
 --filePl VarDict/var2vcf_valid.pl \
 -plpath Platypus/bin/Platypus.py \
 -btpath bcftools/bcftools \
```

```
 -th 4 --ram 12g --MIN_MQ 20 --MIN_BQ 20 --MIN_AF 0.1 --MIN_AO
2 --MIN_COV 8 --MAX_DEPTH 5000 --DP_expected_mean 20 -dp 8 -af
0.1 --AFallData 0.1 --MaxStrBias 0.01 --MinLengthINDELallData
6 --NumCallersallData 2 --minReadsAllDef 5 --minLengthAllele
30 --thresholdSimilarity 0.05
```

3.2.3 SVD Pipeline Output

The pipeline produces the following outputs:

- A list of the annotated EVEs that are also present in the analyzed samples. The frequency of EVE occurrence in the sample(s) can be calculated from this information.
- For each sample, a summary table showing the number of variants found in each locus with the predicted zygosity (homo- or heterozygous) in the sample plus additional information on each allele (e.g., number of reads supporting the variant or reference allele, their orientation and mapping quality, which callers identified the variant). For each EVE, a list of detected SNPs and INDELs and a list of different alleles found across all analyzed samples.

Refer to the manual of the pipeline (https://github.com/BonizzoniLab/SVD) for more information about the output fields.

3.3 Identification of Novel EVEs

When WGS data are available, it is possible not only to study the polymorphism of EVEs that are annotated in the reference genome assembly, but also to search for sample-specific EVEs that are not present in the reference assembly. Hereafter we will call the latter as novel EVEs. Several pipelines have been developed in the context of cancer genetics to identify novel viral integrations (for a review, *see* ref. 36). Among these pipelines, Vy-PER (Virus integration detection by Paired End Reads) [12] emerges for its speed, accuracy of predictions and for the possibility to test integrations derived from more than one viral species simultaneously. The discovery of viral integrations from different viral species in several non-model organisms whose available genome assemblies are often still fragmented, and the enrichment of nrEVEs in repetitive DNA, including piRNA clusters, increases the complexity of EVE identification because reads supporting an EVE could be scattered across repetitive regions. To solve this issue, we developed ViR (Fig. 2) [13]. Vy-PER takes advantage of paired-end reads that are first aligned to the reference assembly using BWA (*See* **Note 2**) [12]. Chimeric read pairs, in which one read maps to the reference genome and the other one does not, are selected. In these pairs, the read mapping to the reference genome is referred to as host read. Among the selected unmapped reads, low-complexity reads are discarded using Phobos. The remaining unmapped reads are aligned with BLAT [28] to a user-defined viral genome database.

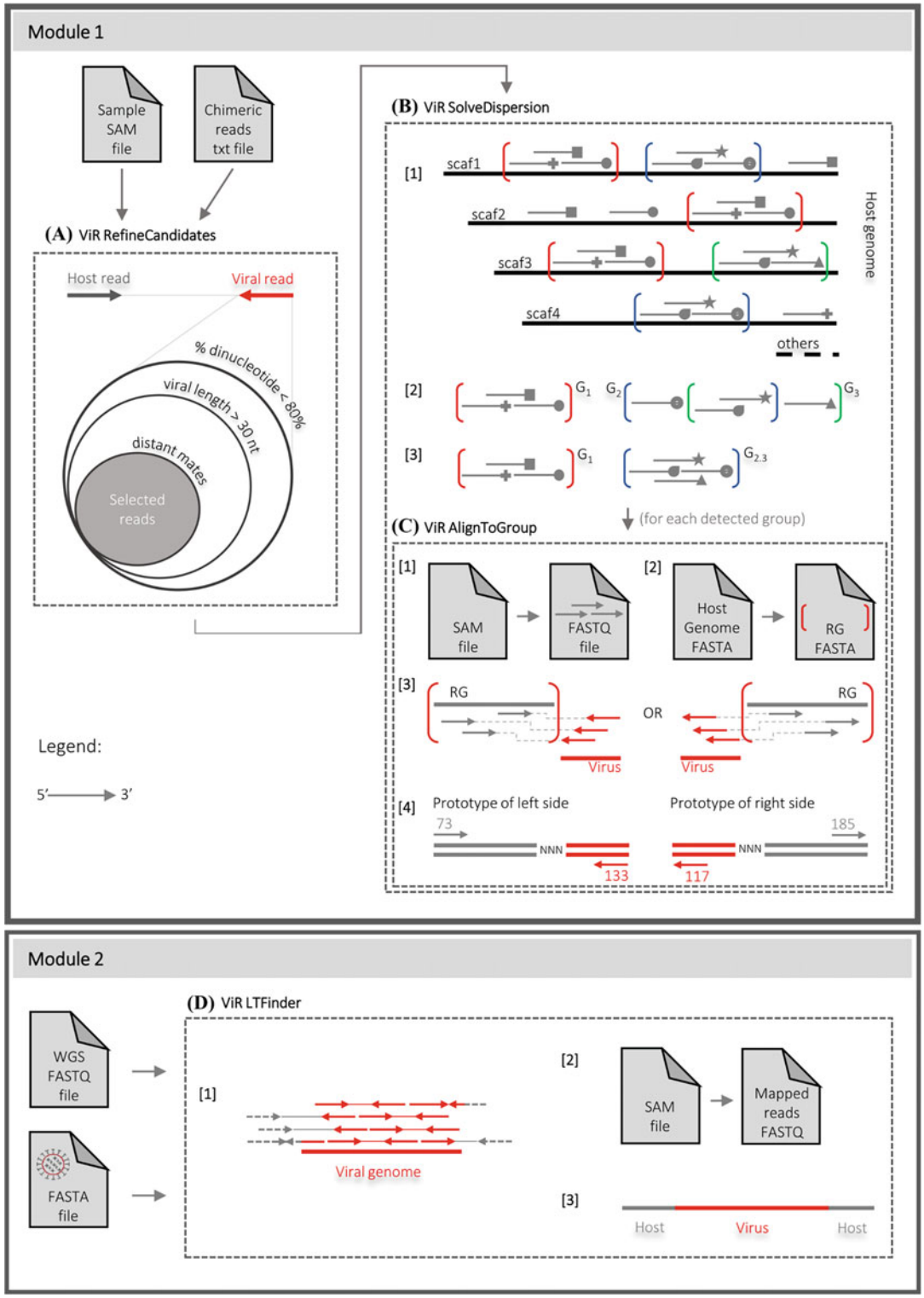


Fig. 2 Conceptual overview of the modules implemented by ViR

Only the top three virus candidates per integration site are retained. Finally, a script refines the output including the final virus integration candidates. ViR works downstream of Vy-PER, or any other paired-end reads-based EVE prediction algorithm, to improve predictions of viral integration sites by addressing the dispersion of reads due to intrasample variability (Fig. 2) [13]. Intrasample variability can be due to repetitive DNA and/or fragmentation in the assembly. ViR is composed of four scripts, which work in two modules (Fig. 2). The first module contains three scripts that help overcoming the dispersion of host reads by grouping together reads that map to regions of the genome with the same sequence, defined as "read groups." The second module includes a single script, "ViR_LTFinder.sh", designed to test for an integration from non-host sequences which have a user-defined percentage of similarity to host sequences. This script can assemble a consensus sequence extending the viral integrations if there are enough unique reads supporting the assembly. Vy-PER output files include a table of the top ten virus candidates, a table of the clusters (genomic windows, number of candidates, virus name and NCBI ID), a detailed table of unfiltered virus candidates and a fasta file for each virus candidate for optional manual alignment/checking. ViR can be integrated with additional information such as (1) a list of the annotated EVEs, to correlate novel EVEs with the ones in the reference assembly; (2) a list of piRNA clusters and TEs to assess the genomic context of novel EVEs.

3.3.1 Preliminary Operations

The input files required by Vy-PER are:

- A reference genome assembly in fasta format.
- Two fastq files for each sample containing the raw paired-end reads. Normally these files have the R1 and R2 suffixes, respectively. Reads should be clean from adapters and low-quality bases should have been trimmed.
- A database of viral genomes or sequences in fasta nucleotide format that will be used by BLAT to select reads of viral origins. Notice that this must be a nucleotide database unlike the one previously used to annotate EVEs, which was a protein database.

ViR requires:

- A reference genome assembly.
- The output from Vy-PER, more specifically:
  - A sam file containing reads aligned on the target genome assembly.
  - A tab separated table of the chimeric reads identified by Vy-PER. The file containing the chimeric reads can be prepared directly from the output of Vy-PER and an example

can be found in the"/ViR examples" folder in the ViR download.

In case sam files from multiple samples must be analyzed with ViR, a list of the sam files full paths must be provided in a text file.

Optional files can be provided to ViR. These files are:

- A bed file with the coordinates of the annotated EVEs in the reference assembly
- A bed file of the piRNA clusters annotated in the reference assembly
- A fasta file containing the nucleotide sequences of predicted TEs

3.3.2 Running Vy-PER

As stated above, Vy-PER is computationally demanding, and it is a good idea to run it on a cluster [12]. With this idea in mind, Forster et al. split Vy-PER into several scripts to allow the user to optimize the number of cores used to run each script and implement a per-lane processing approach that results in better parallelization and faster run time. The Vy-PER download (www.ikmb.uni-kiel.de/vy-per/) includes example scripts for a Linux cluster that can be easily reconfigured for any Unix machine. For the final filtering step, these scripts use a Smith–Waterman implementation on the FPGA-based system RIVYERA (www.sciengines.com), but for general users, Vy-PER can use BLAT on a Linux cluster. We suggest building Vy-PER with a tool or library for parallelized workflow management for high-throughput analysis such as Cosmos [37], uap [38], or HaTSPiL [39]. For clarity, we will describe here the individual Vy-PER scripts and their usage.

1. Align the WGS reads to the host genome using "BWA aln" (*See* **Note 2**) on each fastq file and "BWA sampe" to generate the paired-end sam file:

```
$ bwa aln ReferenceGenome.fasta -n 2 -q 15 -l 5000 -t 8 R1.
fastq > R1.sai
$ bwa aln ReferenceGenome.fasta -n 2 -q 15 -l 5000 -t 8 R2.
fastq > R2.sai
$ bwa sampe ReferenceGenome.fasta R1.sai R2.sai R1.fastq R2.
fastq -a 500 > AllReads.sam
```

The "-a" option in the "bwa sampe" command line defines the maximum insert size and should be changed according to the insert size of the sequenced DNA library. Read groups (RG) can be added using the -r option.

2. Convert the output SAM file into a sorted bam file. If a fai index for the reference genome is not present, create it using the "samtools faidx" command.

```
$ samtools view -@ 4 -u -h -b -T ReferenceGenome.fasta.fai
AlignedSample.bam | samtools sort -@ 4 > SortedReads.bam
```

3. Extract the unmapped read whose mate is mapped on the reference genome to a new sam file, indicating that the original fragment partly matched the reference genome.

```
$ samtools view -f 4 -F 264 > UnmappedReads.sam
```

4. Convert the unmapped reads sam file to fasta using the first script of the Vy-PER package.

```
$ Vy-PER_sam2fas_se UnmappedReads.sam UnmappedReads.fasta
```

5. Identify reads with less than 30 bp (or a user-selected value) of non-repetitive DNA with Phobos and use the "Vy-PER_sam2-fas_se.py" script again to only extract these reads.

```
$ phobos --outputFormat 3 UnmappedReads.fasta phobosReads.fp3
```

```
$ Vy-PER_sam2fas_se -fp3 phobosReads.fp3 30 UnmappedReads.sam
UnmappedReads_passed.fasta
```

6. Run BLAT to compare the extracted reads passing the Phobos filtering to the fasta database of viral genomes. BLAT also requires a 2bit database index that can be created with the "faToTwoBit" executable in the blat suite [28]. Create the 2bit files for the viral genomes database and for the reference assembly fasta file (it will be used by the last Vy-PER script).

```
$ blat -out=blast8 -noTrimA -t=dna -q=dna -maxGap=0 -fastMap
ViralDatabase.2bit UnmappedReads_passed.fasta UnmappedReads_-
blat.txt
```

7. Run the "Vy-PER_blatsam.py" script to extract the reads mapping to viruses from the original paired-end SAM files (these are the novel EVEs) and their mates mapping to the reference genome assembly. A summary table that shows EVEs sorted by their predicted genomic position and the three best virus candidates for each integration site is produced as output.

```
Vy-PER_blatsam SampleID ViralDatabase.fasta UnmappedReads_-
blat.txt UnmappedReads_passed.fasta ReferenceGenome.fasta
AllReads.sam VyPER_OutputTable.txt FastaFolder/
```

8. Run the script "Vy-PER_final_filtering.py" to refine the results and to get the final output. This script removes EVE candidates whose sequence is for more than 50% repetitive DNA and filters the remaining reads to define host/virus chimera clusters.

```
Vy-PER_final_filtering -p 1000 10 10 0.01 0.5 0.95 3 0.90
0 swout.txt 0 VyPER_OutputTable.txt ReferenceGenome.2bit Re-
ferenceGenome.fasta.fai SampleID
```

The "-p" option defines the parameters for clustering the host/virus chimeras and can be adjusted to obtain more or less stringent results. The "swout.txt" file will be empty if no FPGA-based Smith–Waterman [40] is implemented as in this example.

3.3.3 Vy-PER Output

The pipeline produces the following output files:

- A table of the top 10 virus candidates.
- A table of the clusters (including these fields: genomic windows, number of candidates, virus name, and NCBI ID).
- A detailed table of unfiltered virus candidates.
- FASTA files for each virus candidate for optional manual alignment/checking.

In addition, the "rscript_ideogram.R" script can be used to plot in R a summary graph in PDF for the EVEs found by Vy-PER if the hg19 human reference genome assembly (www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) was used.

3.3.4 Running ViR Module 1

The first module of ViR [13] is run on results from Vy-PER [12] or any similar program for the identification of host/virus chimera from paired-end reads. The module is made by three separate scripts which refine Vy-PER predictions and overcome the dispersion of reads due to intrasample variability.

1. Run the "ViR_RefineCandidates.sh" script to select the best candidate pairs from a list of host/virus chimeras supporting a novel EVE. By default, the script excludes viral reads that do not satisfy all of these conditions: (1) the viral portion in the read is shorter than 30 bp; (2) complex and repetitive nucleotides represent more than 80% of each read content; and (3) Read mates do not align together within a 10 kb window in the reference genome. These settings can be modified by the user.

```
$ bash VIR-master/ViR_RefineCandidates.sh \
-work_files_dir VIR-master/ \
-sample_name SampleID \
```

```
-sam_file AllReads.sam \
-chimeric_reads_file sample_chimeric_reads.txt \
-out output_directory_refineCandidates/ \
-reference_fasta ReferenceGenome.fasta \
-path_to_blastn blastn \
-path_to_bedtools bedtools \
-max_percentage_dinucleotide_in_ViralSeq 0.8 -minimum_virus_-
len 30 -blastn_evalue 1e-15 -min_mate_distance 10000
```

The outputs are a table containing the chimeric reads passing the filters and the fasta sequences of the viral and host reads.

2. Run the "ViR_SolveDispersion.sh" script to solve the dispersion of host reads by grouping together reads that map to regions of the genome with the same sequence (Read Groups). Reads mapping to regions of the genome with the same sequence (hereafter called equivalent regions) are identified and merged to define Read Groups. The input for this script is the output folder of the previous script, the reference genome, and a list of samples to be analyzed together. In addition, a fasta file of known transposable element sequences in the species and bed files for known EVEs and piRNA clusters can be provided to correlate the novel EVEs with these genome features. Running this script will automatically launch the third script, "ViR_AlignToGroup.sh".

```
$ bash VIR-master/ViR_SolveDispersion.sh \
 -work_files_dir VIR-master/ \
 -outdict_RefCand output_directory_refineCandidates/ \
 -analysis_name SAMPLE_ID \
 -sample_list sample_list.list \
 -out output_directory_solveDispersion/ \
 -reference_fasta ReferenceGenome.fasta \
  -repreg_fasta  Transposable_elements.fa  -min_TE_al_length
100 \
 -bed_EVE_annotated EVEs.bed -eve_dist 10000 \
 -bed_piwi_clusters piwiClusters.bed -piwi_dist 0 \
 -path_to_blastn blastn \
 -path_to_bedtools bedtools \
 -trinity_exe Trinity \
 -samtools_exe samtools \
 -bwa_exe bwa \
 -merge_dist 1000 -minReads_inRegion 2 -percReadsShared_inGrou-
p_union 0.8
```

3.3.5 ViR Module 1 Output

Five files and a directory "RG/" are produced. The files include detailed information and fasta sequences of the read groups identified in the samples and of all the possible equivalent regions in which they may be positioned in the genome. Also, information about the integration sites and the related boundaries regions are provided. The "RG/" directory includes the realignment of the reads in their relative assigned read group and consensus sequence for the integrations, if enough data was available. A detailed description of the output files can be found in the ViR readme (https://github.com/epischedda/ViR).

3.3.6 Running ViR_LTFinder.sh

The second ViR module contains a single script: "ViR_LTFinder.sh". The script maps WGS reads to a selected non-host fasta sequence using BWA and mapped reads are extracted and used to build de-novo assemblies using Trinity [30]. This module is useful to try to reconstruct and extend the sequence of an EVE that is not present in the reference assembly and can detect any lateral gene transfer event based on the list of non-host sequence used in the search.

```
$ bash ViR_LTFinder.sh \
-analysis_name SAMPLE_ID \
-read_list sample_reads.txt \
-working_dir working_directory/ \
-non_host_fasta non-host_sequence.fasta \
-trinity_exe Trinity
-th 32 -mem 64G \
```

3.3.7 ViR_LTFinder.sh Output

If there are reads supporting a novel integration, the script will produce a de-novo assembly created by Trinity using the consensus of the reads aligned to the sequence. Output of "ViR_LTFinder.sh" includes files for visualization of the aligned reads using a genomic data visualizer (e.g., IGV). Please carefully check the reads supporting the assemblies because multiple, or mosaic, assemblies may be built by Trinity, especially in case of viral integrations occurring in highly repetitive regions such as piRNA clusters.

4 Notes

1. Even though we suggest using a HPC cluster or a server, this protocol could run on a Unix based personal computer. The EVEs annotation (procedure 1) time-limiting commands are the BLASTx searches, which can be run on a personal computer with at least a quad-core CPU and 8 GB of RAM, provided that the genome and the database are not too large. Procedures 2 and 3 could be run on a PC in case of a small genome assembly (<0.5 Gb) and limited size fastq files, but

even in that case we suggest running them on a workstation with at least eight threads.

2. Both the SVD and Vy-PER [12] pipelines use BWA [14] to align WGS reads to a reference genome. Nevertheless, while the SVD pipeline uses BWA mem (which is now considered the standard BWA mapping program), Vy-PER was designed to run with sam alignments produced by BWA aln and sampe. We never tested Vy-PER with alignments produced with BWA mem but it is theoretically possible to use these data with Vy-PER. This possibility should be considered if WGS alignments done with BWA mem are already available or have been produced for the SVD pipeline.

References

1. Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482. https://doi.org/10.1038/nrg3962
2. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618. https://doi.org/10.1038/nrg2386
3. Chen Y, Williams V, Filippova M et al (2014) Viral carcinogenesis: factors inducing DNA damage and virus integration. Cancers 6:2155–2186. https://doi.org/10.3390/cancers6042155
4. Frank JA, Feschotte C (2017) Co-option of endogenous viral sequences for host cell function. Curr Opin Virol 25:81–89. https://doi.org/10.1016/j.coviro.2017.07.021
5. Dheilly NM, Adema C, Raftos DA et al (2014) No more non-model species: the promise of next generation sequencing for comparative immunology. Dev Comp Immunol 45:56–66. https://doi.org/10.1016/j.dci.2014.01.022
6. Blair CD, Olson KE, Bonizzoni M (2020) The widespread occurrence and potential biological roles of endogenous viral elements in insect genomes. Curr Issues Mol Biol 34:13–30. https://doi.org/10.21775/cimb.034.013
7. ter Horst AM, Nigg JC, Dekker FM, Falk BW (2019) Endogenous viral elements are widespread in arthropod genomes and commonly give rise to PIWI-interacting RNAs. J Virol 93:e02124-18. https://doi.org/10.1128/JVI.02124-18
8. Kryukov K, Ueda MT, Imanishi T, Nakagawa S (2019) Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. Virus Res 262:30–36. https://doi.org/10.1016/j.virusres.2018.02.002
9. Horie M, Honda T, Suzuki Y et al (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463:84–87. https://doi.org/10.1038/nature08695
10. Palatini U, Miesen P, Carballar-Lejarazu R et al (2017) Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. BMC Genomics 18:1–15. https://doi.org/10.1186/s12864-017-3903-3
11. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
12. Forster M, Szymczak S, Ellinghaus D et al (2015) Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. Sci Rep 5:11534. https://doi.org/10.1038/srep11534
13. Pischedda E, Crava C, Carlassara M et al (2021) ViR: a tool to solve intrasample variability in the prediction of viral integration sites using whole genome sequencing data. BMC Bioinformatics 22:1–15. https://doi.org/10.1186/s12859-021-03980-5
14. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Mass Genomics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
15. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
16. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

17. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176
18. Whitfield ZJ, Dolan PT, Kunitomi M et al (2017) The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti* genome. Curr Biol 27: 3511–3519.e7. https://doi.org/10.1016/j.cub.2017.09.067
19. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033
20. Kitson E, Suttle CA (2019) VHost-classifier: virus-host classification using natural language processing. Bioinformatics 35:3867–3869. https://doi.org/10.1093/bioinformatics/btz151
21. Shen W, Xiong J (2021) TaxonKit: A practical and efficient NCBI taxonomy toolkit. J Genet Genomics 48(9):844–850. https://doi.org/10.1016/j.jgg.2021.03.006
22. McKenna A, Hanna M, Banks E et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303. https://doi.org/10.1101/gr.107524.110
23. Rimmer A, Phan H, Mathieson I et al (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46:912–918. https://doi.org/10.1038/ng.3036
24. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN]
25. Lai Z, Markovets A, Ahdesmaki M et al (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res 44:1–11. https://doi.org/10.1093/nar/gkw227
26. Danecek P, McCarthy SA (2017) BCFtools/csq: haplotype-aware variant consequences. Bioinformatics 33:2037–2039. https://doi.org/10.1093/bioinformatics/btx100
27. Pischedda E, Scolari F, Valerio F et al (2019) Insights into an unexplored component of the mosquito repeatome: distribution and variability of viral sequences integrated into the genome of the arboviral vector *Aedes albopictus*. Front Genet 10:93. https://doi.org/10.3389/fgene.2019.00093
28. Kent JK (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664. https://doi.org/10.1101/gr.229202
29. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
30. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883
31. Chen S, Senar MA (2019) Exploring efficient data parallelism for genome read mapping on multicore and manycore architectures. Parallel Comput 87:11–24. https://doi.org/10.1016/j.parco.2019.04.014
32. Kondo H, Hirano S, Chiba S et al (2013) Characterization of burdock mottle virus, a novel member of the genus Benyvirus, and the identification of benyvirus-related sequences in the plant and insect genomes. Virus Res 177:75–86. https://doi.org/10.1016/j.virusres.2013.07.015
33. Aguiar ERGR, de Almeida JPP, Queiroz LR et al (2020) A single unidirectional piRNA cluster similar to the flamenco locus is the major source of EVE-derived transcription and small RNAs in Aedes aegypti mosquitoes. RNA 26:581–594. https://doi.org/10.1261/rna.073965.119
34. Fort P, Albertini A, Van-Hua A et al (2012) Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. Mol Biol Evol 29: 381–390. https://doi.org/10.1093/molbev/msr226
35. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6:e1001191. https://doi.org/10.1371/journal.pgen.1001191
36. Chen X, Kost J, Li D (2019) Comprehensive comparative analysis of methods and software for identifying viral integrations. Brief Bioinform 20:2088–2097. https://doi.org/10.1093/bib/bby070
37. Gafni E, Luquette LJ, Lancaster AK et al (2014) COSMOS: Python library for massively parallel workflows. Bioinformatics 30: 2956–2958. https://doi.org/10.1093/bioinformatics/btu385
38. Kämpf C, Specht M, Scholz A et al (2019) uap: reproducible and robust HTS data analysis. BMC Bioinformatics 20:664. https://doi.org/10.1186/s12859-019-3219-1
39. Morandi E, Cereda M, Incarnato D et al (2019) HaTSPiL: a modular pipeline for high-throughput sequencing data analysis.

PLoS One 14:e0222512. https://doi.org/10.1371/journal.pone.0222512

40. Li ITS, Shum W, Truong K (2007) 160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA). BMC Bioinformatics 8:185. https://doi.org/10.1186/1471-2105-8-185

Chapter 19

Bioinformatics Approaches for Determining the Functional Impact of Repetitive Elements on Non-coding RNAs

Chao Zeng, Atsushi Takeda, Kotaro Sekine, Naoki Osato, Tsukasa Fukunaga, and Michiaki Hamada

Abstract

With a large number of annotated non-coding RNAs (ncRNAs), repetitive sequences are found to constitute functional components (termed as repetitive elements) in ncRNAs that perform specific biological functions. Bioinformatics analysis is a powerful tool for improving our understanding of the role of repetitive elements in ncRNAs. This chapter summarizes recent findings that reveal the role of repetitive elements in ncRNAs. Furthermore, relevant bioinformatics approaches are systematically reviewed, which promises to provide valuable resources for studying the functional impact of repetitive elements on ncRNAs.

Key words Repetitive element, Transposable element, Non-coding RNA, ncRNA, Functional element, Bioinformatics

1 Introduction

Non-coding RNA (ncRNA) is a general term for RNAs that are not translated into proteins but exert various functions in cells. Many ncRNAs are involved in a variety of cellular regulations, including transcription and translation, and contribute to the complexity of higher organisms. Among ncRNAs, long ncRNAs (lncRNAs) have recently attracted a great deal of attention [1]. Similar to protein-coding RNAs (mRNAs), most lncRNAs are transcribed by RNA polymerase II, spliced, 5′ capped, and 3′ polyadenylated. In the case of humans, more lncRNA genes than mRNA genes are registered in several databases, such as GENCODE [2], MiTranscriptome [3], FANTOM CAT [4], and NONCODE [5]. While some of these ncRNAs are related to diseases, such as cancers [6, 7], the functions of most lncRNAs remain unknown, and the elucidation of these functions is urgently needed.

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_19,

Typical ncRNAs/lncRNAs have the following different characteristics from mRNAs: (1) they exhibit relatively low expression levels; (2) are expressed in a time- and tissue-specific manner; (3) are frequently nuclear localized; (4) are less conserved among species; and (5) exert their functions by interacting with various biomolecules (DNA, RNA, and protein). These characteristics make ncRNA analysis more difficult than mRNA analysis. Many repeat-derived (especially transposable element (TE)-derived) sequences are significantly enriched in lncRNAs compared with mRNAs [8–10]. These repeat-derived sequences were long thought to be junk, but in recent years, strong evidence has indicated that TEs/repetitive elements are the functional components/domains of lncRNAs (for review, *see* [11–13]). For example, recent studies have shown that the repetitive elements in lncRNAs are related to their expression [14, 15], subcellular localization [16], and others (*see* Subheading 2). In addition, the molecular mechanisms of repetitive elements suggest that they contribute to interactions with other molecules, such as the binding of transcription factors (TFs) [17, 18], RNA-binding proteins (RBPs) [19, 20], DNA [21], and other RNAs [22] (*see* Subheading 3).

Here, we focus on repetitive elements and lncRNA functions and specifically discuss bioinformatics approaches to elucidate their relationship. The following points should be noted from the perspective of bioinformatics: The first is the annotation of repetitive elements utilized in research. Since repeat-derived sequences of functional elements of ncRNAs are assumed to be a small part of the repetitive elements, methods to find the strongly conserved small portions, such as TEs, are necessary. Second, how to utilize high-throughput "omics" data, such as RNA-seq (expression information), ChIP-seq (TF binding), CLIP-seq (RNA–RBP interactions), MARIO (RNA–RNA interactions), GRID-seq (RNA–chromatin interactions), and DRIP-seq (RNA–DNA interactions), is important (see the latter sections for details). When analyzing repeat elements using these data from high-throughput sequencers, multiple-mapped reads to the reference genome/transcriptome should be carefully handled [23]. This is because read sequences derived from repetitive elements may not be uniquely mapped to them. Third, it is important to know what bioinformatics methods/tools and databases are relevant for analyzing repetitive sequences and ncRNA functions. Here, we state the tools and data resources used for investigating the functional impact of repetitive elements on ncRNAs (*see* Table 1).

In this chapter, we describe bioinformatics approaches to elucidate the relationship between repetitive elements and lncRNA functions. The remainder of this paper is organized as follows: Subheading 2 briefly summarizes the emerging roles of repetitive elements in ncRNAs. Subheading 3 describes bioinformatics approaches for studying the role of repetitive elements in ncRNAs,

Table 1
Resources for studying repetitive elements and ncRNAs

Tools/data sources	URL	Description
Non-coding RNA annotations		
GENCODE [2]	https://www.gencodegenes.org/	Continuously updated gene annotations for human and mouse genomes, including protein-coding and non-coding RNAs
MiTranscriptome [3]	http://mitranscriptome.org/	Gene annotations for human ncRNAs predicted from thousands of normal and cancerous RNA-seq data
FANTOM CAT [4]	https://fantom.gsc.riken.jp/cat/	Gene annotations for human lncRNAs predicted from cap analysis of gene expression (CAGE) data
NONCODE [5]	http://www.noncode.org/	Manually retrieved and integrated ncRNA database containing ncRNA gene annotations for 39 species
LNCipedia [73]	https://lncipedia.org/	Manually curated lncRNA annotations for the human genome
Repetitive sequence annotations		
Repbase [42, 74]	https://www.girinst.org/repbase/	Database of representative repetitive sequences for eukaryotic species
Dfam [43, 44]	https://dfam.org/	Database of transposable elements for hundreds of species
RepeatMasker [41]	https://www.repeatmasker.org/	Annotating repetitive sequences of DNA sequences
Tandem Repeat Finder [45]	https://tandem.bu.edu/trf/trf.html	Detecting tandem repeats with k-mer similarity
Tantan [46]	https://gitlab.com/mcfrith/tantan	Detecting tandem repeats and low complexity sequences with HMM
ULTRA [47]	https://github.com/TravisWheelerLab/ULTRA	Detecting tandem repeats with HMM
Red [49]	http://toolsmith.ens.utulsa.edu/	Sensitive to screen transposons and simple repeats
phRAIDER [50]	https://github.com/karroje/phRAIDER	Detecting repeats based on pattern-hunter
P-Clouds [51]	http://www.evolutionarygenomics.com/ProgramsData/PClouds/PClouds.html	Detecting repeats using k-mer counts
RECON [52]	http://eddylab.org/software/recon/	Detecting and classifying repeat sequences from the genome
RepeatScout [53]	http://bix.ucsd.edu/repeatscout/	Detecting repeats by seed-extension strategy

(continued)

Table 1
(continued)

Tools/data sources	URL	Description
LTR_FINDER [58, 59]	http://tlife.fudan.edu.cn/tlife/ltr_finder/	Screening LTRs using suffix-array and the smith–waterman algorithm
LTR_retriever [60]	https://github.com/oushujun/LTR_retriever	Highly accurate and sensitive detection of LTRs
LTRharvest [61]	http://genometools.org/tools/gt_ltrharvest.html	A flexible program for the detection of LTRs.
MITE-hunter [62]	http://target.iplantcollaborative.org/mite_hunter.html	Searching miniature inverted-repeat transposable elements
detectMITE [63]	https://sourceforge.net/projects/detectmite/	MITE detection using the Lempel–Ziv complexity algorithm and CD-HIT
HelitronScanner [64]	https://sourceforge.net/projects/helitronscanner/	Detecting Helitrons with a motif-extracting algorithm
TEclass [65]	http://www.compgen.uni-muenster.de/teclass	Classifying TEs with SVM for k-mer frequencies
REPCLASS [66]	https://sourceforge.net/projects/repclass/	Classifying TEs based on sequence similarity, structural characteristics, and target site duplication
PASTEC [66]	https://urgi.versailles.inra.fr/Tools/PASTEClassifier	Classifying TEs by structural features and sequence similarities
DeepTE [68]	https://github.com/LiLabAtVT/DeepTE	Classifying TEs with convolutional neural networks (CNNs)
TERL [69]	https://github.com/muriloHoracio/TERL	Classifying TEs using CNNs for transformed 2D space data
RepeatModeler2 [54]	https://github.com/Dfam-consortium/RepeatModeler	Ensemble model (LtrHarvest/Ltr_retriever, RepeatScout, and RECON) for repeat detection
EDTA [56]	https://github.com/oushujun/EDTA	Ensemble model for the generation of high-quality and non-redundant TE libraries
REPET [70]	https://urgi.versailles.inra.fr/Tools/REPET	Combined approach for the identification and classification of TEs
PlanTE-MIR DB [75]	http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/	Database of transposable element-related miRNAs for plant genomes
PlaNC-TE [76]	http://planc-te.cp.utfpr.edu.br/	Database of transposable elements for plant genomes
Sequence alignment		
BLAST [55]	https://blast.ncbi.nlm.nih.gov/Blast.cgi	The most commonly used algorithm to compare sequence similarity
LAST [77]	https://gitlab.com/mcfrith/last	Fast genome-level sequence comparison

(continued)

Table 1
(continued)

Tools/data sources	URL	Description
BWA [78, 79]	http://bio-bwa.sourceforge.net/	Mapping of short and long reads against the genome
Bowtie [80, 81]	http://bowtie-bio.sourceforge.net http://bowtie-bio.sourceforge.net/bowtie2	Fast and memory-efficient mapping of short reads to the genome
BLAT [82]	http://genome.ucsc.edu/cgi-bin/hgBlat	BLAST-like alignment tool for the comparison of DNA, RNA, and proteins
STAR [83]	https://code.google.com/archive/p/rna-star/	Fast but memory-intensive mapping tool for RNA-seq data
Tophat [84, 85]	https://ccb.jhu.edu/software/tophat/	Splice-aware mapping of RNA-seq reads to the genome
HISAT [86, 87]	http://www.ccb.jhu.edu/software/hisat https://daehwankimlab.github.io/hisat2/	Fast and sensitive mapping of RNA-seq or DNA-seq reads to the genome
AREM [88]	https://sourceforge.net/projects/arem/	Accounting for multi-mapped reads with expectation-maximum (EM) algorithm
RNA structure		
ViennaRNA package [89]	https://www.tbi.univie.ac.at/RNA/	The most commonly used software suite for the prediction and comparison of RNA secondary structures
CENTROIDFOLD [90]	http://rtools.cbrc.jp/centroidfold/	One of the most accurate tools for the prediction of RNA secondary structures
CapR [91]	https://github.com/fukunagatsu/CapR	Calculating the structural context (stem, exterior loop, hairpin loop, bulge loop, internal loop, and multibranch loop) profiles for RNA sequences
Rtools [92]	http://rtools.cbrc.jp/	Web server for the secondary structural analysis of RNA sequences
RNAz [93]	https://www.tbi.univie.ac.at/software/RNAz/	Predicting the functional RNA secondary structures from multiple sequence alignments
RNAforester [94]	https://bibiserv.cebitec.uni-bielefeld.de/rnaforester	Calculating the RNA secondary structural similarity based on the tree alignment algorithm
RNAmotif [95]	https://github.com/dacase/rnamotif	Identifying structural motifs from RNA sequences
IPknot [96]	http://rtips.dna.bio.keio.ac.jp/ipknot/	Predicting the RNA secondary structure accounting for pseudoknots

(continued)

Table 1
(continued)

Tools/data sources	URL	Description
MXfold2 [97]	https://github.com/keio-bioinformatics/mxfold2/ http://www.dna.bio.keio.ac.jp/mxfold2/	Predicting the RNA secondary structure with deep learning
RNA–RNA, RNA–DNA, RNA–protein interactions		
RIsearch2 [98]	https://rth.dk/resources/risearch/	Predicting large-scale RNA–RNA interactions
RIblast [99, 100]	https://github.com/fukunagatsu/RIblast	Ultrafast prediction of RNA–RNA interactions based on seed-and-extension strategy
LncRRIsearch [101]	http://rtools.cbrc.jp/LncRRIsearch/	Web server for lncRNA-RNA interactome analysis
IntaRNA [102]	http://rna.informatik.uni-freiburg.de/IntaRNA/	Predicting interactions between two RNA sequences
TargetScan [103]	http://www.targetscan.org/	Predicting miRNA target sites in the genome
piRscan [104]	http://cosbi4.ee.ncku.edu.tw/pirScan/	Predicting piRNA targets for a DNA or spliced RNA sequence
Triplexator [105]	http://bioinformatics.org.au/tools/triplexator/	Searching DNA:RNA triplex structures in the genome
TriplexFPP [106]	https://github.com/yuuuuzhang/TriplexFPP	Predicting DNA:RNA triplex based on two-layer CNNs
TDF [107]	http://www.regulatory-genomics.org/tdf	Predicting RNA–DNA interactions in ncRNAs
R-loop DB [108]	http://rloop.bii.a-star.edu.sg/	Computational and experimental data of R-loop-forming sequences
ENCODE [109]	https://www.encodeproject.org/	Datasets, including eCLIP-seq of RNA-binding proteins and RNA-seq in subcellular fractions, etc
MACS [110]	https://github.com/macs3-project/MACS	Originally designed for peak calling for DNA-protein binding sites
Piranha [111]	https://github.com/smithlabcode/piranha	Peak calling for CLIP-seq data
HOMER [112]	http://homer.ucsd.edu/homer/ngs/peaks.html	Peak calling and motif analysis for next-generation sequencing data
MEME [113]	https://meme-suite.org/meme/	Discovering and analyzing sequence motifs
CLAM [114]	https://github.com/Xinglab/CLAM	Analyzing CLIP-seq data while accounting for multi-mapped reads

including de novo identification of repetitive elements (Subheading 3.1), expression (Subheading 3.2), subcellular localization (Subheading 3.3), ncRNA–RNA interaction (Subheading 3.4), ncRNA–DNA interaction (Subheading 3.5), and ncRNA–protein interaction (Subheading 3.6). In Subheading 4, we conclude this review and discuss future research directions.

2 Emerging Roles of Repetitive Elements in ncRNAs

Accumulating evidence suggests that repetitive sequences, including TEs and simple repeats, play an essential role in the biological functions of ncRNAs [11–13]. A part (or all) of a repetitive sequence is a regulatory component (called repetitive element hereafter) that enables host ncRNAs to function through the following three interactions: RNA–RNA, RNA–DNA, and RNA–protein. This section reviews the current studies on repetitive elements in ncRNAs, including nuclear retention, biomolecular condensates, gene expression, and RNA processing and translation.

2.1 Nuclear Retention

Repetitive elements usually facilitate the nuclear localization of host lncRNAs by providing anchoring sites for specific RBPs (Fig. 1a). Lubelsky and Ulisky screened dozens of lncRNAs and untranslated regions (UTRs) for sequences with nuclear retention and found a fragment SIRLOIN (short interspersed nuclear element (SINE)-derived nuclear RNA localization) [16]. This fragment is derived from Alu elements and binds to hnRNPK proteins that can interact with splicing factors, thus promoting the nuclear retention of host RNAs. A subsequent study revealed that hnRNPK triggers nuclear retention only at specific binding sites and sequence contexts in SIRLOIN. Additionally, two binding sites were provided in SIRLOIN for proteins SLTM and SNRNP7. Knockdown of these two proteins affects the nuclear localization of SIRLOIN-containing RNAs [24]. Hacisuleyman et al. introduced the repeating RNA domain (RRD) from mouse *Firre* lncRNA into other cytoplasmic mRNAs and discovered that the RRD could yield a sufficient nuclear retention signal [25]. Of note, this RRD has a significant species specificity with a prominent nuclear retention effect only in rodent lineages. Moreover, hnRNPU protein was observed to bind RRD and affect the nuclear localization of *Firre*. Consistently, LIF3 protein can bind SINEB2 in *Antisense Uchl1* and promote its nuclear retention [26]. Carlevaro-Fita et al. systematically defined four TEs (L1PA16, L2b, MIRb, and MIRc) from lncRNAs associated with nuclear retention and found that their copy number was proportional to the degree of nuclear retention of host lncRNAs [10].

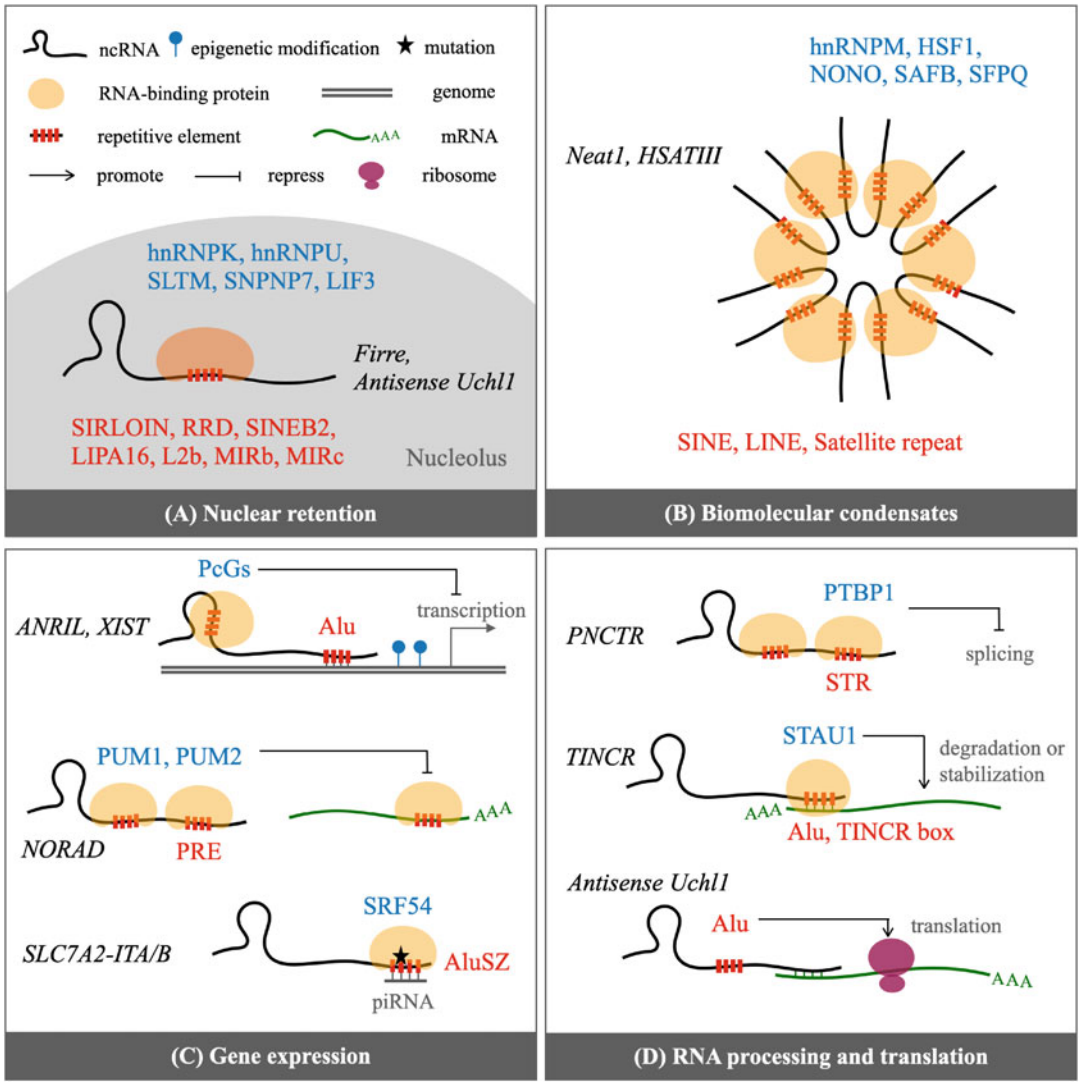


Fig. 1 Repetitive elements in ncRNAs are involved in their biological functions. (**a**) Nuclear retention. (**b**) Biomolecular condensates. (**c**) Gene expression. (**d**) RNA processing and translation. Italic, blue, and red text represent ncRNAs, RNA-binding proteins, and repetitive elements, respectively

2.2 Biomolecular Condensates

Numerous lncRNAs can form biomolecular condensates, such as paraspeckles or stress granules, in cells through repetitive sequences (Fig. 1b). *Neat1* acts as an architectural lncRNA involved in the formation of paraspeckles in cells. Yamazaki et al. found by deletion analysis that the central 8–16 kb region of *Neat1* enriched in long interspersed nuclear elements (LINEs) and SINEs is the critical domain for forming ordered paraspeckles [27]. Moreover, they observed that NONO and SFPQ preferentially bind to this region, making the paraspeckle exhibit phase-separated properties. *HSATIII* consists mainly of satellite repeats and has been reported to be

transcribed under heat shock. It can interact with proteins HSF1, SAFB, and hnRNPM to construct various nuclear stress bodies, which may be associated with chromatin organization and RNA splicing [28, 29]. These results suggest that repetitive sequences may play a determining role in RNA-induced phase separation.

2.3 Gene Expression

LncRNAs regulate gene expression levels in a variety of ways, including chromatin association mediated by repetitive elements, epigenetic modifications, repressor interactions, and Piwi-interacting RNA (piRNA) targeting (Fig. 1c). Several studies have reported that Alu sequences play a vital role in the trans-acting regulation of gene expression in *ANRIL*. He et al. proposed a hypothesis model in which Alu in *ANRIL* can interact directly with target gene enhancers, allowing the Polycomb group proteins (PcGs) bound by *ANRIL* to modify the epigenetic status, thus influencing the expression of downstream genes [21]. Notably, an in silico study analyzed the evolution of *ANRIL* genes in 27 species and observed that TE insertions in exon 3 and exon 8 rendered high conservation of *ANRIL*, suggesting that these two TE insertions may have a biological role [30]. A recent study suggested that the TE in exon 8 is a crucial factor in the genomic association that *ANRIL* possesses [31]. *XIST* is one of the most widely studied lncRNAs. Wutz et al. revealed that the repeat A region (repA) at the 5′ end of *XIST* plays a determinant role in the silencing effect of *XIST* and further showed that the two stem-loop structures formed in this repA are the key to triggering silencing [32]. A subsequent study by Zhao et al. demonstrated that the Polycomb complex PRC2 can be recruited to repA to achieve X-chromosome inactivation [33]. Tichon et al. found that binding sites (PREs) of Pumilio proteins (PUM1 and PUM2) were present in repeat regions of the cytoplasmic lncRNA *NORAD*. Pumilio proteins can repress the expression of mRNAs containing PREs, and *NORAD* decoys Pumilio proteins by repeats to achieve the effect of regulating the expression levels of other mRNAs [34]. Cartault et al. observed that a point mutation in the AluSZ repeat in *SLC7A2-ITA/B* resulted in decreased expression of *SLC7A2-ITA/B* in the brain, leading to neuronal apoptosis [35]. Computational analysis showed that the mutation caused the AluSZ to potentially become a piRNA target or form a stable RNA structure to recruit SRF54, a signal peptide-associated protein. Two mechanistic models elucidated the gain of the deleterious function of AluSZ after point mutation through RNA–RNA or RNA–protein interactions.

2.4 RNA Processing and Translation

LncRNAs are involved in repeat-induced regulation of RNA splicing, degradation, stabilization, and translation (Fig. 1d). *PNCTR* is a lncRNA of concern because of its increased expression in a variety of tumor cells. Yap et al. found that *PNCTR*, which contains multiple short tandem repeats (STRs), can aggregate multiple

PTBP1 proteins in the perinuclear compartment, leading to the modulation of RNA splicing of PTBP1 to promote cell survival [36]. Gong and Maquat discovered that Alu repeats in a lncRNA can form a double-stranded RNA (dsRNA) with Alu in the 3′ UTR of mRNA by intermolecular base-pairing. Such dsRNAs can be recognized by STAU1 proteins that trigger mRNA degradation. Notably, since Alu repeats are widely distributed in the genome, an Alu-containing lncRNA can regulate the decay of multiple mRNAs at the same time, and an Alu-containing mRNA can also be the target of several different Alu-containing lncRNAs [22]. Surprisingly, a similar molecular mechanism has been reported for *TINCR* lncRNAs to stabilize the high expression of target mRNAs. *TINCR* contains a 25 nt repetitive element called the TINCR box, which forms dsRNA with target mRNAs [37]. Note that *TINCR* also functions as a protein-coding RNA and the encoded peptides affect keratinocyte keratinization [38]. Carrier et al. reported that *Antisense Uchl1* shuttled into the cytoplasm when mTORC1 signaling was inhibited and linked to *Uchl1* by base-pairing. SINEB2 triggered the cap-independent translation of *Uchl1*, thus promoting the translation efficiency of *Uchl1* [39]. Further studies suggest that a short hairpin structure in SINEB2 may be a crucial determinant in facilitating translation [40]. Interestingly, as mentioned above, SINEB2 in *Antisense Uchl1* can promote nuclear retention by binding LIF3 [26]. These results suggested that repetitive elements may have distinct regulatory functions owing to their divergent subcellular localization.

3 Bioinformatics Approaches for Studying the Role of Repetitive Elements in ncRNAs

3.1 De Novo Identification of Repetitive Elements

The identification of repetitive elements in the genome is an initial step in the repetitive element analysis. Many bioinformatics methods have been developed to discover repetitive elements, which are split into two categories: library-based and de novo detection. Library-based methods, including RepeatMasker [41], detect repetitive elements by searching sequence similarity against manually curated repeat sequence libraries, such as Repbase [42] and Dfam [43, 44]. De novo detection methods, on the other hand, can find repetitive elements without the use of repeat sequence libraries. Apparently, library-based annotations have high sensitivities for known repetitive elements, whereas de novo detection methods can detect novel repetitive elements. The latter methods can be categorized into the following three types: tandem-repeat detection, interspersed repeat detection, and structured-repeat detection.

Tandem Repeat Finder (TRF) [45] is the most widely used program for tandem repeat detection. TRF searches for candidate

regions in sequences using k-mer (substring with length of k) sliding windows, and then detects tandem repeats by aligning candidates to their surrounding sequences. Other tools, such as tantan [46] and ULTRA [47], build Hidden Markov Models (HMMs) that recognize successive repeating regions in a sequence.

There are two major strategies for interspersed repeat detection: k-mer counting and self-comparison [48]. The premise behind the k-mer counting strategy is that repetitive sequences have similar k-mer profiles. The k-mer counting approach is used by several tools, including Red [49], phRAIDER [50], and P-Clouds [51]. To avoid loss of sensitivities, Red utilizes HMM trained on the distribution of k-mer frequency, phRAIDER adopts spaced seeds that allow mismatches in k-mers, and P-Clouds maps k-mer clusters to the original sequence. Repeat detection based on k-mer counting is faster than that based on self-comparison. On the other hand, RECON [52] and RepeatScout [53] belonging to the self-comparison strategies use sequence alignment scores. Such methods tend to show a higher performance in the evaluation of annotated genomes. RECON, which classifies the results of exhaustive self-alignments by considering the biological characteristics of interspersed repeats, is suitable for discovering interspersed repeats with mutations [54]. In several comparisons of interspersed repeat identification algorithms, RepeatScout, which finds interspersed repeats using the seed-and-extension approach proposed in BLAST [55], has demonstrated high accuracy [48, 56, 57].

Because several TE subclasses have unique structures, there are also approaches to finding structural similarities in sequences rather than sequence homology. For example, LTR_FINDER [58, 59], which is a tool for screening full-length long terminal repeat (LTR) elements, first finds repetitive regions of the terminal and then searches internal domains. Other examples are LTR_retriever [60] and LTR harvest [61] for LTRs, MITE-hunter [62] and detect-MITE [63] for miniature inverted-repeat transposable elements (MITEs), and HelitronScanner [64] for Helitrons. Note that these tools will only detect the targeted subclasses and not all interspersed repeats.

The interspersed repeats were annotated using TE classification methods. TEclass [65] is a standard tool for classifying detected interspersed repeats using support vector machines (SVMs). REPCLASS [66] and PASTEC [67] combined structure-based detection and alignment with repeat sequence libraries. DeepTE [68] and TERL [69] are convolutional neural network (CNN)-based TE classification tools. There are some pipelines that perform a series of steps from interspersed repeat detection to TE annotation by combining multiple existing tools. One such pipeline is RepeatModeler2 [54], which comprises RECON, RepeatScout, LTR_harvest, and LTR_retriever. For other examples, EDTA [56] is a combination of various structure-based methods, and

REPET [70] establishes TE annotations from exhaustive self-alignments (for reviews of repetitive elements detection, see [71, 72]).

3.2 Tissue/Tumor-Specific Expression

Repetitive elements (primarily TEs) in ncRNAs are related to tissue/tumor-specific ncRNA expression. Francescatto et al. confirmed the enrichment of DNA/TcMar-Tigger elements in ncRNAs that are expressed specifically in brain tissue [115]. Based on Nielsen's research [116], they first fitted a linear model (limma [117]) to the expression data from 12 tissues, including the brain, to identify ncRNAs with tissue-specific expression. The occurrence of TEs was then compared using Fisher's exact test between ncRNAs expressed solely in brain tissue (brain-specific ncRNAs) and ncRNAs expressed in two or more different tissues (non-tissue-specific nRNAs). TE enrichment in genomes with substantial changes, such as tumor cells, has been studied via de novo transcriptome assembly. Attig et al. [118] used RNA-seq data from 31 different tumor types and de novo assembled transcripts using Trinity [119]. They then identified ERV elements enriched in tumor-specific ncRNAs using Dfam [43] library-based TE detection. The hypergeometric test has also identified TE subclasses that are enriched in promoters of ncRNAs relative to those of mRNAs in the testis [120]. According to these findings, certain TEs may be responsible for the tissue/tumor-specific expression of ncRNAs.

In tissue/tumor-specific ncRNAs, the role of repetitive elements as cis-regulatory regions has also been investigated. Previous studies have shown that certain TEs can function as tissue/tumor-specific active promoters based on annotated regions [8, 121, 122]. Laurent et al. [122] identified very long non-coding RNAs (vlncRNAs) with tissue/tumor-specific expression based on the fold change between cell lines and primary cells. Based on the frequency of overlap between vlncRNA promoters annotated by ENCODE [109] and TE regions annotated by RepeatMasker [41], they validated the enrichment of TEs in such vlncRNA promoters.

Recently, pipelines have been developed to comprehensively analyze the relationship between tissue/tumor-specific ncRNA expression and repetitive elements. Béguec et al. [123] and Chishima et al. [14] established pipelines to comprehensively capture TEs enriched in tissue-specific ncRNAs. In these pipelines, after detecting specifically expressed lncRNAs and their tissues, the enrichment of TEs in the lncRNAs was investigated based on statistical analysis. To detect tissue-specific lncRNAs, the existing metrics were used for each. Béguec et al. used tau [124], which evaluates the bias of expression levels by normalizing the maximum expression levels among tissues. Chishima et al. adopted ROKU [125], which applies entropy and Akaike's information criterion to detect tissue-specific gene expression patterns. Furthermore, pipelines that focus on dynamic expressions have been developed. Miao

et al. [126] and Shao and Wang [127] constructed pipelines for dynamic TE-containing ncRNA expression analysis at different developmental stages. Miao et al. [126] analyzed the dynamic function of TEs as transcription initiation sites using bulk ATAC-seq data at each developmental stage. Shao and Wang [127] quantified the dynamic expression of TEs at the transcript level using single-cell RNA-seq data obtained from the early stages of embryogenesis. These results revealed the dynamic regulation of TE expression in the preimplantation stage and demonstrated the tissue specificity of TE-containing ncRNA transcripts in early embryogenesis.

3.3 Subcellular Localization

Subcellular localization of lncRNAs predicted by sequence features provides essential clues for analyzing and understanding the biological functions of lncRNAs. Although we have observed that some repetitive elements regulate the localization of lncRNAs [10, 16, 24, 25], studies that incorporate repetitive elements in the prediction of RNA localization are poorly undertaken.

Hamilton et al. proposed an analytical pipeline for identifying RNA secondary structure elements in the genome to detect localization-related sequence features [128]. First, they extracted two similar RNA stem-loop structures from the well-studied GLS (*grk* localization signal) and ILS (*I* factor localization signal) in *Drosophila*. This structure is critical for recognition by the components of the Dynein-dependent localization machinery. After obtaining the sequences by sliding a window over the genome, the sequences were converted into structural data using RNALfold [129]. Finally, RNAdistance [130], RNAforester [94], and RNAmotif [95] were used separately to compare the similarity between the structure data and GLS and ILS stem-loops. They found that G2 and Jockey repeats could form structures similar to GLS and ILS stem-loop structures and validated them using injection assays through which they could induce specific localization in the oocyte. The above approach can encode RNA secondary structure features in repeat elements when predicting lncRNA localization.

Sequencing technologies have allowed us to obtain genome-wide data for mapping lncRNAs to different subcellular compartments. Using these data, we can analyze and characterize the localization of lncRNAs in terms of their sequence features. Zeng et al. exploited ribosome profiling data to define over 1,000 ribosome-associated and ribosome-free lncRNAs in humans and mice [131]. Then, ~100 sequence-related features containing repetitive sequence content were encoded from these lncRNAs. An L1 regularized logistic regression model [132] was used to fit these data to assess various features of ribosome association. Finally, they found that lncRNAs containing LTR repeats were more likely to bind the ribosome, whereas those lncRNAs composing LINE or SINE were more likely to be ribosome-free [133]. Similarly, Nadel

et al. investigated the importance of repetitive elements in chromatin association [134]. They identified DNA:RNA hybrids and density from RNA:DNA immunoprecipitation (RDIP) data in HEK 293T cells. The L1 regression model was used to fit these data and extract crucial sequence features. Sequentially, they found that LINE could facilitate chromatin association.

3.4 ncRNA–RNA Interactions

RNA–RNA interactions based on complementary base pairings are essential mechanisms of action for many ncRNAs. RNA–RNA interactions are more likely to increase target specificity than RNA–protein interactions, and repetitive sequences are important elements for forming the interaction regions between two RNAs. For example, many piRNAs, short RNAs binding with PIWI proteins, have sequences complementary to transposons. These piRNAs silence transposon activities through RNA–RNA interaction with the transposons in animal germ cells to protect the genome from destruction [135]. Another example is microRNA (miRNA). miRNAs are small (approximately 20 nt) RNAs in eukaryotes that suppress the expression of target mRNAs by binding the 3′ UTRs of mRNAs. Some miRNAs and miRNA target sites are derived from transposons, which regulate the expression of various genes, including housekeeping genes, in humans [136, 137]. Furthermore, a transposon acts as a competing endogenous RNA (ceRNA), which binds to miRNAs and maintains mRNA expression by preventing miRNAs from binding to mRNAs. Cho and Paszkwoski discovered a transposon that works as a ceRNA of miRNA171, contributing to root development in rice [138]. As reviewed in Subheading 2, some lncRNAs exert their functions by interacting with other RNAs through intrinsic repetitive elements [22, 37]. Controlling mRNA expression based on RNA–RNA interactions between repetitive elements, as in these cases, seems to be a more common mechanism. Nguyen et al. discovered many transposon–mRNA interactions experimentally and found that the interaction regions in mRNAs were more evolutionarily conserved than the neighboring regions. These results indicated that these interactions have some biological functions [139].

Experimental or computational identification of RNA–RNA interactions is a powerful approach to discover novel repetitive element-associated RNA–RNA interactions. Recently, experimental methods for exhaustive *in vivo* RNA–RNA interaction detection based on high-throughput sequencing have been developed; for example, COMRADES [140], PARIS [141, 142], and RIC-seq [143]. These methods first concatenate interaction regions by cross-linking and proximate ligation, followed by reverse transcription and sequencing the concatenated RNAs. The sequencing reads were aligned by fast RNA-seq aligners, such as STAR [83], and only gapped reads or chiastic mapping reads were extracted (i.e., normally mapping reads were removed from the analysis). Finally,

high-confidence interaction regions were identified by greedy assembling the remaining reads as duplex groups. RNA–RNA interactions identified by these experiments before 2017 have been registered in the RISE database, and the interaction regions can be easily searched using the web interface of the database [144].

Although these experimental methods can detect RNA–RNA interactions with high accuracy, these methods cannot identify interactions involving transcripts with tissue-specific or cell-type-specific expression patterns unless researchers perform the experiments on a particular tissue or cell type. Specifically, as many lncRNAs show tissue-specific expression patterns [145], extensive experiments are required to reveal the whole picture of lncRNA-related RNA–RNA interactions. On the other hand, computational methods can predict the interaction of any RNA, including artificial RNAs that do not exist in nature. However, the prediction accuracy is still not sufficiently high, which means that experimental and computational methods are complementary approaches. For the large-scale interaction prediction of general RNA–RNA interactions, two fast software products, RIsearch2 [98] and RIblast [99, 100], have been developed. Additionally, by using the LncRRIsearch web service, we can search predicted human and mouse lncRNA–RNA interactions by RIblast and investigate tissue-specific or subcellular localized RNA interactions [101]. To detect the RNA–RNA interactions of a specific class of RNA, it is better to use specialized tools for the RNA class. This is because these tools are more accurate when using interaction rules specific to the RNA class. Some examples of such methods are TargetScan for miRNAs [103] and piRscan for *C. elegans* piRNAs [104].

3.5 ncRNA–DNA Interactions

Repetitive sequences can contribute to ncRNA chromatin associations as guides or cofactors. In addition to providing direct RNA–DNA interactions, repetitive sequences can also indirectly trigger ncRNA-chromatin association by binding to RBPs. For example, a technique called silica particle-assisted chromatin enrichment (SPACE) has detected hundreds of RBPs bound to chromatin in mouse embryonic stem (mES) cells [146]. To date, few studies have been conducted on repetitive sequences and ncRNA–DNA interactions that can be briefly categorized as computation-, experiment-, and hybrid-driven.

Computation-driven approaches can predict lncRNA-chromatin associations and analyze the contribution of repetitive sequences in this context [147]. A lncRNA can be directly bound to a specific region of chromatin through base pairing and affects gene expression proximal to that region. Based on this assumption, Deforges et al. predicted candidates for trans-acting lncRNA-chromatin associations using sequence similarity in *Arabidopsis thaliana*. Considering that lncRNAs are also associated with promoters, they extracted all mRNAs with regions containing 2 kb upstream,

5′ UTR, exons, introns, and 3′ UTR, and then utilized BLAST [55] to retrieve lncRNAs with more than 100 nt hits in these regions. Furthermore, hundreds of lncRNA-chromatin associations were identified by positive or negative correlations in expression between lncRNAs and target mRNAs in different samples. Note that a single lncRNA can correspond to multiple distinct chromatin regions. Repetitive sequences may contribute to this multiple-mapping relationship. An intriguing example is the *XLOC_000322* lncRNA bearing SINE repeats, which is predicted to exhibit positive or negative expression correlations for 13 targets. Among these targets, *AT4G04930*, *AT3G234300*, and *AT2G03340* were validated by protoplast transformation to correlate with *XLOC_000322* in terms of expression. Remarkably, this method is not appropriate for predicting cis-acting lncRNA-chromatin associations because of the sequence complementarity between nascent lncRNAs and their loci.

Experiment-driven approaches use high-throughput techniques to detect global RNA–DNA interactions and observe whether repetitive sequences appear remarkably abundant in RNA-DNA-interacting regions employing appropriate controls. The rationale is that the enrichment of certain sequences in these regions means that these specific sequences are subject to evolutionary pressure to undergo selection due to some function. Bonetti et al. developed the RADICL-seq (RNA and DNA-interacting complexes ligated and sequenced) technique to probe RNA–chromatin interactions [148]. In mES cells, nearly 300,000 RNA-DNA-interacting loci were detected using RADICL-seq. Interestingly, more than 95% of the RNAs involved in chromatin association containing small nuclear RNA belong to trans interactions. However, SINE, LINE, and LTR were more likely to appear in a specific pattern in RNAs mapping to cis interactions. Compared with interactions free of repetitive sequences, SINE was enriched in interactions with RNA and DNA distances ranging from 10 kb to 1 Mb, while LINE and LTR were more likely to appear in RNA-chromatin associations of long-range intervals (>100 kb). Likewise, Zeng et al. extracted R-loops (structure of a DNA:RNA hybrid and a displaced DNA) from publicly available DRIP-seq (DNA-RNA immunoprecipitation followed by high-throughput DNA sequencing) data. Considering the distribution of repetitive sequences in the genome, the length and locus-specific distribution of R-loops, and the tendency of R-loops to form in nascent RNAs, the authors established separate control groups to assess the enrichment of repetitive sequences in R-loop-forming regions [149].

Hybrid-driven approaches combine multiple experimental data to validate the predicted results on a computation-driven basis. Bai et al. used a hybrid approach to reveal the contributing role of Alu sequences in enhancer-promoter interactions (EPIs) [150]. First, BLAST [55] was used to establish whether interactions between

enhancers and promoters might be formed by sequence similarity. Sequence-related associations were predicted to be between ~30% of the enhancers and ~40% of the promoters in the human genome. Then, Alu-derived sequence motifs were detected by MEME analysis [113] to be enriched in these EPIs. Intriguingly, Alu depletion was found in these EPIs by comparing the content of Alu in the genome, which was interpreted as serving as a cis-regulatory element that might be subject to some evolutionary restriction. Subsequently, the involvement of Alu repeats in the regulation of gene expression through EPIs was validated using gene and allelic expression data. To elucidate the mechanism by which (e.g., DNA-DNA, RNA-DNA) Alu repeats are mediated in EPI, ChIA-PET [151], GRID-seq [152], and iMARGI [153] for RNA-DNA, Triplexator [105] prediction data for DNA:RNA triplex, and ssDRIP-seq [154] data for R-loop were analyzed. The authors concluded that Alu was involved in the construction of the EPI network by the trans-acting R-loop of promoter and enhancer RNA. Additionally, they found a co-evolutionary relationship between Alus in the enhancer and promoter, further evidencing that Alu plays a regulatory role in the formation of EPI networks.

3.6 ncRNA–Protein Interactions

The human genome comprises approximately 45–60% of TEs [19, 155]. TEs with high sequence similarities are associated with many genomic regions with a regulatory network. Some TE-derived sequences are inserted into RNAs (especially lncRNAs), and RBPs can recognize and bind to these sequences. RNA–RBP interactions can be mapped by CLIP-seq, which uses a strand-specific library and applies a series of databases and algorithms, such as TopHat, GENCODE, CLIP-seq peak calling, and AREM [84, 88, 156, 157]. CLIP-seq reads of specific RBPs were aligned to many TE families using a combination of RepeatMasker, BEDTools, DFAM profile HMM, HMMer, and PoSSuM [41, 43, 158–161]. On average, 12.2% of reads in enhanced CLIP (eCLIP) experiments included repetitive elements annotated by RepBase [19, 74]. Between lncRNAs and mRNAs, RBP-TE associations were similarly enriched or deficient in exons and introns. For STAU1 binding to the Alu sequence, STAU1 sequences were enriched by 3.2–4.1-fold in Alu sequences. For hnRNA C, hnRNP C sequences were 2.3-fold enriched in antisense Alu elements in transcripts. Hundreds of enrichments of sequences were detected in RBP-TE associations. Many RBPs tend to bind to specific sequences and/or structures. The enrichment of RBP-TE pairs showed a clear tendency for RBP to bind to particular subregions within the TE. For instance, hnRNP H1 CLIP-seq reads were aligned to two specific subregions of antisense L2 elements. The conservation of RBP motifs was analyzed using PhyloP [162], and a high mutation rate across RBP motifs was detected. Many non-repetitive sequences in transcripts seemed to accumulate

mutations in RBP motifs. For most RBP motifs, the coverage of CLIP-seq alignments increased in motif instances outside of the repeats. The motifs were conserved significantly in the non-repetitive 3' UTR, intron and lncRNA sequences. Therefore, TE-derived instances of the motifs potentially intercept the sequences of RBPs for the motifs. To examine whether TE binding sites have similar functions to non-repetitive sites, hnRNP C knockdown experiments were performed to define bound and unbound genes from CLIP-seq alignment coverage using peak calling strategies [163–165]. The cumulative distributions of the values of the statistical tests of differential expression by Cuffdiff were plotted for genes bound only in non-repetitive sequences or only in TEs [166]. The expression of genes found only in non-repetitive sequences and only in TEs was similarly increased, which was observed separately for mRNAs and lncRNAs. TE-derived and non-repetitive RBP binding sequences affect the RNA state similarly in RBP knockdown gene expression analyses.

A soft-clustering non-negative matrix factorization (NMF) method for clustering CLIP-seq peaks for RBPs was developed [167]. Soft clustering clusters one RBP into multiple groups, which is necessary for RBP clustering, because many RBPs have several biological functions through binding with cofactor proteins. Conventional hierarchical clustering using cut-tree methods, such as dynamic tree, dynamic hybrid, and static, cannot cluster CLIP-seq peaks properly. For example, the NMF method identified 18 RBP groups, although the conventional methods found only five groups, which were included in the 18 groups. Many known interactions were found using only the NMF method. It also detects binding sequences such that the signal of one RBP peak is weak, and the others are strong, because it takes into account the whole binding strength of the group of RBPs.

Because sequence reads aligned to repetitive sequences were not used to remove multi-mapped reads in the conventional analyses of CLIP-seq data, the frequency of RBP binding to repeat-derived RNA sequences was underestimated [20]. To identify the functional elements of repetitive sequences, subfamily- and nucleotide-based analyses are required using the eCLIP data of repeat-derived RNAs. Novel components of RBP complexes were predicted to regulate the expression of LINE1 sense strand from the analysis of eCLIP data using STAR, Piranha, RepBase, and MACS [74, 83, 110, 111]. The 3′ UTR of L1PA subfamilies contained putative functional elements associated with heterochromatin formation. New candidate components of the splicing complex were found to bind to LINE1 antisense sequences. This method would be useful for predicting functional RNA elements from repeat sequences, including transposons. Previous studies have not focused on the precise patterns of RBP binding to repeat sequences, particularly TE and its subfamily. This study focused on

the patterns of RBP binding to TEs with nucleotide resolution and discovered several short RNA fragments that bind to multiple RBPs and form RBP clusters. RBP binding sites and RNA secondary structures of the short RNA fragments were predicted and evaluated using Rtools, CENTROIDFOLD, CapR, ViennaRNA package, uShuffle, and RNAz [89–93, 168]. These sequences can form stable stem-loop structures.

Without using long sequence similarity, a k-mer-based comparison of lncRNA sequences has been proposed and developed [169]. LncRNAs with the same or similar functions would have sequence similarities, even if sequence alignment algorithms do not identify similarities. The idea is based on the following features of lncRNAs: first, most lncRNAs do not have catalytic activity, and their function is affected by proteins bound to lncRNAs in cells. Second, proteins bind to RNA through 3–8 bases of motifs (k-mers). Third, for a lncRNA, the positions of motifs may not be important for its function. The existence of these motifs may be sufficient for its function, which does not require long sequence similarity.

4 Concluding Remarks

In this review, we summarize the functional roles of repetitive elements in lncRNAs, especially from a bioinformatics viewpoint. Based on bioinformatics analyses of various omics data, we have provided accumulated evidence that repetitive elements contribute to the expression, subcellular localization, binding of other molecules, and so forth. In these studies, basic computational tools, such as read mapping, peak calling, motif detection, RNA secondary structure predictions, and RNA–RNA interaction predictions, as well as databases, such as repeat annotation and lncRNA annotation, play essential roles (cf. Table 1). In the future, further comprehensive analyses integrating large-scale experimental data, bioinformatics tools, and databases will become more important. Bioinformatics techniques needed for further research include methods for finding shorter remnants of repetitive elements with high sensitivity and for more careful handling of multi-map reads (for a review, *see* [23]), which are derived from repetitive sequences, and methods that enable integrative analyses of several omics data. Additionally, several repetitive elements tend to be inserted into one transcript; similar to protein domains, it is important to consider not only a single repetitive element but also the combination of elements to understand their functional relationships.

Acknowledgments

This work was supported by JSPS KAKENHI [grant numbers JP20K15784 to CZ; 16H06279, 16H05879, 17K20032, and JP20H00624 to MH].

References

1. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–159
2. Frankish A, Diekhans M, Ferreira A-M et al (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47:D766–D773
3. Iyer MK, Niknafs YS, Malik R et al (2015) The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47: 199–208
4. Hon C-C, Ramilowski JA, Harshbarger J et al (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. Nature 543: 199–204
5. Zhao L, Wang J, Li Y et al (2021) NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res 49: D165–D171
6. Nguyen TM, Alchalabi S, Oluwatoyosi A et al (2020) New twists on long noncoding RNAs: from mobile elements to motile cancer cells. RNA Biol 17:1535–1549
7. Bao Z, Yang Z, Huang Z et al (2019) LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res 47:D1034–D1037
8. Kelley D, Rinn J (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 13:R107
9. Kapusta A, Kronenberg Z, Lynch VJ et al (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470
10. Carlevaro-Fita J, Polidori T, Das M et al (2019) Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Res 29:208–222
11. Fort V, Khelifi G, Hussein SMI (2021) Long non-coding RNAs and transposable elements: a functional relationship. Biochim Biophys Acta, Mol Cell Res 1868:118837
12. Ali A, Han K, Liang P (2021) Role of transposable elements in gene regulation in the human genome. Life 11:118
13. Johnson R, Guigó R (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA 20:959–976
14. Chishima T, Iwakiri J, Hamada M (2018) Identification of transposable elements contributing to tissue-specific expression of Long non-coding RNAs. Genes 9:23
15. Lynch VJ, Leclerc RD, May G et al (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet 43: 1154–1159
16. Lubelsky Y, Ulitsky I (2018) Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature 555:107–111
17. Sundaram V, Cheng Y, Ma Z et al (2014) Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res 24:1963–1976
18. Sundaram V, Wysocka J (2020) Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philos Trans R Soc Lond Ser B Biol Sci 375:20190347
19. Van Nostrand EL, Pratt GA, Yee BA et al (2020) Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol 21:90
20. Masahiro O, Chao Z, Yukiteru O et al (2021) Binding patterns of RNA binding proteins to repeat-derived RNA sequences reveal putative functional RNA elements. NAR Genom Bioinform 3(3):lqab055
21. Holdt LM, Hoffmann S, Sass K et al (2013) Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. PLoS Genet 9:e1003588
22. Gong C, Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA decay by

duplexing with 3′ UTRs via Alu elements. Nature 470:284–288

23. Deschamps-Francoeur G, Simoneau J, Scott MS (2020) Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J 18: 1569–1576
24. Lubelsky Y, Zuckerman B, Ulitsky I (2021) High-resolution mapping of function and protein binding in an RNA nuclear enrichment sequence. EMBO J 40:e106357
25. Hacisuleyman E, Shukla CJ, Weiner CL et al (2016) Function and evolution of local repeats in the firre locus. Nat Commun 7: 11021
26. Fasolo F, Patrucco L, Volpe M et al (2019) The RNA-binding protein ILF3 binds to transposable element sequences in SINEUP lncRNAs. FASEB J 33:13572–13589
27. Yamazaki T, Souquere S, Chujo T et al (2018) Functional domains of NEAT1 architectural lncRNA induce Paraspeckle assembly through phase separation. Mol Cell 70:1038–1053.e7
28. Jolly C, Metz A, Govin J et al (2004) Stress-induced transcription of satellite III repeats. J Cell Biol 164:25–33
29. Aly MK, Ninomiya K, Adachi S et al (2019) Two distinct nuclear stress bodies containing different sets of RNA-binding proteins are formed with HSATIII architectural noncoding RNAs upon thermal stress exposure. Biochem Biophys Res Commun 516:419–423
30. He S, Gu W, Li Y et al (2013) ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. BMC Evol Biol 13:247
31. Alfeghaly C, Sanchez A, Rouget R et al (2021) Implication of repeat insertion domains in the trans-activity of the long non-coding RNA ANRIL. Nucleic Acids Res 49:4954–4970
32. Wutz A, Rasmussen TP, Jaenisch R (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet 30:167–174
33. Zhao J, Sun BK, Erwin JA et al (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750–756
34. Tichon A, Gil N, Lubelsky Y et al (2016) A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. Nat Commun 7: 12209
35. Cartault F, Munier P, Benko E et al (2012) Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. Proc Natl Acad Sci U S A 109:4980–4985
36. Yap K, Mukhina S, Zhang G et al (2018) A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. Mol Cell 72:525–540.e13
37. Kretz M, Siprashvili Z, Chu C et al (2013) Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature 493:231–235
38. Eckhart L, Lachner J, Tschachler E et al (2020) TINCR is not a non-coding RNA but encodes a protein component of cornified epidermal keratinocytes. Exp Dermatol 29: 376–379
39. Carrieri C, Cimatti L, Biagioli M et al (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature 491:454–457
40. Podbevšek P, Fasolo F, Bon C et al (2018) Structural determinants of the SINE B2 element embedded in the long non-coding RNA activator of translation AS Uchl1. Sci Rep 8: 3189
41. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0, http://www.repeatmasker.org. Accessed 1 May 2021
42. Jurka J, Kapitonov VV, Pavlicek A et al (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467
43. Wheeler TJ, Clements J, Eddy SR et al (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res 41:D70–D82
44. Storer J, Hubley R, Rosen J et al (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA 12:2
45. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580
46. Frith MC (2011) A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res 39:e23
47. Olson D, Wheeler T (2018) ULTRA: a model based tool to detect tandem repeats. ACM BCB 2018:37–46
48. Rodriguez M, Makałowski W (2021) Software evaluation for de novo detection of transposons. bioRxiv. https://doi.org/10.1101/2021.02.08.430290
49. Girgis HZ (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16: 227

50. Schaeffer CE, Figueroa ND, Liu X et al (2016) phRAIDER: pattern-hunter based rapid ab initio detection of elementary repeats. Bioinformatics 32:i209–i215
51. Gu W, Castoe TA, Hedges DJ et al (2008) Identification of repeat structure in large genomes using repeat probability clouds. Anal Biochem 380:77–83
52. Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269–1276
53. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1): i351–i358
54. Flynn JM, Hubley R, Goubert C et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451–9457
55. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST : architecture and applications. BMC Bioinformatics 10:421
56. Ou S, Su W, Liao Y et al (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20:275
57. Saha S, Bridges S, Magbanua ZV et al (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36: 2284–2294
58. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265–W268
59. Ou S, Jiang N (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA 10:48
60. Ou S, Jiang N (2018) LTR_retriever: a highly accurate and sensitive program for identification of Long terminal repeat retrotransposons. Plant Physiol 176:1410–1422
61. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18
62. Han Y, Wessler SR (2010) MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199
63. Ye C, Ji G, Liang C (2016) detectMITE: a novel approach to detect miniature inverted repeat transposable elements in genomes. Sci Rep 6:19688
64. Xiong W, He L, Lai J et al (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A 111:10263–10268
65. Abrusán G, Grundmann N, DeMester L et al (2009) TEclass--a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329–1330
66. Feschotte C, Keswani U, Ranganathan N et al (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1: 205–220
67. Hoede C, Arnoux S, Moisset M et al (2014) PASTEC: an automatic transposable element classification tool. PLoS One 9:e91929
68. Yan H, Bombarely A, Li S (2020) DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics 36:4269–4275
69. MHP d C, Domingues DS et al (2021) TERL: classification of transposable elements by convolutional neural networks. Brief Bioinform 22:bbaa185
70. Flutre T, Duprat E, Feuillet C et al (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526
71. Goerner-Potvin P, Bourque G (2018) Computational tools to unmask transposable elements. Nat Rev Genet 19:688–704
72. Satovic E (2020) Tools and databases for solving problems in detection and identification of repetitive DNA sequences. Period Biol 121-122:7–14
73. Volders P-J, Anckaert J, Verheggen K et al (2019) LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res 47:D135–D139
74. Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6: 11
75. Lorenzetti APR, de Antonio GYA, Paschoal AR, Domingues DS (2016) PlanTE-MIR DB: a database for transposable element-related microRNAs in plant genomes. Funct Integr Genomics 16:235–242
76. Pedro DLF, Lorenzetti APR, Domingues DS et al (2018) PlaNC-TE: a comprehensive knowledgebase of non-coding RNAs and transposable elements in plants. Database 2018:bay078
77. Kielbasa SM, Wan R, Sato K et al (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21:487–493

78. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-Wheeler transform. Bioinformatics 25:1754–1760
79. Li H, Durbin R (2010) Fast and accurate long-read alignment with burrows-Wheeler transform. Bioinformatics 26:589–595
80. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
81. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359
82. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12:656–664
83. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21
84. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111
85. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36
86. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360
87. Kim D, Paggi JM, Park C et al (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915
88. Newkirk D, Biesinger J, Chon A et al (2011) AREM: aligning short reads from ChIP-sequencing by expectation maximization. J Comput Biol 18:1495–1505
89. Lorenz R, Bernhart SH, Höner Zu Siederdissen C et al (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:26
90. Sato K, Hamada M, Asai K et al (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic Acids Res 37:W277–W280
91. Fukunaga T, Ozaki H, Terai G et al (2014) CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol 15:R16
92. Hamada M, Ono Y, Kiryu H et al (2016) Rtools: a web server for various secondary structural analyses on single RNA sequences. Nucleic Acids Res 44:W302–W307
93. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 102: 2454–2459
94. Höchsmann M, Töller T, Giegerich R et al (2003) Local similarity in RNA secondary structures. Proc IEEE Comput Soc Bioinform Conf 2:159–168
95. Macke TJ, Ecker DJ, Gutell RR et al (2001) RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 29:4724–4735
96. Sato K, Kato Y, Hamada M et al (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics 27: i85–i93
97. Sato K, Akiyama M, Sakakibara Y (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun 12:941
98. Alkan F, Wenzel A, Palasca O et al (2017) RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. Nucleic Acids Res 45:e60
99. Fukunaga T, Hamada M (2017) RIblast: an ultrafast RNA–RNA interaction prediction system based on a seed-and-extension approach. Bioinformatics 33:2666–2674
100. Fukunaga T, Hamada M (2018) A novel method for assessing the statistical significance of RNA-RNA interactions between two Long RNAs. J Comput Biol 25:976–986
101. Fukunaga T, Iwakiri J, Ono Y et al (2019) LncRRIsearch: a web server for lncRNA-RNA interaction prediction integrated with tissue-specific expression and subcellular localization data. Front Genet 10:462
102. Mann M, Wright PR, Backofen R (2017) IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. Nucleic Acids Res 45:W435–W439
103. Agarwal V, Bell GW, Nam J-W et al (2015) Predicting effective microRNA target sites in mammalian mRNAs. elife 4:e05005
104. Wu W-S, Huang W-C, Brown JS et al (2018) pirScan: a webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. Nucleic Acids Res 46:W43–W48
105. Buske FA, Bauer DC, Mattick JS et al (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome Res 22:1372–1381
106. Zhang Y, Long Y, Kwoh CK (2020) Deep learning based DNA:RNA triplex forming potential prediction. BMC Bioinformatics 21:522
107. Kuo C-C, Hänzelmann S, Sentürk Cetin N et al (2019) Detection of RNA-DNA binding sites in long noncoding RNAs. Nucleic Acids Res 47:e32

108. Jenjaroenpun P, Wongsurawat T, Yenamandra SP et al (2015) QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. Nucleic Acids Res 43:W527–W534
109. Davis CA, Hitz BC, Sloan CA et al (2018) The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46:D794–D801
110. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
111. Uren PJ, Bahrami-Samani E, Burns SC et al (2012) Site identification in high-throughput RNA-protein interaction data. Bioinformatics 28:3013–3020
112. Heinz S, Benner C, Spann N et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576–589
113. Bailey TL, Boden M, Buske FA et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–W208
114. Zhang Z, Xing Y (2017) CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res 45: 9260–9271
115. Francescatto M, Vitezic M, Heutink P et al (2014) Brain-specific noncoding RNAs are likely to originate in repeats and may play a role in up-regulating genes in cis. Int J Biochem Cell Biol 54:331–337
116. Nielsen MM, Tehler D, Vang S et al (2014) Identification of expressed and conserved human noncoding RNAs. RNA 20:236–251
117. Ritchie ME, Phipson B, Wu D et al (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47
118. Babaian A, Mager DL (2016) Endogenous retroviral promoter exaptation in human cancer. Mob DNA 7:24
119. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652
120. Davis MP, Carrieri C, Saini HK et al (2017) Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. EMBO Rep 18: 1231–1247
121. Jang HS, Shah NM, Du AY et al (2019) Transposable elements drive widespread expression of oncogenes in human cancers. Nat Genet 51:611–617
122. St Laurent G, Shtokalo D, Dong B et al (2013) VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol 14:R73
123. Le Béguec C, Wucher V, Lagoutte L et al (2018) Characterisation and functional predictions of canine long non-coding RNAs. Sci Rep 8:13444
124. Yanai I, Benjamin H, Shmoish M et al (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21: 650–659
125. Kadota K, Ye J, Nakai Y et al (2006) ROKU: a novel method for identification of tissue-specific genes. BMC Bioinformatics 7:294
126. Miao B, Fu S, Lyu C et al (2020) Tissue-specific usage of transposable element-derived promoters in mouse development. Genome Biol 21:255
127. Shao W, Wang T (2021) Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data. Genome Res 31:88–100
128. Hamilton RS, Hartswood E, Vendra G et al (2009) A bioinformatics search pipeline, RNA2DSearch, identifies RNA localization elements in drosophila retrotransposons. RNA 15:200–207
129. Hofacker IL, Priwitzer B, Stadler PF (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics 20:186–190
130. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429–3431
131. Zeng C, Fukunaga T, Hamada M (2018) Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. BMC Genomics 19:414
132. Cox DR (1959) The regression analysis of binary sequences. J R Stat Soc Series B Stat Methodol 21:238–238
133. Zeng C, Hamada M (2018) Identifying sequence features that drive ribosomal association for lncRNA. BMC Genomics 19:906
134. Nadel J, Athanasiadou R, Lemetre C et al (2015) RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. Epigenetics Chromatin 8:46
135. Iwasaki YW, Siomi MC, Siomi H (2015) PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem 84:405–433

136. Petri R, Brattås PL, Sharma Y et al (2019) LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLoS Genet 15:e1008036
137. Piriyapongsa J, Mariño-Ramírez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337
138. Cho J, Paszkowski J (2017) Regulation of rice root development by a retrotransposon acting as a microRNA sponge. elife 6:e30038
139. Nguyen TC, Cao X, Yu P et al (2016) Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. Nat Commun 7: 12023
140. Ziv O, Gabryelska MM, Lun ATL et al (2018) COMRADES determines in vivo RNA structures and interactions. Nat Methods 15: 785–788
141. Zhang M, Li K, Bai J et al (2021) Optimized photochemistry enables efficient analysis of dynamic RNA structuromes and interactomes in genetic and infectious diseases. Nat Commun 12:2344
142. Lu Z, Zhang QC, Lee B et al (2016) RNA duplex map in living cells reveals higher-order transcriptome structure. Cell 165: 1267–1279
143. Cai Z, Cao C, Ji L et al (2020) RIC-seq for global in situ profiling of RNA–RNA spatial interactions. Nature 582:432–437
144. Gong J, Shao D, Xu K et al (2018) RISE: a database of RNA interactome from sequencing experiments. Nucleic Acids Res 46: D194–D201
145. Iwakiri J, Terai G, Hamada M (2017) Computational prediction of lncRNA-mRNA interactions by integrating tissue specificity in human transcriptome. Biol Direct 12:15
146. Rafiee M-R, Zagalak JA, Sidorov S et al (2021) Chromatin-contact atlas reveals disorder-mediated protein interactions. Nucleic Acids Res 49:13092–13107
147. Deforges J, Reis RS, Jacquet P et al (2019) Prediction of regulatory long intergenic non-coding RNAs acting in trans through base-pairing interactions. BMC Genomics 20:601
148. Bonetti A, Agostini F, Suzuki AM et al (2020) RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. Nat Commun 11:1018
149. Zeng C, Onoguchi M, Hamada M (2021) Association analysis of repetitive elements and R-loop formation across species. Mob DNA 12:3
150. Bai X, Li F, Zhang Z (2021) A hypothetical model of trans-acting R-loops-mediated promoter-enhancer interactions by Alu elements. J Genet Genomics 48:1007–1019
151. Fullwood MJ, Liu MH, Pan YF et al (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462:58–64
152. Li X, Zhou B, Chen L et al (2017) GRID-seq reveals the global RNA–chromatin interactome. Nat Biotechnol 35:940–950
153. Wu W, Yan Z, Nguyen TC et al (2019) Mapping RNA–chromatin interactions by sequencing with iMARGI. Nat Protoc 14: 3243–3272
154. Xu W, Xu H, Li K et al (2017) The R-loop is a common chromatin feature of the Arabidopsis genome. Nat Plants 3:704–714
155. Kelley DR, Hendrickson DG, Tenen D et al (2014) Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. Genome Biol 15:537
156. Harrow J, Frankish A, Gonzalez JM et al (2012) GENCODE: the reference human genome annotation for the ENCODE project. Genome Res 22:1760–1774
157. Kelley D, CLIP-Seq peak calling, https://github.com/davek44/CLIP-Seq. Accessed 1 May 2021
158. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
159. Wheeler TJ, Eddy SR (2013) Nhmmer: DNA homology search with profile HMMs. Bioinformatics 29:2487–2489
160. Beckstette M, Homann R, Giegerich R et al (2006) Fast index based algorithms and software for matching position specific scoring matrices. BMC Bioinformatics 7:389
161. Stegmaier P, Kel A, Wingender E et al (2013) A discriminative approach for unsupervised clustering of DNA sequence motifs. PLoS Comput Biol 9:e1002958
162. Pollard KS, Hubisz MJ, Rosenbloom KR et al (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121
163. Polymenidou M, Lagier-Tourenne C, Hutt KR et al (2011) Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci 14:459–468
164. Glaz J, Pozdnyakov V, Wallenstein S (2009) Scan statistics: methods and applications.

Springer Science & Business Media, New York

165. Lee H, Schatz MC (2012) Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28:2097–2105
166. Trapnell C, Hendrickson DG, Sauvageau M et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31:46–53
167. Li YE, Xiao M, Shi B et al (2017) Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA–protein binding sites. Genome Biol 18:169
168. Jiang M, Anderson J, Gillespie J et al (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics 9:192
169. Kirk JM, Kim SO, Inoue K et al (2018) Functional classification of long non-coding RNAs by k-mer content. Nat Genet 50:1474–1482

Chapter 20

Extending and Running the Mosquito Small RNA Genomics Resource Pipeline

Gargi Dayama, Katia Bulekova, and Nelson C. Lau

Abstract

The Mosquito Small RNA Genomics (MSRG) resource is a repository of analyses on the small RNA transcriptomes of mosquito cell cultures and somatic and gonadal tissues. This resource allows for comparing the regulation dynamics of small RNAs generated from transposons and viruses across mosquito species. This chapter covers the procedures to set up the MSRG resource pipeline as a new installation by detailing the necessary collection of genome reference and annotation files and lists of microRNAs (miRNAs) hairpin sequences, transposon repeats consensus sequences, and virus genome sequences. Proper execution of the MSRG resource pipeline yields outputs amenable to biologists to further analyze with desktop and spreadsheet software to gain insights into the balance between arthropod endogenous small RNA populations and the proportions of virus-derived small RNAs that include Piwi-interacting RNAs (piRNAs) and endogenous small interfering RNAs (siRNAs).

Key words Small RNA deep sequencing, RNAi, Genomics, Transposons, Viruses

1 Introduction

Mosquitoes greatly impact human health in many temperate regions around the world by serving as a vector of prolific pathogens like flaviviruses and bacterial parasites causing malaria. Although we do not fully understand why mosquitoes are such competent transmission vectors for pathogenic arboviruses, one hypothesis is that mosquitoes use RNA interference (RNAi) pathways very efficiently to suppress and tolerate arbovirus infections [1]. To monitor the mosquitoes' RNAi responses, one needs to examine the small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), and microRNAs (miRNAs) from deep sequencing data that are plentiful in the National Computational Biotechnology Institute's Sequencing Read Archive (NCBI's SRA) [2].

Several groups have previously shown that mosquitoes can generate siRNAs and piRNAs derived from arbovirus RNA, and

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_20,

that arbovirus infection and blood feeding can stimulate interesting miRNA changes [3–11]. However, a consistent bioinformatics pipeline to quantitate the small RNA types and genomic locations has only recently been described in a new genomics resource for mosquito small RNAs [12]. Although the MSRG resource currently covers four mosquito species with sequenced and annotated genomes, multiple other biomedically relevant mosquito species with emerging genomic information will need to be incorporated in future studies such as the recently re-sequenced and annotated genomes of *Anopheles stephensi* and *Culex tarsalis* [13, 14].

The origin of our small RNA genomics pipeline began with a comparative genomics study of piRNA cluster loci across various animals in different orders (i.e., vertebrates) or families (i.e., dipterans) [15]. The initial pipeline design used reference files and directory paths referring to the UCSC Genome Browser [16], which had a standardized format of downloadable genome sequences, transcriptome sequences, and Repeatmasker files [17]. However, as we extended this pipeline to other organisms like mosquitoes whose genome sequences and annotations are stored in other databases like VectorBase [18], some legacy code and placeholder files remained as vestiges of the initial pipeline.

This eight-step pipeline was designed for three main types of genomics analyses of small RNAs: including (1) annotation of genic and intergenic small RNAs, such as miRNA being matched to a defined miRNA hairpin database while siRNAs (~18–23 nt) and piRNAs (~24–32 nt) are distinguished by length as a proxy; (2) transposable elements consensus sequences, virus genome sequences, structural RNAs and *Wolbachia* genomes sequences analyses; and (3) piRNA ping-pong and phasing patterns analyses. The pipeline is constructed with a series of shell, Perl, Python, and C scripts that utilize various short read-mapping packages like Bowtie as well as BLAST and BLAT. An overview of the pipeline's eight steps is portrayed in Fig. 1.

2 Materials

2.1 Installation and Dependencies

For sharing and portability, we have packaged the pipeline and developed an image with the required tools and dependencies using a secure container system, Singularity. This package can be deployed on various platforms, including high-performance clusters (HPC). Source code of the algorithms and a list of all the required dependencies (such as python, bioperl, Bowtie [19], Cutadapt [20], etc.) can also be found on GitHub repository (https://github.com/laulabbumc/MosquitoSmallRNA). Instructions for installation of the container on the user's machine are detailed in the GitHub repository. Launching the msrg.simg singularity image allows users to execute MSRG pipeline scripts on other

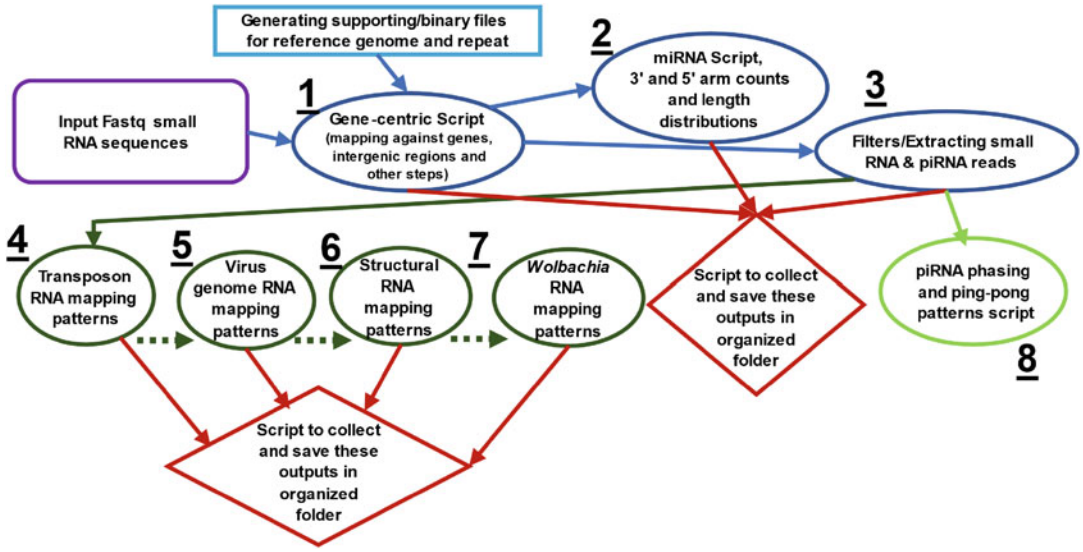


Fig. 1 The small RNA genomics pipeline workflow applied to fruit flies and mosquitoes. The numbers mark the eight steps of the pipeline process

LINUX/UNIX clusters. All dependencies for the pipeline are included in the singularity image and do not need to be installed separately.

2.2 Reference and Supporting Files

Although the pipeline has been developed for mosquito small RNAs, it can be utilized to analyze small RNAs for other species as long as adequate genome sequence assemblies and annotation files are available. We have included an example of all the reference and supporting files required to run the pipeline on the GitHub repository: https://github.com/laulabbumc/MosquitoSmallRNA.

There are seven required reference files needed to build the accessory reference files for the pipeline: (1) the reference genome sequence of the species of interest, (2) genome transcript annotations in Gene Transfer Format (GTF) file in BED format, (3) Transposable Element (TE) consensus sequences fasta file, (4) miRNA hairpins fasta file, (5) virus genomes fasta file, (6) *Wolbachia* genomes fasta file, and (7) a dummy one-line repeat masker file in BED format. The GTF file is converted to "refseq.bed" using the "*gtf2bed*" module. Users of this pipeline can find the genome sequence and annotation files of their species of interest from public databases like the NCBI GenBank, miRbase [21] or VectorBase [18] repositories.

Using the shell script *Update_Database_Refs* the "refseq.bed" is then converted to "structureRNA.bed" and genic/intergenic annotation files in BED format (3′ regulatory, 3′ UTR, 5′ promoter, 5′ UTR, coding region (cds), exon, gene, intron, etc.). All

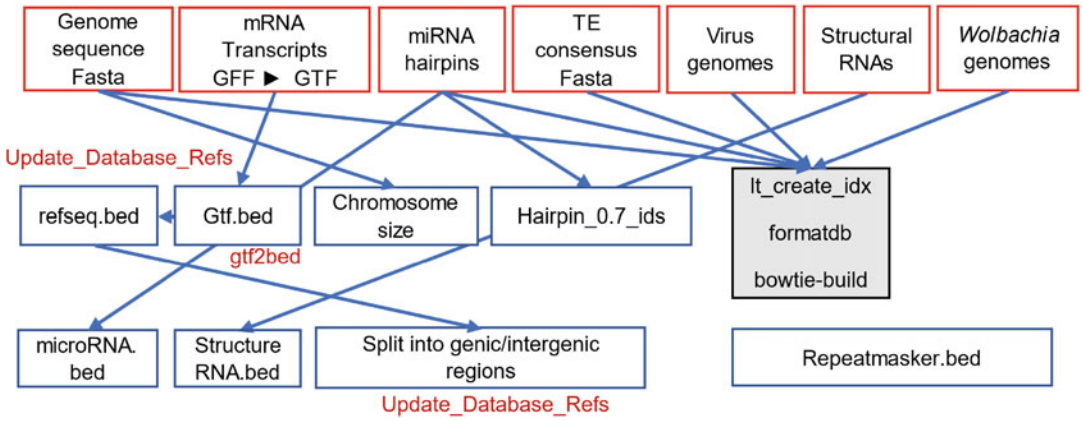


Fig. 2 Generating reference and supporting files required to run the pipeline. Red labels are scripts that convert the source databases in red boxes into supporting reference files. Repeatmasker.bed is a required dummy file from legacy code

the reference and consensus files need to be indexed, formatted and built using Bowtie [19], which we have found to perform better for mapping small RNAs compared to Bowtie2 because the original Bowtie has a strict mapping model allowing up to 3 mismatches and optimized for very short (<40 nt) reads [19], which were the typical limit of high-throughput sequencing data of early small RNA studies such as in [23–26]. Bowtie2 was designed to allow gap extensions and accommodate longer (>50 nt) reads [22], but this can cause false mapping events with small RNAs. Lastly, a file with the chromosome size for the genome needs to be generated in addition to the file containing identification entries of the hairpin file. A diagram of the reference and supporting files is shown in Fig. 2.

2.3 Sample Naming Convention for Input Files and Computing Cluster Space Considerations

Attention should be paid to naming the input small RNA fastq files so that samples and outputs can be readily identified by species, tissue or cell type, lab origin, and experiment date. Some files downloaded from the NCBI SRA are first named in the outputs saving folder by the BioProject identifier (PRJNA#####) or Gene Expression Omnibus (GEO) accession (GSE#####). Our convention is to then name files by the abbreviation for the species, the tissue or cell type, timepoints or dates, and initials for the lab principal investigator. Examples of file naming conventions are displayed here for the *Aedes aegypti* mosquitoes, currently the most commonly studied species: https://laulab.bu.edu/msrg/MSRG_AeAeg.html.

A ~20 M read deep small RNA library will require ~80 GB of disk space to accommodate a series of intermediate files from very large SAM and BAM files generated in the mapping process to a series of BED and custom format intermediate files. The script that

saves the outputs to the sharing folder consolidates and organizes the user-friendly final results files, while all these intermediate files can generally be recreated from running the pipeline again from the source fastq and reference files, although the computer processing time could be extensive (a few days). The rate limiting step in the MSRG pipeline (Fig. 1) is the Gene-centric script that has several input-output hard disk-writing and scratch-space writing steps that are time consuming and cannot be parallelly multitasked to separate servers.

3 Methods

3.1 The Gene-Centric Script, a Central Initial Processing Step

This is one of the main parts of the pipeline to determine read-length distributions, assign reads to defined lists of miRNAs and structural RNAs such as transfer and ribosomal RNAs, then map remaining reads to the genome with annotation overlays that allow for binning and counting of reads mapping to genes and predicted gene models, Transposable Element (TE) consensus sequences, and intergenic regions. Within the genic/intergenic small RNA pipeline, small RNA reads are first trimmed by Cutadapt [20] program that is hardcoded in the pipeline under default parameters as a step to remove the adaptor sequences in the 3′ end.

Trimmed reads are then mapped to a collection of virus sequences using Bowtie [19] allowing 2 mismatches. Reads which are mapped to the exogenous virus genomes list are removed, but reads derived from putative endogenous viral elements [27–30] not present in repeat libraries like Dfam [17] and Repbase [31] would be treated similarly as transposon repeats, and mapped in the following TE counting steps. Next, reads are mapped to miRNAs and structural RNAs, e.g., snRNAs, tRNAs, rRNAs, snoRNAs using Bowtie with 2 mismatches. Reads which are mapped to miRNAs and structural RNAs are removed. Finally, reads are mapped to genomes using Bowtie with 2 mismatches to get the genic/intergenic counts using the genome GTF file. Genic counts are further categorized into 5′ UTR counts, CDS counts, 3′ UTR counts.

Step 1. To find genes and intergenic regions with mapping small RNAs:

```
gene-centric.sh <fastq> <species name> <adapter> <cutoff
(filter reads count)> <cluster extension cutoff>
```

The cutoff (filter reads count) parameter is divided by the value of 1 billion and corresponds to how many reads a gene must have in small RNA reads mapping to it to be counted in the outputs. The cluster extension cutoff is how many reads (i.e., 5 reads) per the next 25 bp window to scan from the annotated end of the 3′UTR to look for more genic 3′UTR small RNAs reads if the

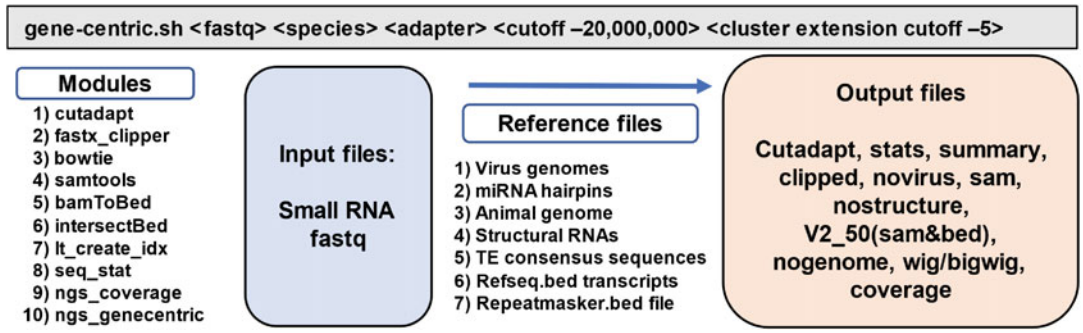


Fig. 3 Diagram of the Gene-centric Script, the most important initial script because all downstream procedures require the results from this script to run properly

transcriptome annotations of 3′UTRs are lacking, such as in [15, 32]. These values are also noted in Subheading 3.4 below as an empirical example for finding genic piRNA across several metazoan genomes [15], although a user can adjust these parameters. Additionally, WIG files are generated and record the normalized read counts within every window of 25 bases for positive strand and negative strands.

See Subheading 4 below for tips in making sure the Gene-centric pipeline's critical components and outputs diagrammed in Fig. 3 are all completed with data outputs.

3.2 miRNA Counting and Other Read-Lengths Separation Scripts

Trimmed reads between 18–23 nt length are extracted and mapped to miRNA hairpin sequences using Bowtie [19] with 2 mismatches. These reads are counted for either mapping to the 5′ or 3′ end of the hairpin miRNA precursor sequence, depending on where the more abundant guide strand is generated by the cells or tissues. Usually, the source file is obtained from a repository like miRbase [33], or from a de novo predicted list using the *miRDeep* program [34]. The counts can be fed into downstream Principal Component Analysis (PCA) and hierarchical clustering using R scripts.

Step 2. Mapping counts for miRNAs in step2_miRNA_pipeline.sh.

```
process_smRNA.sh <fastq> <adapter>
Ngs_smrna_pipeline.sh <fastq> <species>
And other Python and Perl scripts to tabulate miRNA counts and
read lengths
```

The small RNA reads are then apportioned into 18–23 nt long reads corresponding to endogenous siRNAs and 24–35 nt long reads corresponding to likely piRNAs. In addition, other low complexity reads checked from the Gene-centric pipeline are removed before the intermediate files are inputted into this step.

Step 3. Removing low complexity counts, extracting 18–23 nt long small RNA reads, and extracting 24–35 nt long piRNA reads in step3_extract-s_and_pi-RNA.sh:

```
dust_prinseq_local.sh <sample.trim.fastq.uq.polyn>
perl extract_fasta_sequence_given_len_range.pl <sample.trim.
fastq.uq.polyn> 18 23 <sample.18_23.trim.fastq.uq.polyn>
perl extract_fasta_sequence_given_len_range.pl <sample.trim.
fastq.uq.polyn> 24 35 <sample.24_35.trim.fastq.uq.polyn>
```

3.3 Transposable Elements, Viruses, Structural RNAs, and Wolbachia Genomes (TVSW) Scripts

The input file for the TVSW scripts should include reads where adaptor sequences and miRNA reads have been removed. These reads are mapped to TE consensus sequences using Bowtie with 2 mismatches and to virus genomes using Bowtie with 1 mismatch to reduce the frequency of tRNA fragments potentially mapping to viruses (the validity of these tRNA fragment mappings is still being investigated). Finally, the mapping patterns with respect to TVSW are plotted with an R script.

The outputs from these steps can be further analyzed with hierarchical clustering with the Python Seaborn *Clustermap* function using Euclidean distance and average linkage clustering method. PCA can also be carried out by R *prcomp* function, with plots generated by the *ggplot* function. Manual approaches are then needed for curating genic and intergenic piRNA Cluster Loci (piRCL) and predicting the piRNA targets. The pipeline is now set up to save the results from each of the TVSW scripts that is completed before the next script is called.

In these next four steps (Step 4, 5, 6, and 7), each of the scripts that get TEs, viruses, structural RNAs, and *Wolbachia* counts are run individually with a results-saving step that has to be completed before proceeding to the next script. To avoid generating too many files that may consume a lot of disk space, these scripts overwrite the previous scripts generated files and re-use the set of the same base files in the plotting function to generate the small RNA coverage pattern plots across the entries. Thus, the scripts carry out the copying of a multi-fasta file of TE sequences, virus sequences, etc., over to a reference database for building the index file, and then a read-mapping mismatch number is passed onto the "align_-to_repeat_virus" script for the specific type of small RNA being mapped: "0" mismatches for *Wolbachia*, "1" mismatch for virus and structure RNA, and "2" mismatches for TEs.

General command structure for Steps 4, 5, 6, and 7:

```
step4_transposon.sh <sample_name>
step5_virus.sh <sample_name>
step6_srna_structure.sh <sample_name>
step7_wolbachia.sh <sample_name>
```

3.4 Saving Results to Sharing Folder

The counts and length distributions generated during various steps such as gene-centric and miRNA analysis are saved during each of the steps (Fig. 1). These scripts now send final results files generated in steps above to a consolidated results folder for sharing or posting.

In addition, the *wigToBigWig* script is used to convert the fixed step WIG file to the Bigwig file which can be visualized on the Broad Institute Integrative Genomics Viewer (IGV, [35]) together with the genome assembly and GTF files. Reads mapped to the intergenic regions are progressively clustered together if normalized read counts are over a 0.02 reads per million thresholds within a sliding window of 25 base. To reduce the redundancy in the genic table caused by different isoforms of a gene, the *mergeBed* program [36] is used to consolidate different isoforms by providing the genomic location of each isoform. The isoform with the highest read counts is chosen as the representative of the gene.

3.5 piRNA Ping-Pong and Phasing Patterns Analysis

In Step 8 reads are checked again with the Cutadapt program [20], and trimmed reads longer than 23 nt are aligned to the genome using Bowtie with no mismatch. The genomic location and the number of times of mapped reads are recorded. Using this information, autocorrelation analysis is executed to identify periodic peaks based on a previous script from [37]. For 5′-to-5′ phasing analysis, autocorrelation analyses of 5′-to-5′ distance on the same genomic strands are carried out [37]. For 3′-to-5′ phasing analysis, the autocorrelation analyses of 3′-to-5′ distance on the same genomic strands are carried out and Z-score at distance 0 is calculated. For 24–35 nt small RNA ping-pong pattern determinations, autocorrelation analyses of 5′-to-5′ distance on the opposite genomic strands are carried out and Z-score at distance 10 was calculated, noting Z-scores over 2 as significant. The 18–23 nt RNA (siRNA) duplex analysis is similar except that Z-score at distance 21 is calculated for reads 18–23 nt long.

Step 8: Phasing and ping-pong patterns analyses scripts within step8_phasing.sh:

```
sRNA_Phasing_pipeline.sh <sample.24_35.trim.fastq.uq.polyn>
<species>
siRNA_Phasing_pipeline.sh <sample.18_23.trim.fastq.uq.polyn>
<species>
```

4 Notes

A critical first step in the MSRG pipeline is to trim the 3′ adapter sequence in the small RNA libraries with the Cutadapt [20] tool, but different labs' libraries have distinct linker sequences that may

4.1 Noting the MSRG Input Parameter for the 3′ Adapter Sequence in the Small RNA Libraries.

not be obvious in the publications or data submissions. The incorrect adapter sequence input prevents the Bowtie aligner from aligning the untrimmed reads and this causes outputs to be largely missing the bulk of the small RNA patterns. For example, "TGGAATTCTC" is the correct input sequence from the Illumina TruSeq small RNA library construction kit while "AGATCGGAAG" is the correct input sequence for the NEBNext small RNA library construction kit. Our typical approach to determining the linker sequence from a downloaded library dataset is to apply the FastQC tool [38] and then inspecting the "Overrepresented sequences" section to look for a common string in these sequences likely representing the 3′ adapter sequence. Small RNA libraries that have already had linkers pre-trimmed of 3′ adapter sequences can also be valid input files for the MSRG pipeline regardless of the linker sequence input parameter, provided that those pre-trimmed RNA sequences map well to the reference genome.

4.2 Visualizing WIG/BigWig Plots on the Integrate Genome Viewer (IGV) Browser Requires Matched Set GTF File and Genome File

For the small RNA coverage plots, there are visualization files like WIG and BigWig files that can be viewed with the IGV browser from the Broad Institute [35]. However, the GTF and genome files used for each small RNA analysis run through this pipeline need to be matched to load properly in the IGV browser. For example, the matched genome fasta and GTF files from the MSRG database are included for download [12], because we have found that different updated versions of mosquito genomes and GTF files from VectorBase contained different coordinates that were inconsistent with the small RNA genomics outputs processed on the earlier genome and GTF files.

The utility of the WIG/BigWig coverage plots is for cross comparisons with the counts recorded in the gene-centric-consolidated tables and the intergenic count tables generate in the first gene-centric step. Typically, we set the Y-scale of the tracks on the auto-scale so that the IGV browser window dynamically adjusts the peaks being displayed, with one track each for Plus and Minus genomic strands. The Normalized counts files (*.norm) display all the small RNA reads coverage where read frequency is normalized against the reads number of mapping sites, whereas Unique counts files (*.unq) only display the read coverage for the reads with a single genomic mapping location, so the majority of transposon and satellite repeats small RNAs are absent from the Unique count files.

4.3 Why Are the Gene-Centric Script and Phasing Scripts So Slow?

These two scripts rely on BED tools [36] when mapping reads to gene functional annotations or genomic intervals for gene-centric and phasing, respectively. Because these commands operate at the speed limit for input-output writing commands to the data disk, the processor demands are low and scripts cannot be sped up with requesting more processors. Thus, it is practical to specify in the

job submission script to allow for at least 48 h and up to a few days for the process not to terminate from a time out.

Although the gene-centric pipeline is absolutely essential to run for all the other downstream steps, the phasing script can be considered optional for some users who may not be interested in the highly periodic phasing of small RNA biogenesis patterns in mosquitoes [12, 37]. This last phasing step can be commented out to save time in obtaining the final results, or the last phasing step can also be run as a stand-alone script using the original small RNA read library as an input. Optimizations of the phasing script are ongoing, and updates will appear in GitHub.

Acknowledgements

We thank the patience of the editors for pandemic-related extensions and Augustine Abaris from Boston University Research Computing Services for assistance with the setup of the singularity image. This study was also supported by Boston University School of Medicine Startup funds, the BU Genome Science Institute, the Wing-Tat Lee foundation, and the NIH grant fund R01-GM135215 to N.C.L.

References

1. Gamez S, Srivastav S, Akbari OS, Lau NC (2020) Diverse defenses: a perspective comparing dipteran Piwi-piRNA pathways. Cell 9(10): 2180. https://doi.org/10.3390/cells9102180
2. Leinonen R, Sugawara H, Shumway M (2011) International nucleotide sequence database C. the sequence read archive. Nucleic Acids Res 39(Database issue):D19–D21. https://doi.org/10.1093/nar/gkq1019
3. Brackney DE, Scott JC, Sagawa F, Woodward JE, Miller NA, Schilkey FD et al (2010) C6/36 Aedes albopictus cells have a dysfunctional antiviral RNA interference response. PLoS Negl Trop Dis 4(10):e856. https://doi.org/10.1371/journal.pntd.0000856
4. Adelman ZN, Anderson MA, Liu M, Zhang L, Myles KM (2012) Sindbis virus induces the production of a novel class of endogenous siRNAs in Aedes aegypti mosquitoes. Insect Mol Biol 21(3):357–368. https://doi.org/10.1111/j.1365-2583.2012.01141.x
5. Vodovar N, Bronkhorst AW, van Cleef KW, Miesen P, Blanc H, van Rij RP et al (2012) Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. PLoS One 7(1):e30861. https://doi.org/10.1371/journal.pone.0030861
6. Schnettler E, Donald CL, Human S, Watson M, Siu RWC, McFarlane M et al (2013) Knockdown of piRNA pathway proteins results in enhanced Semliki Forest virus production in mosquito cells. J Gen Virol 94 (Pt 7):1680–1689. https://doi.org/10.1099/vir.0.053850-0
7. Liu Y, Zhou Y, Wu J, Zheng P, Li Y, Zheng X et al (2015) The expression profile of Aedes albopictus miRNAs is altered by dengue virus serotype-2 infection. Cell Biosci 5:16. https://doi.org/10.1186/s13578-015-0009-y
8. Miesen P, Girardi E, van Rij RP (2015) Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. Nucleic Acids Res 43(13): 6545–6556. https://doi.org/10.1093/nar/gkv590
9. Saldana MA, Etebari K, Hart CE, Widen SG, Wood TG, Thangamani S et al (2017) Zika virus alters the microRNA expression profile and elicits an RNAi response in Aedes aegypti mosquitoes. PLoS Negl Trop Dis 11(7): e0005760. https://doi.org/10.1371/journal.pntd.0005760
10. Varjak M, Maringer K, Watson M, Sreenu VB, Fredericks AC, Pondeville E et al (2017) Aedes aegypti Piwi4 is a noncanonical PIWI protein

involved in antiviral responses. mSphere 2(3): e00144-17. https://doi.org/10.1128/mSphere.00144-17

11. Halbach R, Miesen P, Joosten J, Taşköprü E, Rondeel I, Pennings B et al (2020) A satellite repeat-derived piRNA controls embryonic development of Aedes. Nature 580(7802): 274–277. https://doi.org/10.1038/s41586-020-2159-2
12. Ma Q, Srivastav SP, Gamez S, Dayama G, Feitosa-Suntheimer F, Patterson EI et al (2021) A mosquito small RNA genomics resource reveals dynamic evolution and host responses to viruses and transposons. Genome Res 31(3):512–528. https://doi.org/10.1101/gr.265157.120
13. Chakraborty M, Ramaiah A, Adolfi A, Halas P, Kaduskar B, Ngo LT et al (2021) Hidden genomic features of an invasive malaria vector, Anopheles stephensi, revealed by a chromosome-level genome assembly. BMC Biol 19(1):28. https://doi.org/10.1186/s12915-021-00963-z
14. Main BJ, Marcantonio M, Johnston JS, Rasgon JL, Brown CT, Barker CM (2021) Whole-genome assembly of Culex tarsalis. G3 (Bethesda) 11(2):jkaa063. https://doi.org/10.1093/g3journal/jkaa063
15. Chirn GW, Rahman R, Sytnikova YA, Matts JA, Zeng M, Gerlach D et al (2015) Conserved piRNA expression from a distinct set of piRNA cluster loci in eutherian mammals. PLoS Genet 11(11):e1005652. https://doi.org/10.1371/journal.pgen.1005652
16. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS et al (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39(Database issue): D876–D882. https://doi.org/10.1093/nar/gkq963
17. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J et al (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res 41(Database issue):D70–D82. https://doi.org/10.1093/nar/gks1265
18. Giraldo-Calderon GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P et al (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res 43(Database issue): D707–D713. https://doi.org/10.1093/nar/gku1117
19. Langmead B (2010) Aligning short sequencing reads with bowtie. Curr Protoc Bioinformatics. Chapter 11:Unit 11 7. https://doi.org/10.1002/0471250953.bi1107s32
20. Chen C, Khaleel SS, Huang H, Wu CH (2014) Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med 9:8. https://doi.org/10.1186/1751-0473-9-8
21. Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res 47(D1): D155–DD62. https://doi.org/10.1093/nar/gky1141
22. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9(4):357–359. https://doi.org/10.1038/nmeth.1923
23. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP et al (2006) Characterization of the piRNA complex from rat testes. Science 313(5785): 363–367
24. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C et al (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 127(6):1193–1207
25. Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442(7099):199–202
26. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R et al (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 128(6):1089–1103
27. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6(11):e1001191. https://doi.org/10.1371/journal.pgen.1001191
28. Parrish NF, Fujino K, Shiromoto Y, Iwasaki YW, Ha H, Xing J et al (2015) piRNAs derived from ancient viral processed pseudogenes as transgenerational sequence-specific immune memory in mammals. RNA 21(10): 1691–1703. https://doi.org/10.1261/rna.052092.115
29. Suzuki Y, Frangeul L, Dickson LB, Blanc H, Verdier Y, Vinh J et al (2017) Uncovering the repertoire of endogenous Flaviviral elements in Aedes Mosquito genomes. J Virol 91(15): e00571-17. https://doi.org/10.1128/JVI.00571-17
30. Tassetto M, Kunitomi M, Whitfield ZJ, Dolan PT, Sanchez-Vargas I, Garcia-Knight M et al (2019) Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. eLife 8:e41244. https://doi.org/10.7554/eLife.41244

31. Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9(5):411–412.; ; author reply 4. https://doi.org/10.1038/nrg2165-c1
32. Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S et al (2009) A broadly conserved pathway generates 3'UTR-directed primary piRNAs. Curr Biol 19(24): 2066–2076. https://doi.org/10.1016/j.cub.2009.11.064
33. Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42(Database issue):D68–D73. https://doi.org/10.1093/nar/gkt1181
34. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S et al (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 26(4):407–415. https://doi.org/10.1038/nbt1394
35. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G et al (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26. https://doi.org/10.1038/nbt.1754
36. Quinlan AR (2014) BEDTools: the Swiss-Army tool for genome feature analysis. Curr Protoc Bioinformatics 47:11 2 1–11 2 34. https://doi.org/10.1002/0471250953.bi1112s47
37. Gainetdinov I, Colpan C, Arif A, Cecchini K, Zamore PD (2018) A single mechanism of biogenesis, initiated and directed by PIWI proteins, explains piRNA production in Most animals. Mol Cell 71(5):775–90 e5. https://doi.org/10.1016/j.molcel.2018.08.007
38. Andrews S. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/2010

Chapter 21

Preparation of Non-overlapping Transposable Elements (TEs) Annotation by Interval Tree

Shohei Kojima

Abstract

Transposable elements (TEs) are a major source of PIWI-interacting RNAs (piRNAs), therefore properly assigning piRNA library sequencing reads to the TEs from which they were derived is important for accurate assessment of piRNA biology. When calculating the abundance of small RNA-seq reads mapping to various TEs, a non-overlapping TE annotation is preferable because reads mapping to more than one genomic feature will often be excluded when counting reads. However, most unmodified TE annotations contain some degree of overlap between TE features. Here, I outline the principle and provide all scripts needed to resolve such overlapping regions of TE annotations to a single best TE annotation leveraging a computationally efficient tree algorithm. Non-overlapping annotations generated by this method can be directly used in commonly used read counting software.

Key words Transposon annotation, Mobile genetic elements, Interval tree, RepeatMasker, Repetitive DNA

1 Introduction

piRNAs processed from TEs mediate important functions across a wide variety of animals. In arthropods, somatic piRNAs silence TEs, controlling metabolic homeostasis and neuronal heterogeneity [1–3]. In mammals, about half of pre-pachytene piRNAs derive from TEs and are central in silencing TEs during spermatogenesis [4]. Furthermore, it is known that different classes of TEs give rise to piRNAs present in fetal and postnatal gonads [5]. To investigate the genomic source of piRNAs, precise assignment of piRNA reads onto TE annotations is required. Calculation of TE-derived piRNA expression from small RNA-seq is conceptually similar to counting any NGS reads mapping to TEs, a well-acknowledged challenge, albeit further limited by the length of piRNAs [6]. Proper assignment of NGS reads onto non-overlapping TE features is thus very

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_21,

important for accurate measurement of TE-derived piRNA expression.

Annotation of TEs in reference genomes is the first step in identifying sources of TE-derived piRNAs. RepeatMasker is de facto standard to annotate repetitive elements, including TEs, at nucleotide resolution [7]. RepeatMasker can wrap multiple search engines, such as sequence alignment-based similarity search method (RMBlast) or methods using probabilistic model (hmmer [8]). Those programs calculate and report similarity scores (Smith–Waterman score or bit score, respectively) between stretches of genome sequence and repeat consensus sequences or models. It is possible that the same genome region may have high similarity to more than one repeat. In such cases, RepeatMasker reports multiple annotations for the same region. It is also possible that the borders of repeats (i.e., the start and end positions of the repeat) may be ambiguous, making two adjacent repeats partially overlap each other. Reads mapping to more than one overlapping TE annotations can be excluded during read counting. For example, featureCounts [9] and htseq-count [10], the two commonly used programs to count reads on genomic features, exclude such reads by default. To solve this problem, a non-overlapping TE annotation consisting of the single best TE annotation based on similarity score should be used [11].

Searching for overlaps between genomic intervals can be slow when the dataset is large, thus a scalable and efficient approach is desirable. An interval tree is a data structure frequently used to identify overlap of genomic intervals [12]. It is an extension of a binary search tree. In a binary search tree, an internal node can have two child nodes and can be conceptualized as follows: If one assumes there is a sorted array of numbers and inserts one number in the array, where the number has been inserted can be found by iterating a cycle in which the array is split in half, and a judgment is made regarding which half the number has fallen into. In the case of an interval tree, nodes hold positions of intervals (e.g., start and end positions of repeat annotation). Because of the binary nature of the data structure, the time needed to search for a specific interval intersecting with another interval is shorter than pairwise comparison. The tree can also be constructed such that it satisfies "heap property"; that is, each node has a key (e.g., a value such as similarity score) and the key in each internal node is equal or greater than all the keys in the node's subtrees. Trees of such properties are able to be efficiently searched for nodes having the maximum key intersecting with a certain interval.

2 Materials

2.1 Computer

One advantage of this method is that it can be accomplished quickly on a personal computer, including Linux, Mac, or Windows machines. The script uses only single thread. In the case of mouse genome (mm10), it requires less than 100 MB RAM, and takes 3 min on Linux with Intel Core i8 using either RMBlast or hmmer3 repeat masks.

2.2 Computational Environment

The following software and files should be installed or collected prior to beginning:

1. Python 3.6 or later.
2. Python built-in modules, os, sys, gzip, datetime, collections, argparse, errno.
3. Repeat-masked genome of the organism of interest. Pre-masked genomes are available (http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html) and users can generate their own custom-masked genome.
4. Script to generate non-overlapping repeat annotation: https://github.com/shohei-kojima/collapse_RepeatMasker_annotation

3 Method

3.1 Preparation of Repeat Annotation

RepeatMasker will use either RMBlast or hmmer3, depending on the repeat library provided (*see* **Note 4.1**). RepeatMasker outputs repeat annotation in "[file_name].out" file. In this file, similarity score (Smith–Waterman score or bit score) will be reported with genome interval, strand, and repeat class of each annotation. Those scores reflect similarity of repeat to consensus repeat sequence. Some repeat annotations partially or completely overlap with one or more neighboring ones, due to similarity to more than one repeat sequence present in the library. RepeatMasker can be run after the configuration of the search engine according to the detailed software instructions provided elsewhere (repeatmasker.org).

3.2 Remove Overlapping Regions

3.2.1 Principle

Conceptually, removing overlapping regions of a repeat annotation involves selecting the annotation with the highest likelihood of reflecting the actual nature of the genomic range in question, which is evaluated using similarity score. Computationally, this selection can be accomplished by building interval trees satisfying max-heap property (Fig. 1). The first step to building such a tree is to sort the repeat annotation based on start coordinates (Fig. 1a). When there are overlapping annotations, the script described in

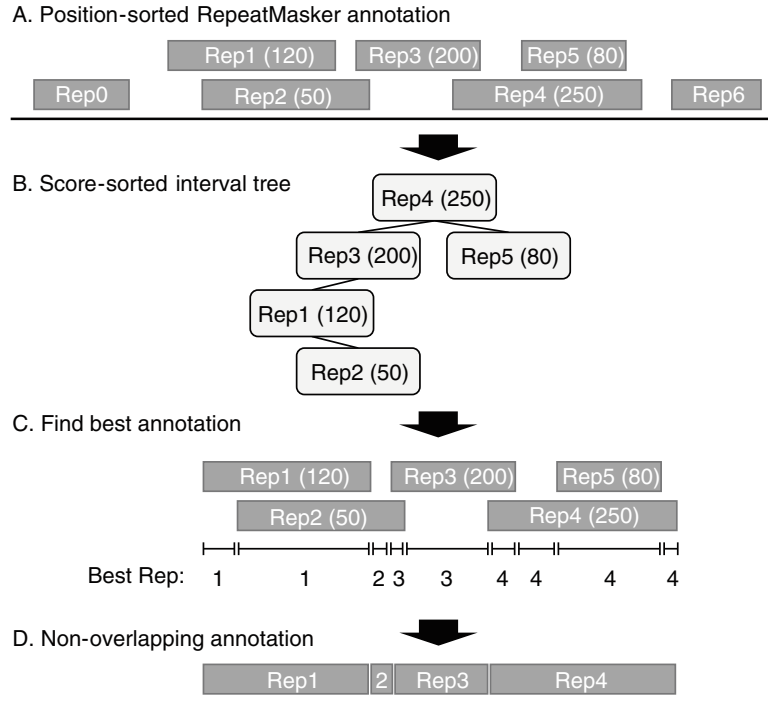


Fig. 1 Generation of non-overlapping repeat annotation by interval tree. (**a**) Repeat annotation from RepeatMasker will be sorted by start coordinates. Numbers in parenthesis show the similarity scores of annotations (either bit score or Smith–Waterman score). (**b**) It will generate interval tree when there are overlapping repeat annotations. The similarity scores are used as keys to order nodes to build a heap. (**c**) It will find annotation with best similarity score for all intervals. (**d**) It will connect intervals with the same annotation and report non-overlapping annotation

Subheading 2.2 makes an interval tree comprising nodes that store the repeat information (Fig. 1b). When constructing this tree, similarity scores will be assigned as the keys to build a heap. After constructing the interval tree, the script will find the repeat annotations with the highest score for all intervals using the property of max-heap (Fig. 1c). Finally, it will merge intervals annotated as the same repeat and report a non-overlapping annotation.

3.2.2 Usage

The script can be downloaded from GitHub. Here is an example using "git clone" command to clone the repository.

```
$ git clone https://github.com/shohei-kojima/collapse_Repeat-
Masker_annotation
```

The command above will create a directory named "collapse_RepeatMasker_annotation." In this directory, the script "collapse_RM_annotation.py" can be found.

Next, change the current working directory to "collapse_RepeatMasker_annotation" by "cd" command, and type "python collapse_RM_annotation.py -h" to see a help message. If the help message is output, the script is working.

```
$ cd collapse_RepeatMasker_annotation
$ python collapse_RM_annotation.py -h # issue help message
```

To generate a non-overlapping repeat annotation, specify the input file ("[file_name].out" file from RepeatMasker) with the "-i" flag, and the output file basename with the "-o" flag. These two flags are always required to use main functions.

```
$ python collapse_RM_annotation.py -i genome.fa.out -o genome.
fa.out.collapsed
```

3.3 Output Files

The script outputs two files below:

1. [output_basename].gtf.gz
2. [output_basename].bed.gz.

These two files are gzipped-compressed files. To decompress those files, use "gzip -d" command.

```
$ gzip -d file.gtf.gz
```

- GTF file

The GTF file generated by the script can be directly used in software which count reads mapping to genome features (e.g., featureCounts, htseq-count). Each annotation in the GTF file consists of three features, one gene, one transcript, and one exon. The "gene_id" in the attribution files are formatted as shown in Fig. 2. Four notations are separated with a dot; unique ID, name of the repeat (the 10th column of the input ".out" file), repeat family/class (the 11th column of the input ".out" file), and the original ID (the 15th column of the input ".out" file). The "transcript_id" and the "exon_id" in the attribution files are formatted as "t_gene_id" and "e_gene_id," respectively. In the attribution field, the similarity score (the 1st column of the input ".out" file) is also recorded as "bit_score."

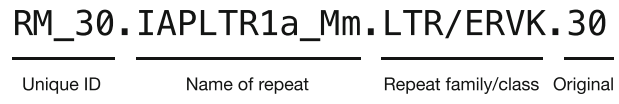


Fig. 2 Format of gene_id in the output GTF file. The gene_id in the GTF file generated by the provided script consists of four identifying features: unique ID, name of repeat, repeat family or class, and the original ID

- BED file

The BED file generated by the script has 6 columns. The 4th column records the same "gene_id" used in the GTF file. The 5th column records the similarity score (the 1st column of the input ".out" file). The 6th column is the orientation of TE annotation.

3.4 Options

This script has several options that affect results.

- "keep_simple_repeat" option

  By default, the script does not output annotations that are assigned as "Simple_repeat" and "Low_complexity" in the 11th column in the input ".out" file. With the -keep_simple_repeat option, it will also output those annotations. Example code is below.

```
$ python collapse_RM_annotation.py \
 -i genome.fa.out -o genome.fa.out.collapsed \
 -keep_simple_repeat
```

- "min" option

  By default, this script outputs all TE annotations regardless of length (i.e., "-min 1"). The minimal length of TE annotations that will be output can be set by specifying the length as an integer with this option. Below is the example of specifying minimal length 50.

```
$ python collapse_RM_annotation.py \
 -i genome.fa.out -o genome.fa.out.collapsed \
 -min 50
```

- "gap" option

  Occasionally, a TE copy that is recognizably (i.e., by manual inspection) derived from a single insertion is annotated as multiple fragments by RepeatMasker. To address such cases, this script can connect TE fragments if these are annotated as the same element (i.e., same repeat name) and are closer to each other than the specified gap length. The gap length can be specified by the "-gap [integer]" option. By default, if the same repeats are nested to each other without a gap, the provided script connects those fragments and reports using as a single unique annotation (i.e., "-gap 0"). Below is the example of specifying gap length 5.

```
$ python collapse_RM_annotation.py \
 -i genome.fa.out -o genome.fa.out.collapsed \
 -gap 5
```

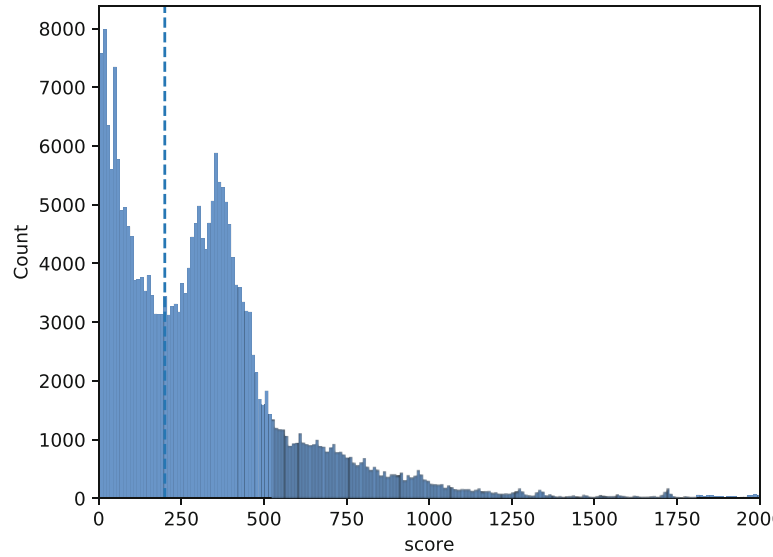


Fig. 3 Distribution of the bit scores of LTR/ERVK in mouse genome. The mouse genome, mm10, was masked by RepeatMasker with the Dfam repeat library (Dfam3.2). The distribution of the bit scores for LTR/ERVK annotations longer than 50 bp is visualized as a histogram. An example arbitrary threshold, 200, is shown as a vertical dotted line

3.5 Further Filtering

TE annotation generated by "collapse_RM_annotation.py" should be further filtered when only TEs having high similarity to repeats in the library should be used (e.g., Ref. 11). An arbitrary threshold for filtering can be determined by manually inspecting the distribution of the similarity scores of the TE annotations of your interest (Fig. 3). An example script for visualization can be found in the GitHub repository (https://github.com/shohei-kojima/collapse_RepeatMasker_annotation/plot_histogram.py).

4 Notes

4.1 Repeat Library

Choice of repeat library plays a major role in the quality of TE annotation, especially for non-model organisms. RepeatMasker can take two different types of libraries, including ones from Repbase and Dfam. Dfam is an open database and current default library for RepeatMasker, while Repbase is currently only available with a paid subscription. Although Dfam covers a smaller number of organisms than Repbase, the number is growing (552 species in July 2021). When the species of one's interest is not found in Dfam, the repeat should be supplemented from Repbase. If the species is also not available in Repbase, a custom repeat library can be constructed by identifying de novo repeats, for example using RepeatModeler which wraps multiple programs that discover different types of repeats [3, 13].

Acknowledgments

The author thanks Nicholas F. Parrish for comments and extensive editing of this manuscript.

References

1. Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M et al (2013) Transposition-driven genomic heterogeneity in the drosophila brain. Science 340:91–95. https://doi.org/10.1126/science.1231965
2. Jones BC, Wood JG, Chang C, Tam AD, Franklin MJ, Siegel ER et al (2016) A somatic piRNA pathway in the *Drosophila* fat body ensures metabolic homeostasis and normal lifespan. Nat Commun 7:13856. https://doi.org/10.1038/ncomms13856
3. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA et al (2018) Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol 2:174–181. https://doi.org/10.1038/s41559-017-0403-4
4. Ernst C, Odom DT, Kutter C (2017) The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. Nat Commun 8:1411. https://doi.org/10.1038/s41467-017-01049-7
5. Gainetdinov I, Skvortsova Y, Kondratieva S, Funikov S, Azhikina T (2017) Two modes of targeting transposable elements by piRNA pathway in human testis. RNA 23:1614–1625. https://doi.org/10.1261/rna.060939.117
6. Lanciano S, Cristofari G (2020) Measuring and interpreting transposable element expression. Nat Rev Genet 21:721–736. https://doi.org/10.1038/s41576-020-0251-y
7. Tarailo-Graovac M, Chen N (2009;Chapter 4: Unit 4.10) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. https://doi.org/10.1002/0471250953.bi0410s25
8. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD (2018) HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. https://doi.org/10.1093/nar/gky448
9. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656
10. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169. https://doi.org/10.1093/bioinformatics/btu638
11. Sakashita A, Maezawa S, Takahashi K, Alavattam KG, Yukawa M, Hu Y-C et al (2020) Endogenous retroviruses drive species-specific germline transcriptomes in mammals. Nat Struct Mol Biol 27:967–977. https://doi.org/10.1038/s41594-020-0487-4
12. Li H, Rong J (2021) Bedtk: finding interval overlap with implicit interval tree. Bioinformatics 37:1315–1316. https://doi.org/10.1093/bioinformatics/btaa827
13. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451–9457. https://doi.org/10.1073/pnas.1921046117

Chapter 22

Statistical Thermodynamics Approach for Intracellular Phase Separation

Tomohiro Yamazaki and Tetsuya Yamamoto

Abstract

Phase separation is one of the fundamental processes to compartmentalize biomolecules in living cells. RNA–protein complexes (RNPs) often scaffold biomolecular condensates formed through phase separation. We here present a statistical thermodynamics approach to investigate intracellular phase separation. We first present the statistical thermodynamic theory of the liquid-liquid phase separation (LLPS) of two molecules (such as proteins and solvent molecules) and of a polymer solution (such as RNPs and solvent molecules). Condensates produced by LLPS show coarsening and/or coalescence to minimize their total surface area. In addition to the LLPS, there are other types of self-assembly, such as microphase separation, micellization, emulsification, and vesiculation, with which the growth of the assembly stops with optimal size and shape. We also describe a scaling theory of micelles of block copolymers, where their structures are analogous to the core-shell structure of paraspeckle nuclear bodies scaffolded by RNPs of NEAT1_2 long noncoding RNAs (lncRNAs) and RNA-binding proteins (RBPs). These theories treat the self-assembly of polymers in the thermodynamic equilibrium, where their concentrations and compositions do not change with time. In contrast, RNPs are produced according to the transcription of RNAs and are degraded with time. We therefore take into account the dynamical aspect of the production of RNPs in an extension of the theory of the self-assembly of soft matter. Finally, we discuss the structure of paraspeckles as an example to demonstrate that an approach combining experiment and theory is powerful to investigate the mechanism of intracellular phase separation.

Key words Architectural RNA, Biomolecular condensates, Flory–Huggins theory, Liquid–liquid phase separation (LLPS), Macroscopic phase separation, Micellization, Microphase separation, NEAT1_2 lncRNA, Polymer physics, Soft matter physics

1 Introduction

1.1 Phase Separation and Biomolecular Condensates

Accumulating evidence suggests that various cellular structures, such as a variety of nuclear and cytoplasmic structures, nuclear pore complex, and pre-autophagosomal structure, form through phase separation, a physical process that separates dense and dilute phases [1–3]. This process creates membraneless compartments, which are often called biomolecular condensates, concentrating biomolecules in crowded cellular circumstances. The cell nucleus

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_22,

is rich in subnuclear condensates, including nucleoli, nuclear speckle, paraspeckle, Cajal body, Gem, perinucleolar compartment, nuclear stress body, and histone locus body [1, 4]. These condensates contain specific sets of proteins and RNAs and act as reaction crucibles, molecular sponges, and genomic hubs [5]. Thus, it would be important to understand mechanisms for the formation and function of the intracellular biomolecular condensates through phase separation. In addition, as pathological aggregates are also formed through phase separation, it would be crucial to understand the mechanism of their formation and control the phase separation/aggregation in the treatment of the diseases [6, 7].

1.2 RNA and Phase Separation

RNAs play critical roles in the formation and regulation of biomolecular condensates [8]. Although RNAs inhibit aggregations of a set of RBPs by their nonspecific interactions [9], RNAs can induce phase separation by specifically scaffolding RBPs via their RNAs sequences and structures. It has been proposed that RNP condensates form via phase separation when the summation of RNA–RNA, RNA–protein, and protein–protein interactions exceeds a certain threshold [10]. Various RNAs, including lncRNAs, can be essential scaffolds of a subset of biomolecular condensates with their partner RBPs [8, 11–14]. As such condensates are found in various species from yeast to humans, this scaffolding function is thought to be one of the fundamental functions of RNAs. We thus termed such scaffolding RNAs architectural RNAs (arcRNAs) [8, 11]. The arcRNAs induce phase separation by increasing local concentrations of RBPs typically containing multimerization domains, such as intrinsically disordered regions and low-complexity domains, via their RNA sequences and structures that these RBPs interact [8, 11]. Especially, nuclear condensates scaffolded by arcRNAs are constructed as the arcRNAs are transcribed. Therefore, transcription dynamics govern the behaviors of the condensates.

In Subheading 3, we introduce the statistical thermodynamics of phase separation and self-assembly, which are widely used in soft matter physics. The concepts presented in the Subheading 3 are written any standard textbooks of soft matter physics [15, 16] and polymer physics [17–19], but we reduce them so that they are more accessible to experimental biologists, who are interested in the physics of phase separation. Readers can also take into account the features of the system of their interest in an extension of these theories. Soft matter physics usually treats the phase separation and self-assembly of polymer systems in which the concentration and composition of polymers do not change. In our recent works, we take into account the transcription dynamics, which change the concentration and composition in the system, in an extension of the phase separation [15] and self-assembly of RNA [20, 21]. In Subheading 3.6, we show an example of paraspeckles to show that

an approach combining experiments and theory is powerful to study the mechanism of the assembly of nuclear condensates [20].

2 Materials

Computer programs for numerical calculations can be written by any languages. Mathematica® or Maple® are useful because one has to take many derivatives in the calculations of phase separation. An example of mathematica code is given in the Supplementary Materials.

3 Methods

3.1 Basic Principle of Phase Separation

LLPS is understood as the competition between the interactions between molecules and thermal fluctuations. In this section, we outline the statistical thermodynamics of phase separation of a two-component system, such as the phase separation of proteins in an aqueous solution. This theory predicts the condition (the temperature and the concentration of proteins) with which the phase separation happens and the concentration of proteins in each of the coexisting phases. We extend this theory to outline the Flory–Huggins theory, which is the standard theory of the phase separation in a polymer solution, in Subheading 3.2 and the scaling theory of micelles of block copolymer in Subheading 3.4. Many nuclear condensates are scaffolded by arcRNA. The roles played by arcRNA can be understood on the basis of the Flory–Huggins theory. Some nuclear condensates, such as paraspeckles, form the core-shell structure, which is analogous to block copolymer micelles. We recently showed that this analogy is very powerful to understand the mechanisms of the assembly of paraspeckles [20].

In the statistical thermodynamics, the lattice model is commonly used to treat phase separation, *see* Fig. 1a [15, 16]. This model represents the system as a lattice, where each lattice site is occupied by one molecule. In this section, we treat a system composed of two types of molecules, A and B. For the case of an aqueous solution of proteins, A molecules are proteins and B molecules are water molecules (as you will see later, one can instead think of proteins as B molecules and of water molecules as A molecules). Molecules change their positions (the lattice sites that they occupy) with time t with the following conditions: not more than one molecule occupies one lattice site (due to the excluded volume interactions between molecules) and all the lattice sites are occupied by molecules (because the system is filled with molecules); the property of the system which mixing does not change the volume is called incompressibility. The number of each molecular species is constant. The probability of molecules

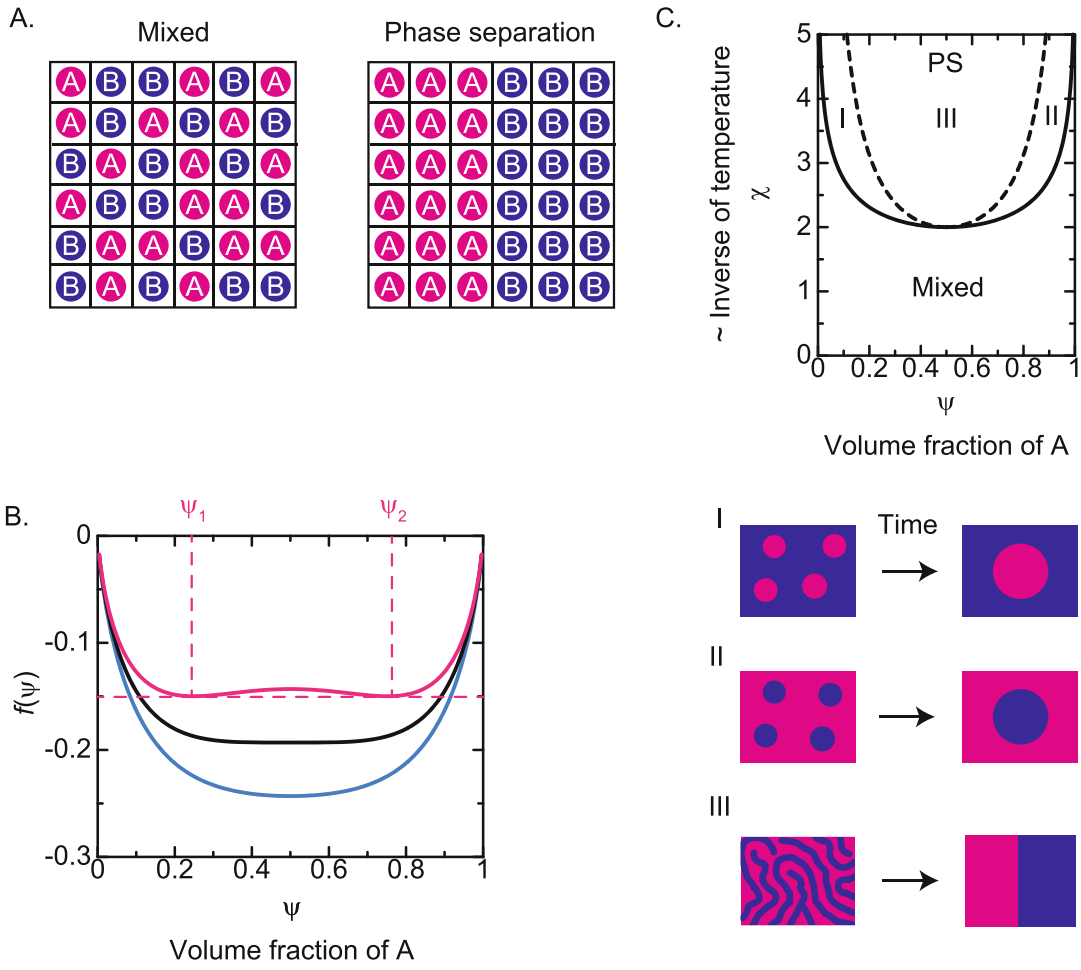


Fig. 1 The statistical thermodynamics of phase separation. (**a**) The lattice model of a mixture of A and B molecules. With this model, each lattice site is occupied by either of A (magenta) or B (cyan) molecule. The two molecules at the nearest neighbors can exchange the positions due to the thermal fluctuation and the probability of the exchange depends on the interactions between these molecules. The magnitudes of the interactions between these molecules are represented by the interaction parameter χ, *see* Eq. (4). The system can be completely mixed (left) or show two coexisting phases (right). (**b**). The free energy per lattice site is shown as a function of the volume fraction of A molecules for $\chi = 1.8$ (cyan), 2.0 (black), and 2.2 (magenta). There is only a single minimum for $\chi > 2$, whereas there are two minima at $\psi = \psi_1$ and ψ_2 for $\chi > 2$. The phase separation happens in the range $\psi_1 < \psi < \psi_2$ because the free energy becomes smaller when the volume fraction of one phase is $\psi = \psi_1$ and the volume fraction of the other phase is $\psi = \psi_2$: the volume fraction of the two phases does not depend on the fraction ψ of A molecules in the entire system as long as it is in the range of $\psi_1 < \psi < \psi_2$. (**c**) The miscibility phase diagram with respect to the interaction parameter χ and the volume fraction ψ of A molecules in the system. The binodal curve is shown by the solid line and the spinodal curve is shown by the broken line. In the thermodynamic equilibrium, the phase separation happens in the region delineated by the binodal curve. The mixed state is still metastable in the region between the binodal curve and the spinodal curve, whereas this state is unstable in the region delineated by the spinodal curve

to move to the neighboring sites depends on the interactions with neighboring molecules and the magnitudes of the thermal fluctuation. We ask the structure of this system (whether A and B are completely mixed or show phase separation) at the thermodynamic equilibrium ($t \to \infty$). There are several approaches to derive the answer to this question. For example, Monte Carlo simulation directly switches the positions of molecules with the above rule. We here use the mean field approximation to address the physical insight into the phase separation.

The second law of thermodynamics predicts that the stable structure of a thermodynamic system is determined by the minimum of the free energy. The (Helmholtz) free energy F has the form

$$F = E - TS. \tag{1}$$

The first term in the right side E represents the interaction energy, which is due to the interactions between molecules in the system. The second term in the right side of Eq. (1) S is the entropy of the system, which represents the thermal fluctuation of the system. T is the absolute temperature. Equation (1) represents the intuitive fact: the thermal fluctuation is not significant for low temperature and the interaction energy is not significant for high temperature. For the case of low temperature, the structure of the system is determined to minimize the total of the interaction energy between molecules in the system: when the attractive interactions between the same molecular species are larger than those between different molecular species (which is usually the case), the system shows phase separation. For the case of high temperature, the structure of the system is determined to maximize the entropy: the thermal fluctuation tends to make the molecules randomly mixed. This may be analogous to the fact that your room becomes messy (your things are "randomly mixed") as time elapses if you do not constantly take care to order the objects therein.

In a uniform system, the volume fraction ψ of molecule A does not depend on position. The volume fraction of molecule B is $1 - \psi$ because each lattice site is occupied either by A or B. The probability that you find an A molecule at a given site is ψ. The free energy F is derived as a function of the volume fraction ψ. We derive the volume fraction ψ that makes the free energy minimum.

The mathematical derivation of the free energy is given in, for example, the book by Safran (Subheading 1.4 in [15]) and the book by Doi (Subheading 2.3 in [16]). We instead show the form of the free energy and investigate its property. The interaction energy has the form

$$\frac{E}{M} = \frac{1}{2} z J_{\mathrm{AA}} \psi^2 + \frac{1}{2} z J_{\mathrm{BB}} (1-\psi)^2 + z J_{\mathrm{AB}} \psi (1-\psi), \tag{2}$$

where J_{AA}, J_{BB}, J_{AB} are the magnitudes of A–A interaction (the interaction between two A molecules at adjacent sites), B–B interactions (the interaction between two B molecules at adjacent sites), and A–B interactions (the interaction between A and B molecules at adjacent sites), respectively. M in the denominator of the left side is the number of lattice sites that compose the system and is proportional to the volume of the system. The A–A interaction is attractive for $J_{\mathrm{AA}} < 0$ (the free energy decreases when two A molecules are at the adjacent sites) and repulsive for $J_{\mathrm{AA}} > 0$ (the free energy increases when two A molecules are at the adjacent sites). We here do not ask the type of intermolecular interactions between molecules (such as electrostatic interactions or van der waals interactions), but J_{AA}, J_{BB}, and J_{AB} can be calculated if the details of the intermolecular interactions are known (*see*, for example, [15]). z is the number of adjacent sites in the lattice.

To understand, Eq. (2), we first count the interaction energy experienced by a molecule at an arbitrary lattice site (say, the m-th lattice site). The first term in the right side is designed such that the energy increases by J_{AA} when this lattice site and its adjacent site are both occupied by A molecules, where it happens with the probability ψ^2 (because the probability with which an arbitrary site is occupied by A molecule is ψ, the probability with which two sites are both occupied by A molecules is ψ^2). The energy increases by J_{BB} when this site and its adjacent sites are both occupied by B molecules, where it happens with the probability $(1-\psi)^2$ (see the second term in the right side of Eq. (2)). The energy increases by J_{AB} when this site and its adjacent site are occupied by one A molecule and one B molecule, where it happens with the probability $2\psi(1-\psi)$; 2 accounts for the fact that A molecule can be at the m-th lattice site and B molecule at the adjacent site or vice versa (see the third term of the right side of Eq. (2)). The molecule at a given lattice site interacts with z molecules because this site is adjacent to z lattice sites (this is why the right side is multiplied by z). To count the interaction energy of all the M lattice site is derived by multiplying the interaction energy of a molecule in an arbitrary lattice by M. $1/2$ in the right side of Eq. (2) accounts for the fact that each interaction is count twice.

Eq. (2) is rewritten in the form

$$\frac{E}{M} = k_{\mathrm{B}} T \chi \psi (1-\psi) + \frac{1}{2} z J_{\mathrm{AA}} \psi + \frac{1}{2} z J_{\mathrm{BB}} (1-\psi), \tag{3}$$

with

$$\chi = \frac{z}{k_{\mathrm{B}} T} \left(J_{\mathrm{AB}} - \frac{J_{\mathrm{AA}} + J_{\mathrm{BB}}}{2} \right). \tag{4}$$

Indeed, the second and third terms of Eq. (3) do not affect the system because the total number ψM of molecule A and the total number $(1-\psi)M$ of molecule B are both fixed. This implies that only the net interaction (represented by χ), the difference between A–B interaction and the average of A–A and B–B interaction, is relevant to phase separation. For cases in which the magnitudes of the attractive interactions between the same molecular species (namely, A–A and B–B interactions) are larger than the magnitudes of the attractive interactions between different molecular species (namely, A–B interactions), the interaction parameter χ is positive. The energy E becomes minimum when $\psi = 0$ or $\psi = 1$ for cases in which χ is positive. This represents the simple fact that A molecules tend to aggregate and B molecules aggregate when the magnitudes of the attractive interactions between the same molecular species are larger than the magnitude of the attractive interactions between different molecular species. Cases in which χ is negative represent a net repulsion between the same molecular species (it is the case when the molecules have net electric charges). The repulsion can lead to ordering and we do not treat such cases here. $k_{\mathrm{B}}T$ is the thermal energy ($k_{\mathrm{B}} = 1.38 \times 10^{-23}\ J/K$ and T is the absolute temperature). The parameter χ therefore represents the magnitude of the net interaction energy, relative to the thermal energy. The parameter χ is called the (Flory) interaction parameter or simply χ-parameter.

The entropy of mixing S is derived as

$$\frac{S}{M} = -k_{\mathrm{B}}(\psi \log \psi + (1-\psi)\log(1-\psi)), \tag{5}$$

where the way to derive Eq. (5) by using the Boltzmann principle is shown in **Note 1**. As we see in the discussion below Eq. (3), the configuration with the minimum energy is $\psi = 0$ (all of the molecules are B) or $\psi = 1$ (all of the molecules are A) for cases in which χ is positive. The free energy contribution $-TS$ of this entropy is minimum when $\psi = 1/2$ and thus deviate the free energy minimum from $\psi = 0$ and $\psi = 1$. This reflects the fact that the mixing entropy represents the thermal fluctuation that mixes A and B molecules.

By substituting Eqs. (3) and (5) into Eq. (1), the free energy density $f(\psi) = F/M$ (the free energy per site) of a uniform system is derived as

$$f(\psi) = k_{\mathrm{B}}T(\psi \log \psi + (1-\psi)\log(1-\psi) + \chi\psi(1-\psi)). \tag{6}$$

The structure of the system is determined by the minimum of the free energy. It should be noted that the number of molecules and the number of lattice sites (the volume of the system) are both constant and are given by the preparation condition of the system. By taking into account these conditions, the free energy of a phase separated system is constructed as

$$F = f(\psi_1)M_1 + f(\psi_2)M_2 - \mu(\psi_1 M_1 + \psi_2 M_2) + \Pi v_0(M_1 + M_2), \tag{7}$$

where ψ_1 and ψ_2 are the volume fractions of A molecules in phases 1 and 2 and M_1 and M_2 are the number of lattice sites that consist of phases 1 and 2. The first and second terms in Eq. (7) are the free energy of phases 1 and 2. The third and fourth terms are introduced to fix the total number of molecule A (this also fixes the number of molecule B) and the total number of lattice sites in the system, respectively. The trick here is that we determine the constants μ and Π so that the total number of molecule A and the total number of lattice sites in the system are fixed to given values (such mathematical device is called Lagrange multiplier). Physically, μ and Π correspond to the chemical potential and the osmotic pressure. v_0 is the volume per lattice site. Eq. (7) is now a function of the volume fractions, ψ_1 and ψ_2, and the number, M_1 and M_2, of lattice sites.

The derivatives of the free energy with respect to ψ_1, ψ_2, M_1, and M_2 are all zero at the minimum of the free energy F. This leads to the conditions

$$\mu = \frac{\partial f(\psi_1)}{\partial \psi_1} = \frac{\partial f(\psi_2)}{\partial \psi_2} \tag{8}$$

$$\Pi v_0 = \psi_1^2 \frac{\partial}{\partial \psi_1}\left(\frac{f(\psi_1)}{\psi_1}\right) = \psi_2^2 \frac{\partial}{\partial \psi_2}\left(\frac{f(\psi_2)}{\psi_2}\right) \tag{9}$$

Physically, Eqs. (8) and (9) represent the fact that the chemical potential μ and the osmotic pressure Π are equal in the two coexisting phases (this is the condition of thermodynamic equilibrium). Graphically, Eqs. (8) and (9) represent the fact that the free energy has common tangent at the volume fractions ψ_1 and ψ_2, see the magenta broken line in Fig. 1b; it is rather straightforward to see that the volume fractions ψ_1 and ψ_2 that give the common tangent are indeed the minimum of the free energy. In the window of volume fractions, $\psi_1 < \psi < \psi_2$, the free energy of mixed state is larger than the free energy at $\psi = \psi_1$ and $\psi = \psi_2$, *see* Fig. 2. This implies that if the total volume fraction ψ of molecule A in the system (which is the total number of (or the total volume occupied by) molecule A in the system divided by the total number M (or the total volume) of lattice sites in the system) is within the window, $\psi_1 < \psi < \psi_2$, the system shows phase separation, where the volume fraction of molecule A in one phase is ψ_1 and the volume fraction of molecule A in the other phase is ψ_2. Note that the volume fractions ψ_1 and ψ_2 of the two phases do not depend on the total number ψM of molecule A in the system: the total number ψM of molecule A only affects the size of domains of these phases.

The volume fractions, ψ_1 and ψ_2, are often plotted as functions of the interaction parameter χ as the solid line in Fig. 1c. The

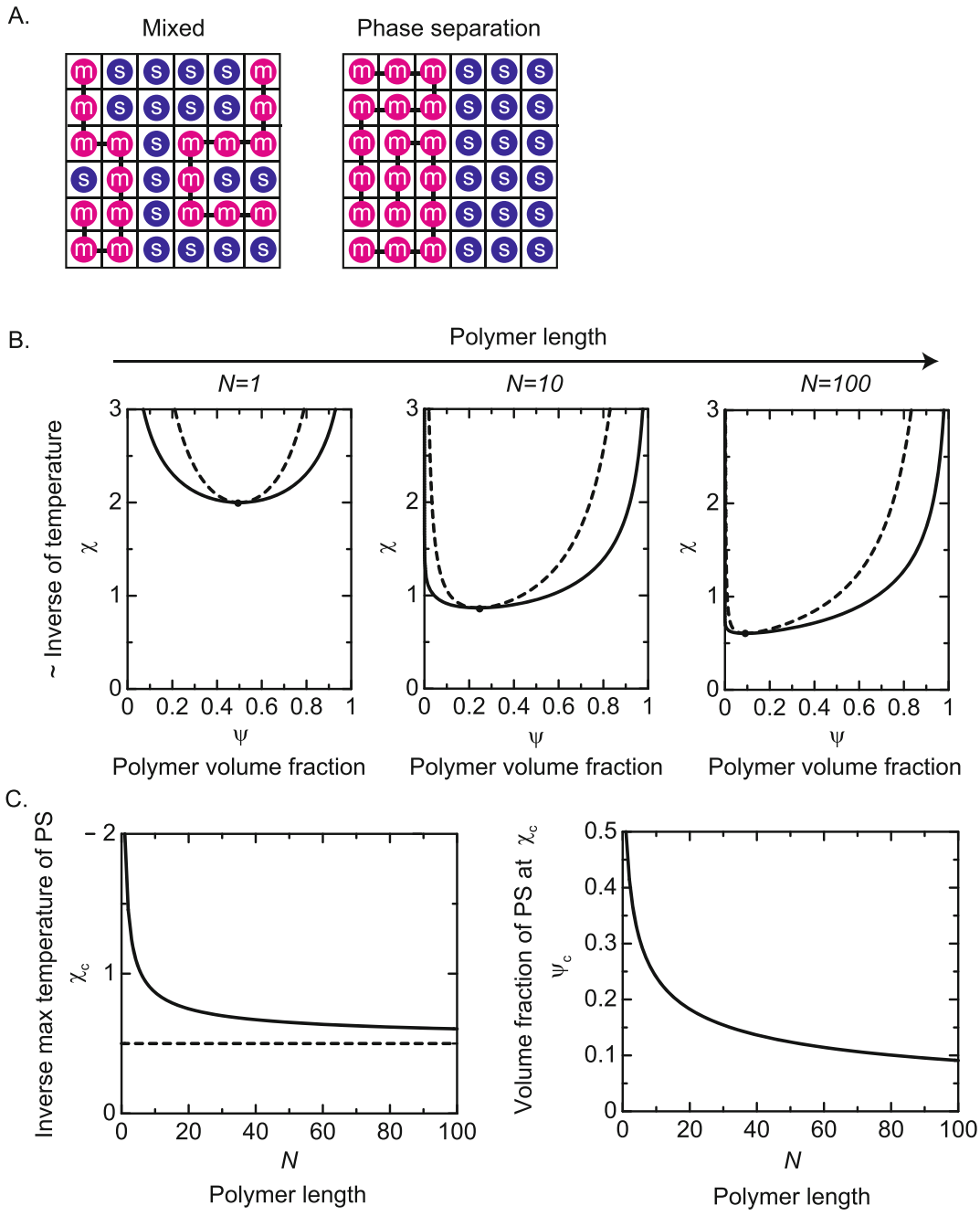


Fig. 2 Phase separation of polymer solution, Flory–Huggins theory. (**a**) The lattice model of a polymer solution. Each lattice site is occupied by either of a monomer (magenta) in a polymer or a solvent molecule (cyan). Monomers occupy consecutive lattice sites to represent the connectivity of monomers in polymers. (**b**) The phase diagram of polymer solutions with respect to the interaction parameter χ and the volume fraction ψ of polymers for $N = 1$ (left), 10 (middle), and 100 (right), as predicted by the Flory–Huggins theory. The binodal curve is shown by the solid line and the spinodal curve is shown by the broken line. (**c**) The interaction parameter χ_c (left) and the polymer volume fraction ψ_c (right) at the critical point are shown as functions of the number of monomers in a polymer, as predicted by the Flory–Huggins theory, *see* Eqs. (12) and (13)

interaction energy contributions, J_{AA}, J_{BB}, and J_{AB}, are usually fixed by the chemistry of A and B molecules (note that in some case, the magnitudes of interactions between some molecules can depend on the temperature due to their conformational changes or the changes of their solvation state). Remember that χ is the magnitude of net interaction energy divided by the thermal energy, where the latter is proportional to the absolute temperature T. Changing χ usually corresponds to the changes of the temperature. The volume fraction ψ of A molecules is proportional to the concentration of these molecules (you can change one to the other by using the density and molar mass). The diagram as Fig. 1c (which is called miscibility phase diagram) tells us the region of the total volume fraction ψ of molecule A and the interaction parameter χ in which the two-phase coexistent state is stable and the volume fractions, ψ_1 and ψ_2, of molecule A in the two coexisting phases.

The solid curve in Fig. 1c is called binodal curve. In the region between the solid line and the broken line in Fig. 1c, the mixed state is still metastable; the free energy of two-phase coexisting phases is smaller than the free energy of mixed state, but there is an energy barrier that has to be overcome by thermal fluctuations for the system to show phase separation. The formation of a droplet by overcoming the energy barrier is called nucleation and then the droplet grows as time elapses. This process is called nucleation and growth (*see* Subheading 3.3 for the physical mechanism that drives the growth of the droplet). When the second derivative of the free energy is negative, $\frac{\partial^2 f(\psi)}{\partial \psi^2} < 0$ (where the free energy is convex), the mixed state becomes unstable; there is no energy barrier and phase separation happen immediately. A mosaic pattern of two phases emerges instantaneously and shows time evolution toward the thermodynamically stable state. This process is called spinodal decomposition. The border between metastable and unstable regions is called spinodal curve and is given by

$$\frac{\partial^2 f(\psi)}{\partial \psi^2} = 0, \tag{10}$$

see the broken line in Fig. 1c. The binodal curve and spinodal curve intersect at the critical point, *see* Fig. 1c. The volume fraction of molecule A and the interaction parameter at the critical point are given by $\psi_c = 1/2$ and $\chi_c = 2$, respectively.

The above theory can be extended to systems of interest by taking into account features of the systems in the free energy as long as one treats the phase separation in the thermodynamic equilibrium. Indeed, one can even predict the surface tension by using an extension of this theory (*see* Chapter 2 in [15]).

3.2 Phase Separation in Polymer Solution

3.2.1 Flory–Huggins Theory

Polymers, such as DNA, RNA, and the complexes of these nucleic acids with proteins (e.g., chromatin, RNP), are rich in cells. A polymer is composed of many repeating units (monomers) that are connected in one dimension. Because of the connectivity, monomers in a polymer chain move together; this makes more difficult for polymers to mix by thermal fluctuations. The phase separation in a polymer solution can be treated by using the lattice model as Subheading 3.1, *see* Fig. 2a. Each lattice site of a polymer solution is occupied by a monomer in a polymer chain or a solvent molecule. Monomers adjacent along a polymer chain occupy the adjacent lattice sites; this represents the connectivity of polymers. We treat cases in which each polymer is composed of N monomers. To be explicit, we think of molecule A as a monomer and molecule B as a solvent molecule and describe the system by the volume fraction ψ of a monomer in a lattice site. The lattice model of a polymer solution is called Flory–Huggins theory and the mathematical derivation can be found in many textbooks of polymer physics [15–19].

As in the case of Subheading 3.1, we here find the minimum of the free energy to predict the phase separation in a polymer solution. The interaction energy has the same form as Eq. (3) (if one thinks of molecule A as a monomer and molecule B as a solvent molecule). The entropy of mixing of a polymer solution has the form

$$S = -k_{\mathrm{B}}\left(\frac{\psi}{N}\log\psi + (1-\psi)\log(1-\psi)\right). \tag{11}$$

Eq. (11) returns to Eq. (5) if one sets $N = 1$, which corresponds to cases in which monomers are not connected. This may be intuitively understood if one thinks the first term of Eq. (11) represents the entropy of mixing of polymers and the second term of Eq. (11) represents the entropy of mixing of solvent. The first term of Eq. (11) represents the fact that the center of mass of a polymer move along with all N monomers in the polymer. The Flory–Huggins theory uses the interaction energy of the form of Eq. (3). As it is shown in Subheading 3.1, this corresponds to the interaction energy when one chops off monomers and distributes them randomly in the lattice. This approximation may look too crude in the first glance, but is effective for a polymer solution, where polymers are interpenetrating each other (because in such case, one can forget about the fact that the probability of finding a monomer next to a monomer of the same polymer due to their connectivity). One can use Eq. (11), instead of Eq. (5), and follow the same argument as Subheading 3.1 to derive the phase diagram of a polymer solution.

Examples of the phase diagram of a polymer solution predicted by the Flory–Huggins theory are shown in Fig. 2b. This result

suggests that the phase separation of a polymer solution happens at a higher temperature (the temperature is the inverse of the interaction parameter, *see* Eq. (4)) and a lower monomer concentration than a mixture of monomers and solvent. It is because monomers in a polymer move together and this makes polymers more difficult to be mixed with solvent as the length of each polymer (which is count by the number N of monomers in each polymer) increases. The monomer volume fraction and the interaction parameter at the critical point also shift to

$$\psi_c = \frac{1}{1+\sqrt{N}} \tag{12}$$

$$\chi_c = \frac{\left(1+\sqrt{N}\right)^2}{2N}, \tag{13}$$

Eq. (12) predicts that the monomer volume fraction at the critical point decreases as the number N of monomers in a polymer increases, *see* Fig. 2c. The interaction parameter χ_c at the critical point decreases as the number N of monomers in a polymer increases (χ_c saturates to $1/2$ for $N \to \infty$). These reflect the tendency of polymers in a solution to show phase separation as the number N of monomers in a polymer increases.

The Flory–Huggins theory implies one feature of arcRNA. Even if the concentration of proteins is smaller than the concentration necessary to drive the phase separation, the phase separation can happen when these proteins bind to arcRNA: the bound proteins are effectively "connected" by arcRNA and this connectivity reduces the concentration that is necessary to drive phase separation. It is the universal role of the connection to phase separation.

3.2.2 *Excluded Volume Interactions Between Monomers in Polymers*

The repulsive interaction between hydrophilic portions of arcRNA plays an important role in the structure of paraspeckles, which are the nuclear bodies discussed in Subheading 3.6. The origin of the repulsive interaction can be found from the analysis of the Flory–Huggins theory. This section is important for deeper understanding of Eq. (22) in Subheadings 3.4 and 3.5 and thus can be skipped in the first reading.

In the limit of dilute solution ($\psi \ll 1$), the free energy (per lattice site) has the form

$$f(\psi) = \frac{\psi}{N}\log\psi + \frac{1}{2}a_2\psi^2 \tag{14}$$

with

$$a_2 = 1 - 2\chi, \tag{15}$$

where we omitted the term $(\chi - 1)\psi$ from Eq. (14) because in the free energy $f(\psi)M$, $M\psi$ is the total number of monomers in the solution and is a constant (*see* also the discussion below Eq. (3) in

Subheading 3.1). The parameter a_2 represents the effective interaction between monomers in a solvent and is called the second virial coefficient. The effective interaction is attractive for $a_2 < 0$ and is repulsive for $a_2 > 0$. This may be understood by the fact that the system favors smaller values of the free energy (the principle of minimum free energy, *see* the discussion of Eq. (1)); for cases in which $a_2 < 0$,the free energy decreases as the volume fraction ψ of polymers increases (namely, the system favors the condensation of polymers), in contrast, for cases in which $a_2 > 0$, the free energy increases as the volume fraction of polymer increases (namely, the system does not favor the condensation of polymers). We note that the first term, 1, in the right side of Eq. (15) results from the entropy of mixing; even when the net interaction between monomers is attractive, $\chi > 0$ (which is indeed usually the case unless monomers have electric charges), the interaction between monomers can be effectively repulsive when the attractive interaction is smaller than the mixing tendency of monomers and solvent molecules due to the thermal fluctuation. Indeed, phase separation does not happen for cases in which $a_2 > 0$, *see* $\chi_c > 1/2$ from Eq. (13).

The interactions between polymers make a significant contribution to the osmotic pressure of a polymer solution. The osmotic pressure of a polymer solution is calculated by using the thermodynamic relationship, $\Pi v_0 = \psi^2 \frac{\partial}{\partial \psi}\left(\frac{f(\psi)}{\psi}\right)$, see also Eq. (9). It is instructive to see the osmotic pressure of a polymer solution in the dilute limit, $\psi \ll 1$,

$$\frac{\Pi v_0}{k_B T} = \frac{\psi}{N} + \frac{1}{2} a_2 \psi^2 \tag{16}$$

The first term in the right side of Eq. (15) represents so-called van't Hoff law (which one may remember that the osmotic pressure has a similar law to the equation of state of ideal gas). This term results from the thermal fluctuation of polymers in a solvent. This term is usually dominant for the cases of low molecular-weight molecules, $N = 1$, (or monomers). However, Eq. (13) predicts that this term is greatly diminished for the polymers, where usually N is very large. The second term in the right side of Eq. (15) represents the contribution of the interactions between monomers to the osmotic pressure. The second virial coefficient a_2 therefore can be characterized by measuring the osmotic pressure of polymer solutions. The osmotic pressure is positive for cases in which the interactions between molecules are repulsive, which corresponds to $a_2 > 0$, because the repulsive interactions prevent the condensation of polymers.

3.3 Phase Separation and Self-Assembly

In Subheadings 3.1 and 3.2, we discussed the condition with which the phase separation happens in multicomponent systems,

including polymer solutions, and the volume fraction of components in each of the coexisting phases. The shape and size of condensates produced by the phase separation, which are one of the features of biomolecular condensates, is another feature that can be characterized experimentally. The type of phase separation discussed in Subheadings. 3.1 and 3.2 are called *macroscopic* phase separation, which is the case of LLPS. The properties and dynamics of condensates assembled by such phase separation are governed by the interfacial tension γ, which is indeed equal to the interfacial energy per unit interfacial area. The free energy (per lattice site) is minimum at the volume fractions, ψ_1 and ψ_2, of molecules in the condensate and the nucleoplasm at far away from the surfaces, *see* Fig. 1b and also Subheading 3.1. The volume fraction at the vicinity of the interfaces is between ψ_1 and ψ_2, and thus the free energy (per lattice site) at these regions is larger than the minimum of the free energy, *see* Fig. 1b. The free energy of the system (which is the sum of the free energy over all the lattice sites) thus increases as the total area A of the interface increases:

$$F_{\mathrm{sur}} = \gamma A. \tag{17}$$

The shape with which the interfacial area of a condensate becomes minimum (with a given volume) is sphere. The condensates produced by macroscopic phase separation are therefore spherical (the interface may show thermal fluctuations and the magnitudes of the fluctuations is determined by the ratio between the thermal energy $k_{\mathrm{B}}T$ and the interfacial tension γ), *see* Fig. 3a. The total area of interfaces in the system decreases by the growth of condensates. The condensates produced by macroscopic phase separation therefore grow as long as molecules are available. There are two modes of the growth of condensates: coarsening (Ostwald ripening) and coalescence, see Fig. 3b. These phenomena may be observed by live imaging.

An AB diblock copolymer is a block of A monomers connected to a block of B monomers, see also the schematic figure in Fig. 4a. For cases in which these two blocks are immiscible, diblock copolymers in a melt or a solution form ordered structures, such as sphere (spherical domains of A monomers form a periodic lattice in the matrix of B monomers), cylinder (cylindrical domains of A monomers form periodic lattice in the matrix of B monomers), and lamellar (the layer structure of A and B monomers), *see* Fig. 3c. The pattern, the size of domains, and the distance between domains are mostly determined by the fraction of A monomers in each copolymer. This phase separation with which domains have optimal size and shape is called *microphase separation*.

There are other types of self-assembly, where assemblies have optimal size and shape, such as *micellization*, *emulsification*, *and vesiculation*, *see* Fig. 3d. Such self-assembly is observed in an

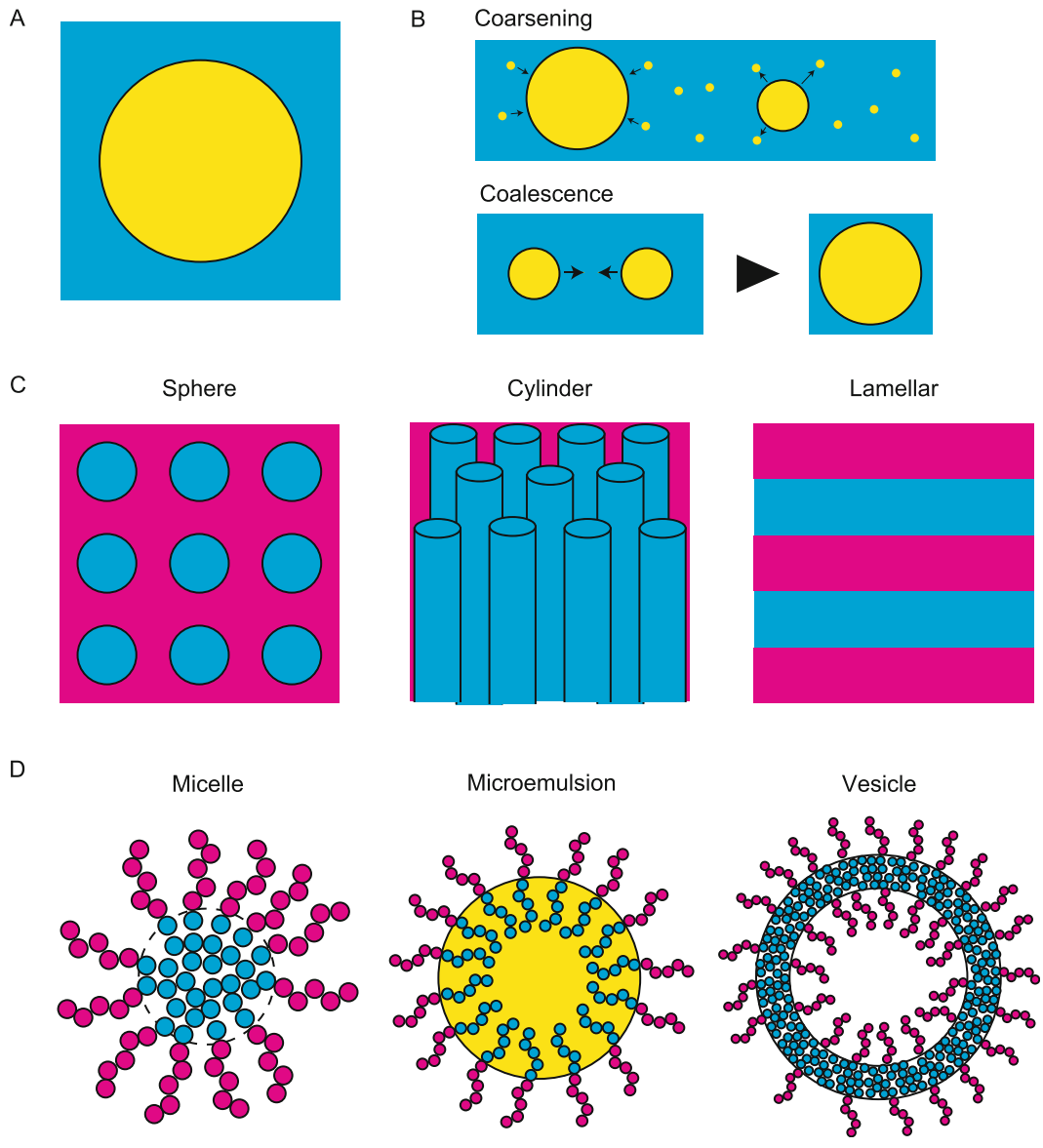


Fig. 3 Phase separation and self-assembly. (**a**) A condensate (a domain) formed by phase separation. It is spherical to minimize the interface. (**b**) The size of condensates increases by coarsening (RNA complexes tend to dissociate from smaller condensates and to associate with larger condensates) or coalescence (condensates fuse when they collide each other). (**c**) Microphase separation of block copolymers in melts or solutions. (**d**) Block copolymers show micellization, emulsification, and vesiculation in solutions

aqueous solution of amphiphiles that have hydrophilic heads and hydrophobic tails and AB block copolymers in a selective solvent (solvent is selective when it dissolves one block of the copolymer, but does not dissolve the other block). The hydrophilic heads have affinity to water, but repel from oil. In contrast, hydrophobic tails have affinity to oil, but repel from water. Amphiphiles in an aqueous

solvent are therefore localized at the water surface to avoid their hydrophobic tails from water (and thus form monomers). If the concentration of amphiphiles is larger than so-called critical micelle concentration (often abbreviated as CMC), amphiphiles form micelles, with which hydrophobic tails of amphiphiles form the micelle core to avoid the hydrophobic tails from water and the hydrophilic heads form the shell. Indeed, AB block copolymers in a selective solvent assemble similar structures. Micelles can be sphere, cylinder, or lamellar, depending on the packing of the hydrophobic tails in the core and the packing of the hydrophilic heads in the shell. One may think that micellization and microphase separation are very similar. However, microphase separation is a discontinuous transition (the system is uniform above the transition temperature and forms microphases below this temperature), while micellization is continuous (the number of amphiphiles that assemble micelles increases continuously with increasing the concentration of amphiphiles through CMC).

3.4 Micellization of Block Copolymers

Condensates produced by LLPS are disordered liquid. Some nuclear condensates, such as paraspeckles, form the characteristic core-shell structure, which is analogous to micelles of block copolymers. There are several approaches to treat micelles of block copolymers, such as the self-consistent field theory and the phase field theory. We here use a simple scaling theory to address the physical insight in micellization [22, 23]. In the following, we show only the theory of spherical micelles. An extension to this theory to cylindrical micelles and lamellars is shown in, for example Ref. 23. This theory predicts the factors involved in the growth of micelles and the factors that limit the growth of micelles. It also predicts the number of block copolymers associated to each micelle as a function of the concentration of block copolymers and the length of each polymer block (which is quantified by the number of monomers in each polymer block).

3.4.1 Free Energy of a Block Copolymer Micelle

We treat a spherical micelle of AB block copolymers, where A block is composed of N_A monomers and B block is composed of N_B monomers, see the magenta and cyan beads in Fig. 4a. We assume that the length of both monomers is b (the volume of a monomer is b^3). B monomers are hydrophobic and are packed in the core of the micelle. The radius of the core r_c is thus given by

$$\frac{4\pi}{3} r_c^3 = n b^3 N_B. \tag{18}$$

The free energy F_n of a micelle composed of n block copolymers has the form

$$F_n = F_{cr} + F_{sur} + F_{shl}, \tag{19}$$

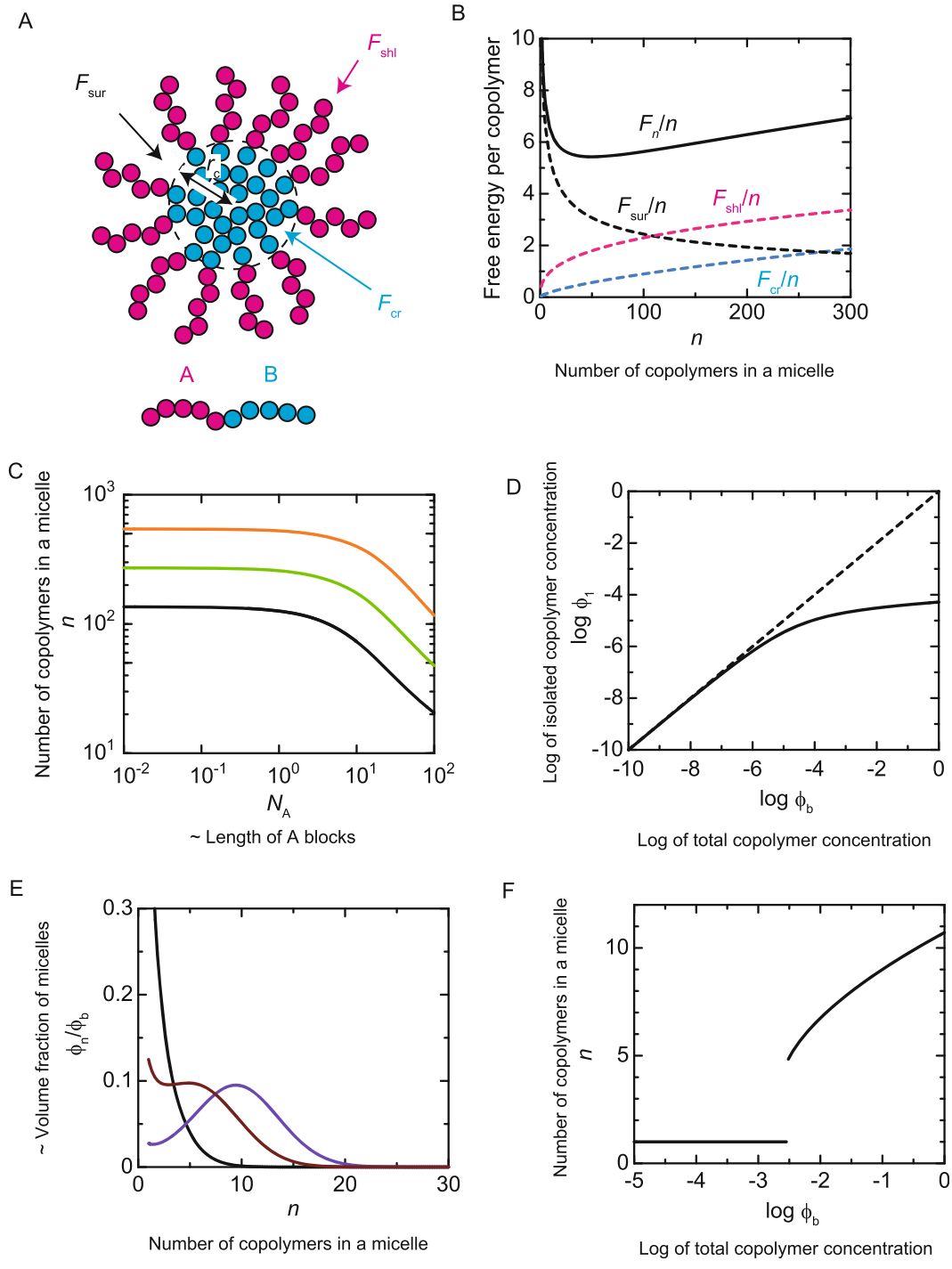


Fig. 4 Micellization of AB block copolymers in a solution. (**a**) The free energy of a micelle composed of AB block copolymers (A and B monomers are shown by magenta and cyan beads, respectively) is composed of the free energy F_{cr} of the core, the surface free energy F_{sur}, and the free energy F_{shl} of the shell. B. (**b**) The free energy contributions per block copolymer shown as functions of the number n of block copolymers in the micelle. The free energy of the core, the surface free energy, and the free energy of the shell are shown by the cyan, black, and magenta broken lines, respectively. The total of the three free energy contributions is shown

where F_{cr} is the free energy of the core, F_{sur} is the free energy at the surface of the micelle core, and F_{shl} is the free energy of the shell, *see* Fig. 4a.

The free energy of the core has the form

$$\frac{F_{cr}}{k_B T} = \frac{3}{2}\lambda_s n \frac{r_c^2}{N_B b^2}, \tag{20}$$

where $\lambda_s (=\pi^2/40)$ is a numerical constant due to the distribution of the free ends of B blocks. Eq. (20) is valid only for spherical micelles (*see* Ref. [23] for the other cases). A polymer behaves as a spring and stores energy when it is stretched. The free energy of a polymer has the same form of the elastic free energy of a spring with spring constant $3k_B T/Nb^2$ (N is the number of monomers in the polymer), *see* **Note 2** for a more detailed description. Eq. (20) represents the fact that B blocks are stretched with increasing the radius of the core; one end of a B block is at the surface of the core because the A block of the copolymer is outside of the core and the other end is at the vicinity of the core so that the core is filled with B monomers. The free energy of the core is thus the stretching free energy of B blocks. The stretching free energy F_{cr} of B blocks is proportional to $n^{5/3}/N_B^{1/3}$ (substitute the core radius r_c, given by Eq. (18), into Eq. (20)). This free energy contribution becomes less significant as the number N_B of monomers in B blocks increases; while it is rather difficult to extend a short string, it is relatively easier to extend a long crumpled string.

The surface free energy has the form

$$F_{sur} = 4\pi r_c^2 \gamma, \tag{21}$$

see also Eq. (17). Equation (21) represents the fact that B monomers at the surface of the core touches water or A monomers. The surface free energy is proportional to $N_B^{2/3} n^{2/3}$ (substitute the core radius r_c, given by Eq. (18), into Eq. (21)). This free energy contribution F_{sur} becomes more significant as the number N_B of monomers in B blocks increases.

The free energy of the shell has the form

Fig. 4 (continued) by the black solid line. (**c**) The number n of block copolymers in a micelle, with which the free energy per block copolymer becomes minimum, shown as functions of the number N_A of monomers in A block for $N_B = 40$ (black), 80 (light green), and 160 (orange). (**d**) The volume fraction ϕ_1 of isolated block copolymers shown as a function of the total volume fraction ϕ_b of block copolymers (solid line). The dashed lines are cases in which all block copolymers are isolated, $\phi_1 = \phi_b$. (**e**) The volume fraction ϕ_n of micelles composed of n block copolymers shown as functions of n for $log\ \phi_b = -5.17$ (black), -4.60 (brown), and -4.36 (purple). (**f**) The number n of block copolymers in the micelles with the maximum volume fraction, ϕ_n, shown as functions of the total volume fraction ϕ_b of block copolymers. We used $N_A = 20$ and $N_B = 40$ for the calculations of **b** and **d**–**f** and used $\gamma b^2/(k_B T) = 0.2$, $v/b^3 = 1.0$, and $C_s = 1.5$ for all the calculations (**b**–**f**)

$$\frac{F_{\mathrm{shl}}}{k_{\mathrm{B}}T} = \frac{3}{5}\frac{n^{3/2}}{(4\pi)^{1/2}} C_s \log\left(1 + \frac{5}{3}\frac{h}{r_{\mathrm{c}}}\right) \tag{22}$$

with

$$h = N_{\mathrm{A}} b \left(\frac{n}{4\pi r_{\mathrm{c}}^2}\frac{v}{b}\right)^{1/3}. \tag{23}$$

The constant C_{s} is a numerical constant that cannot be determined by the scaling theory. Experiments suggest $C_{\mathrm{s}} \simeq 1.38$ [23]. h is a length scale that returns to the height of A blocks for $h \ll r_{\mathrm{c}}$. v ($\equiv b^3 a_2$) is the excluded volume that represents the interactions between A monomers and is proportional to the second virial coefficient of A block, see the discussion below Eq. (15). Eq. (22) takes into account both the excluded volume interactions between A monomers and the stretching free energy of A blocks. Eq. (22) is valid for cases in which the excluded volume interactions between A monomers is repulsive and relatively large, $a_2 > 0$. The excluded volume interactions result from the mixing tendency between A monomers and solvent, *see* the discussion below Eq. (15). Solvent of such tendency is called a *good* solvent. For cases in which the excluded volume interactions between A monomers are small ($a_2 \approx 0$), the free energy has different form, *see* **Note 3**. Solvent of such tendency is called *θ-solvent* when the (two-body) excluded volume interactions between A monomers, represented by a_2, are dominated by the three-body excluded volume interactions and is called *marginal* solvent when the two-body excluded volume interactions are still dominant. In the following, we use Eq. (22).

The surface free energy F_{sur}/n per block copolymer decreases with increasing the number n of block copolymers in a micelle, see the black broken line in Fig. 4b. This implies that when other free energy contributions, F_{shl} and F_{cr}, are absent, as in the cases of macroscopic phase separation, the size of the micelle increases until all the block copolymers in the system are associated with the micelle (because the free energy becomes minimum at the thermodynamic equilibrium, *see* the discussion below Eq. (1)). The free energy contributions, F_{shl}/n and F_{cr}/n, of the shell and the core per block copolymer both increase with increasing the number of block copolymers in a micelle, see the magenta and cyan broke lines in Fig. 4b. There is a minimum of the free energy F_n/n per block copolymer, see the solid line in Fig. 4b. This implies that the surface free energy increases the number of block copolymers in a micelle and the free energy of the core and the shell limits the number of block copolymers in a micelle.

The number n of block copolymers in a micelle, with which the free energy per block copolymer is minimum, does not change much with increasing the number N_{A} of monomers in A blocks for small values of N_{A}, *see* Fig. 4c. It is because the number of block

copolymers in a micelle is mostly limited by the stretching free energy of B blocks in the core, which does not depend on N_A, *see* Eq. (20). One end of a B block is at the surface of the core (because it is connected with the A block) and the other end is at the proximity of the center to fill the micelle core with B monomers; the B blocks in the core are thus stretched to the radius r_c of the micelle core, which increases as the number of block copolymers increases, *see* Eq. (18). Quantitatively, the destabilization of the system by overstretching the B blocks is represented by the increase of the stretching free energy, *see* the cyan broken line in Fig. 4b; this free energy thus limits the number of block copolymers to suppress the overstretching of the B blocks. The number of block copolymers in a micelle decreases with increasing the number N_A of monomers for large values of N_A, *see* Fig. 4c. It is because the number of block copolymers in a micelle is mostly limited by the free energy of the shell. The surface density, $n/\left(4\pi r_c^2\right)$, of A blocks in the shell increases with increasing the number of block copolymers in a micelle and thus A monomers that show repulsive excluded volume interactions are confined in smaller space. Quantitatively, the destabilization of the system by increasing the density of repulsive monomers is represented by the increase of the free energy of the shell, see the magenta broken line in Fig. 4b; this free energy thus limits the number of block copolymers in a micelle to suppress overconcentration of A monomers. The asymptotic solutions are shown in **Note 4**.

3.4.2 Free Energy of a Block Copolymer Solution

In general, micelles in a solution are not composed of the same number of block copolymers. The number of block copolymers in micelles depends on the concentration of block copolymers in the solution. One therefore has to analyze the free energy of the entire solution, not just one micelle. The free energy of a block copolymer solution has the form

$$\frac{F_{sol}}{k_B T} = \sum_{n=1}^{\infty} \left[\frac{\phi_n}{Nn} \log \frac{\phi_n}{Nn} + \frac{F_n}{k_B T} \frac{\phi_n}{Nn} + \frac{\phi_n}{Nn} \log n! - \frac{\mu}{k_B T} \frac{\phi_n}{N} \right], \tag{24}$$

where ϕ_n is the volume fraction of micelles composed of n block copolymers. $(\phi_n/(Nn))$ is proportional to the number of micelles composed of n block copolymers) [15]. The first term represents the fact that n block copolymers move together when a micelle composed of n block copolymers move (it is the same logic as the first term of Eq. (14)). The second term is the free energy F_n of a micelle, *see* Eq. (19). The third term is related to the fact that each block copolymer in a micelle has the same probability to be dissociated from the micelle; the probability with which either of the block copolymers is dissociated is n times this probability (this term is not essential, but we added to be precise). The fourth term is introduced to fix the number of block copolymers in the system

constant by fine tuning μ, *see* in the same spirits as Eq. (7). Equation (24) does not take into account the interactions between micelles.

The volume fraction ϕ_n of micelles composed of n block copolymers at the minimum of the free energy, Eq. (24), (with respect to ϕ_n) has the form

$$\log \frac{\phi_n}{Nn} + 1 + \log n! + \frac{F_n}{k_B T} = \frac{n\mu}{k_B T}. \tag{25}$$

Equation (25) is valid for any values of n; the constant μ is represented by the volume fraction ϕ_1 of isolated block copolymers ($n = 1$)

$$\frac{\mu}{k_B T} = \log \frac{\phi_1}{N} + 1 + \frac{F_1}{k_B T}. \tag{26}$$

Combining Eqs. (25) and (26) leads to the form

$$\frac{\phi_n}{nN} = e^{-F_n/(k_B T)}. \tag{27}$$

with

$$F_n = F_n - nF_1 + k_B T \log n! - nk_B T \log \frac{\phi_1}{N}. \tag{28}$$

The volume fraction ϕ_1 of isolated block copolymers is determined by the condition

$$\phi_b = \sum_{n=1}^{\infty} \phi_n, \tag{29}$$

where ϕ_b is the total volume fraction of block copolymers in the solution. The critical micelle concentration (CMC) is usually defined by the copolymer volume fraction ϕ_b^{cmc}, with which $\phi_b^{cmc} = 2\phi_1$ - half of the copolymers composes micelles.

For cases in which the volume fraction ϕ_b of block copolymers is very small, block copolymers do not form micelles; the volume fraction ϕ_1 of isolated block copolymers is equal to the total volume fraction ϕ_b of block copolymers, see the solid and broken lines in Fig. 4d. The volume fraction $\phi_b - \phi_1$ of micelles increases with increasing the total volume fraction ϕ_b of block copolymers. The volume fraction ϕ_n of micelles composed of n block copolymers is a monotonically decreasing function of n when the total volume fraction ϕ_b of block copolymers is small, see the black line in Fig. 4e. In contrast, there is a maximum of the volume fraction ϕ_n at finite value of n when the total volume fraction ϕ_b of block copolymers becomes large enough, see the purple line in Fig. 4e. The number n of block copolymers in micelles of the maximum volume fraction increases with increasing the total volume fraction ϕ_b of block copolymers, *see* Fig. 4f.

3.5 Summary and Extension to Intracellular Phase Separation Research

Here we show theories of (macroscopic) phase separation and micellization. There are a couple of important concepts, which may be useful for intracellular phase separation researches:

- The system shows phase separation when the thermal energy (that mixes molecules) becomes smaller than the interaction energy (that attracts molecules of the same type).
- Polymers tend to show phase separation due to the connectivity of monomers in polymers (e.g., DNA, RNA).

In a dilute limit, the contribution of the mixing entropy is effectively taken into account in the interaction energy. The mixing tendency between polymers and solvent molecules give rise to effective repulsive interactions between polymers. The "interactions" in polymer physics include the contribution from the mixing entropy.

- The size of condensates produced by macroscopic phase separation (e.g., LLPS) increases to minimize the total area of their surfaces and these condensates are spherical. Amphiphiles and block copolymers show microphase separation and self-assembly, with which the condensates have the optimal size and shape.
- The number of block copolymers in a micelle increases due to the surface free energy and it is limited by the stretching free energy of blocks in the core and the repulsive excluded volume interactions between blocks in the shell.
- Most of block copolymers do not form micelles unless the concentration of block copolymers in the solution is large enough. The number of block copolymers in a micelle also depends on the concentration of block copolymers in the solution.

These theories treat the phase separation and the micellization at the thermodynamic equilibrium—for example, one asks whether an aqueous solution of polymers shows phase separation if one wait for a long enough time. In such problems, one assumes that the concentration, composition, and length of polymers do not change with time. This is not the case of arcRNA that composes nuclear condensates; arcRNA transcripts are synthesized continuously at the transcription site and they can degrade gradually while they diffuse in the condensate and the external solution. Recently, such dynamical feature of arcRNA has been taken into account in an extension of the Flory–Huggins theory [24].

Paraspeckle is one of the well-studied biomolecular condensates with RNA scaffolds. Paraspeckles are scaffolded by the RNPs consisting of NEAT1_2 lncRNA and RNA-binding proteins that have potentials to undergo phase separation via multimerization [25–29]. Transcription of the NEAT1_2 lncRNAs triggers

formation of the paraspeckle at the *NEAT1* gene loci with the interacting partner RBPs [29, 30]. The paraspeckles have characteristic internal core-shell architectures and form cylindrical as well as spherical shapes, which are distinct from typical condensates formed through LLPS [31, 32]. Within the paraspeckles, the 5′ and 3′ terminal regions of NEAT1_2 are localized in the shell and the middle region of the NEAT1_2 is localized in the core. Such structural features are analogous to micelles of amphipathic ABC triblock copolymers in water, where the A and C blocks are hydrophilic and the B block is hydrophobic. In the NEAT1_2 mutants, with which part of the terminal regions are deleted, the terminal regions are distributed between the core and the shell or, in some cases, are localized in the core [20] (*see* Subheading 3.6). The diameter of wild-type paraspeckles is smaller than the diameter of paraspeckles in deletion mutants (*see* Subheading 3.6). This implies that paraspeckles are assembled by the attractive interactions between proteins bound to the middle blocks of NEAT1_2, such as NONO and FUS [24], and the number of NEAT1_2 in a paraspeckle is limited by the stretching free energy of middle blocks and the repulsive excluded volume interactions between terminal blocks, which are probably due to the mixing tendency of terminal blocks and water molecules. One can therefore treat the assembly of paraspeckles in an extension of the theory of triblock copolymer micelles.

We have constructed the triblock copolymer model of paraspeckles by taking into account the following points [20, 21]: (1) The terminal blocks localized in the core suppress the attractive interactions between the middle blocks in the core. (2) NEAT1_2 is folded because its two terminal blocks are localized in the shell while the middle blocks are packed in the core. The localization of terminal blocks unfolds NEAT1_2 and partially releases the stretching of the middle blocks. (3) There is a contribution of thermal fluctuations that distribute terminal regions equally to the shell and the core. (4) Paraspeckles are probably assembled by the interactions between nascent NEAT1_2 transcripts, the production of which is in progress. Our theory takes into account the fact that nascent NEAT1_2 transcripts therefore can assemble paraspeckles without translational entropy cost, *c.f.* the first term of Eq. (24).

3.6 Example of Experimental Validations of Theoretical Predictions

We here present our recent study of the paraspeckle nuclear body as an example of how the theoretical model is useful for experimental investigations of the RNA-driven phase separation process. In the experiments, we established several NEAT1_2 mutant cell lines by using CRISPR/Cas9 and observed fine structures of the paraspeckles by electron microscope and super-resolution microscope [20] (*see* **Note 5**). The summary of the experimental results consistent with predictions of the block copolymer micelle model is as follows.

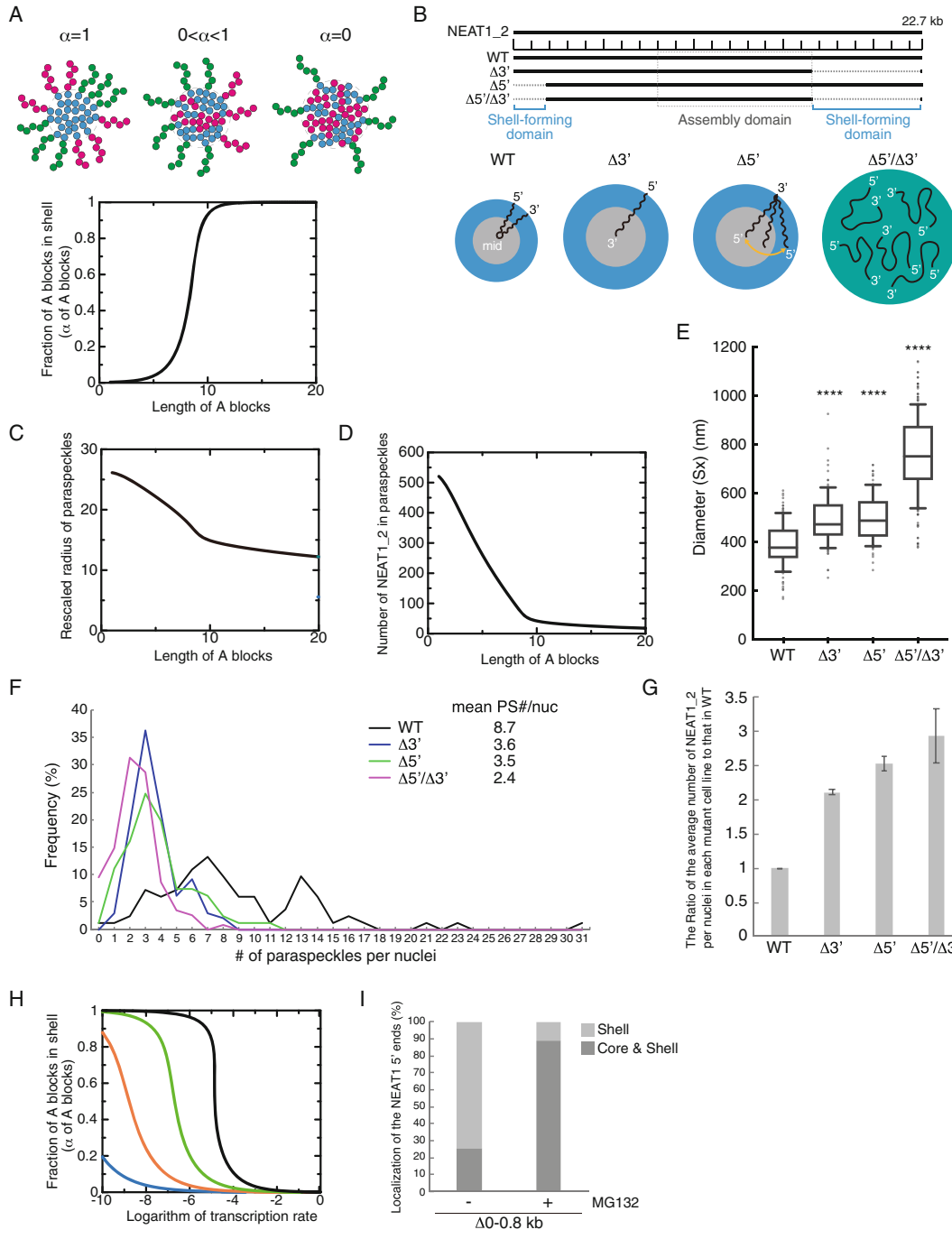


Fig. 5 Example of the experimental validations of block copolymer micelle model. (**a**) The upper schematics show that, in the copolymer micelle model, B blocks (blue) are localized in the core and C blocks (green) are localized in the shell. A fraction, α, of A blocks (magenta) are localized in the shell, and the other fraction, $1-\alpha$, are localized in the core. A lower graph shows the fraction α of the A blocks in the shell versus the length of the A blocks (represented by the number of segments) of spherical paraspeckles in the steady state. (**b**) Functional NEAT1_2 RNA domains for paraspeckle assembly and shell-localization of the 5′ and 3′ ends of the NEAT1_2 within the paraspeckles. Summary of the NEAT1_2 organization within the paraspeckle and the size of the paraspeckles in the NEAT1_2 mutants. (**c**) Theoretical calculation of the radius of paraspeckles versus the

- The terminal regions of the NEAT1_2 are redistributed into the core of the paraspeckle as the length of the terminal regions (the A or C blocks) decrease (Fig. 5a and b).
- The deletion of 5′ and/or 3′ regions of the NEAT1_2, corresponding to A and/or C blocks, leads to increase incorporated number of NEAT1_2 RNP per paraspeckle and increase the size and number of the paraspeckle (Fig. 5c–g).
- As transcription rate (the expression levels of NEAT1_2) increases, the terminal regions of the NEAT1_2 are redistributed to the core (Fig. 5h and i).
- Paraspeckles in the NEAT1_2 lacking both 5′ and 3′ region (NEAT1_2 Δ5′/Δ3′), corresponding to A and C blocks, are spheres without internal core-shell architectures (Figs. 5b and 6c).

An experimental result from RNA-driven phase separation model based on the Flory–Huggins theory is shown below.

- Disordered spherical condensates constructed by NEAT1_2 Δ5′/Δ3′ mutant become larger as the NEAT1_2 transcription upregulation (Fig. 6a–d). In contrast, short axis of the paraspeckles constructed by NEAT1_2 WT does not change significantly as the NEAT1_2 expression is enhanced (Fig. 6b–d).

These experiments show that our theoretical models effectively represent and predict features of the paraspeckle. From a combination of theoretical and experimental analyses, we have shown that

Fig. 5 (continued) length of the A blocks (represented by the number of segments) of spherical paraspeckles in the steady state. The radius was rescaled by the segment length. (**d**) Theoretical calculation of the number of NEAT1_2 transcripts versus the length of the A blocks (represented by the number of segments) of spherical paraspeckles in the steady state. (**e**) Diameters (Sx) of the paraspeckles in WT, Δ3′, Δ5′, and Δ5′/Δ3′ cells treated with MG132 (5 μM for 6 h) determined by SRM. WT (mean size: 387.5 nm), Δ3′ (mean size: 492.4 nm), Δ5′ (mean size: 497.7 nm), and Δ5′/Δ3′ mutant (mean size: 753.3 nm). (****: $P < 0.0001$, compared with WT: Kruskal–Wallis test with Dunn's multiple comparison test). Each box plot shows the median (inside line), 25–75 percentiles (box bottom to top), and 10–90 percentiles (whisker bottom to top). (**f**) The number of paraspeckles per nuclei in each cell line. The mean numbers are shown (mean PS (paraspeckle) #/nuc). Statistical analysis showed a significant reduction in the paraspeckle numbers in the Δ3′ ($P < 0.0001$), Δ5′ ($P < 0.0001$), and Δ5′/Δ3′ ($P < 0.0001$) mutants compared with the WT. The number of paraspeckles in the Δ5′/Δ3′ was significantly fewer than that in the Δ3′ and Δ5′ ($P < 0.0001$ compared with Δ3′, $P = 0.0023$ compared with Δ5′). (**g**) The ratio of the average number of NEAT1_2 per nuclei in each mutant cell line to that in the WT. Data are represented as mean ± SD ($n = 3$). (**h**) Theoretical calculation of the fraction of A (or C) blocks in the shell vs. the logarithm of the transcription rate (rescaled by the rate at which the transcripts are spontaneously incorporated in the paraspeckle). The numbers of segments of A blocks were 5.0 (cyan), 7.5 (orange), 10.0 (green), and 12.5228 (black). (**i**) Graph showing the proportion of paraspeckles with localization of the NEAT1 5′ ends to the core and shell or the shell in Δ0–0.8 kb cells under MG132-untreated (steady state) and -treated (NEAT1_2 upregulated) conditions. (The data presented in this Figure have been modified from our paper [20])

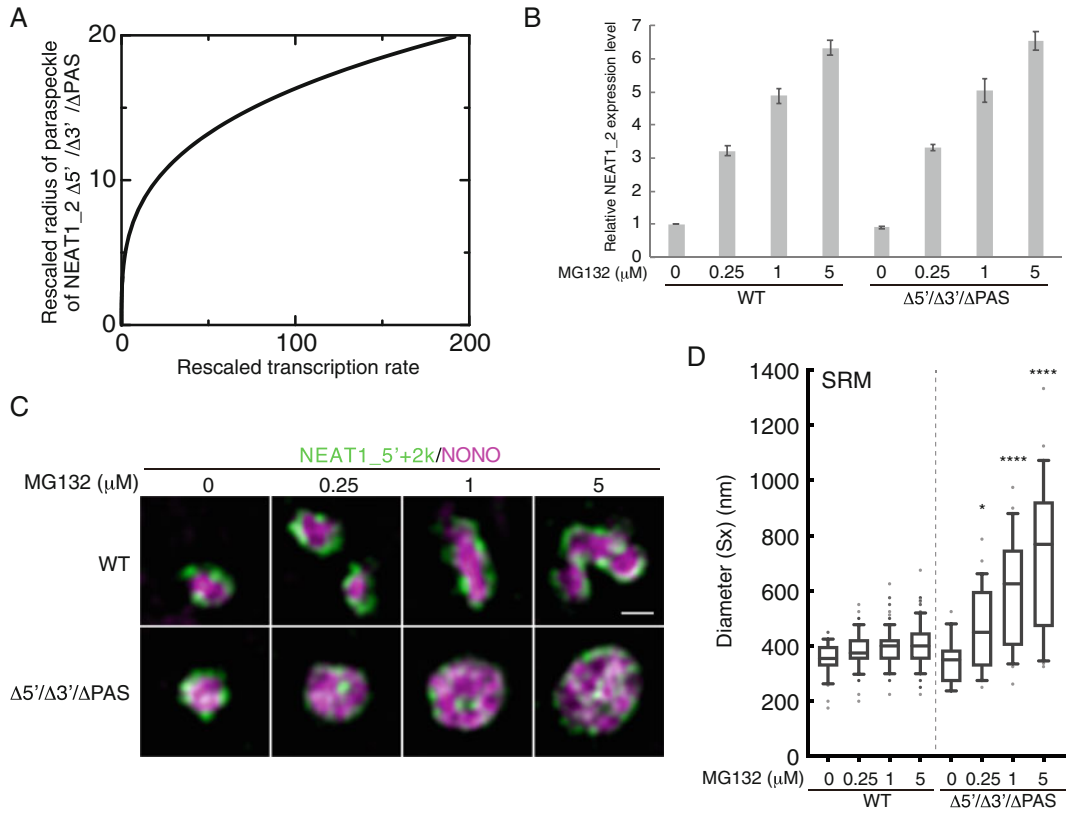


Fig. 6 Examples of the experimental validations of RNA-driven phase separation model based on the Flory–Huggins theory. (**a**) Theoretical calculation of the rescaled radius of paraspeckles versus the rescaled transcription rate for the case in which the A and C blocks were deleted by CRISPR/Cas9 in the steady state. The graph is derived by assuming that NEAT1_2 is produced at a constant rate. The produced NEAT1_2 RNPs diffuse in the solution and the free diffusion is hindered by the attractive interactions between NEAT1_2 RNPs with the interaction parameter χ (we used $\chi = 1.0$). NEAT1_2 was degraded at the constant rate k_0. The radius was rescaled by the length scale $\sqrt{D/k_0}$ and the transcription rate was rescaled by the inverse time scale $k_0(D/k_0)^{3/2}/v_0$. (**b**) Quantitation of NEAT1_2 by RT-qPCR in WT and Δ5′/Δ3′ cells with or without MG132 treatment (6 h). Data are represented as mean ± SD ($n = 3$). (**c**) The paraspeckles in WT and Δ5′/Δ3′/ΔPAS cells detected by SRM using NEAT1_5′ and 2k FISH probes (green), and NONO IF (magenta) under the conditions shown in (**b**). Scale bar, 500 nm. (**d**) Quantitation of the diameter (Sx) observed by SRM in WT and Δ5′/Δ3′ cells under the conditions shown in (**b**). WT [mean size: MG132 (−), 354.2 nm; MG132 0.25 μM, 380.5 nm; MG132 1 μM, 394.1 nm; MG132 5 μM, 409.6 nm], Δ5′/Δ3′ [mean size: MG132 (−), 344.4 nm; MG132 0.25 μM, 472.1 nm; MG132 1 μM, 601.4 nm; MG132 5 μM, 721.2 nm]. (*: $P = 0.029$, ****: $P < 0.0001$, compared with WT: Kruskal–Wallis test with Dunn's multiple comparison test). Each box plot shows the median (inside line), 25–75 percentiles (box bottom to top), and 10–90 percentiles (whisker bottom to top). (The data presented in this figure have been modified from our paper [20])

RNPs behave as block copolymers and form micelles. This micellization mechanism would be potentially important for several aspects of condensate formation and function: (1) internal core-shell architecture, (2) restricted size of the condensates (almost a constant size and not too larger condensates), (3) larger number of

condensates and total surface area, and (4) rare coalescence of condensates [20]. Finally, investigations using both theoretical and experimental approaches would enable us to design experiments from predictions of theoretical models. Such investigations strongly facilitate our understanding of underlying mechanisms for the formation and function of biomolecular condensates.

4 Notes

1. One can derive Eq. (5) by using the Boltzmann's principle

$$S = k_{\mathrm{B}} \log W, \tag{30}$$

where W is the number of ways to accommodate A and B molecules in the lattice sites. The number W of ways is calculated as

$$W = \frac{M!}{N_{\mathrm{A}}!(M - N_{\mathrm{A}})!} \tag{31}$$

by using the number N_{A} of A molecules. We note that the volume fraction ψ is related by the number N_{A} of A molecules as $\psi = N_{\mathrm{A}}/M$. By substituting Eq. (31) into Eq. (30), the entropy is derived as

$$S = -k_{\mathrm{B}} \left[N_{\mathrm{A}} \log \frac{N_{\mathrm{A}}}{M} + (M - N_{\mathrm{A}}) \log \left(1 - \frac{N_{\mathrm{A}}}{M} \right) \right]. \tag{32}$$

Eq. (32) is derived by using the Stirling's approximation

$$\log n! \simeq n \log n - n, \tag{33}$$

which is effective for large values of n. By using the relationship $\psi = N_{\mathrm{A}}/M$, Eq. (32) is rewritten in the form of Eq. (5).

2. The entropy due to the conformational fluctuation of polymers (which is called the conformational entropy) is derived by using the Boltzmann's principle, Eq. (30). In this case, we use the number of conformation of the chain for W. We think of a polymer chain composed of N repeating units, the length of each is b. The conformation of the chain is represented by using vectors linking the repeating units, b_1, b_2, $\cdots$, b_N. We assume that the correlations between any of the two vectors are zero

$$\langle b_n \cdot b_m \rangle = 0 \tag{34}$$

if $m \neq n$ and $\langle b_n \cdot b_n \rangle = b^2$. The vector R between the two ends of the polymer chain (which is called the end-to-end vector) has the form

$$R = \sum_{n=1}^{N} b_n. \tag{35}$$

The thermodynamic average of the end-to-end vector is zero, $\langle R \rangle = 0$, because the probability with which the vector b_n orients to one direction is the same as the probability with which the vector b_n orients to the opposite direction. Therefore, the size of a polymer chain is usually characterized by the mean square of the end-to-end vector

$$\left\langle R^2 \right\rangle = \langle R \cdot R \rangle = \sum_{n=1}^{N} \sum_{m=1}^{N} \langle b_n \cdot b_m \rangle = \sum_{n=1}^{N} \langle b_n \cdot b_n \rangle = Nb^2. \quad (36)$$

The end-to-end vector is represented by the components of the three directions, x, y, and z, in the isotropic 3d space, $R = (R_x, R_y, R_z)$, where the vector calculus tells you $R^2 = R \cdot R = R_x^2 + R_y^2 + R_z^2$. Because the space is isotropic, the mean square average of each component is identical

$$\left\langle R_x^2 \right\rangle = \left\langle R_y^2 \right\rangle = \left\langle R_z^2 \right\rangle = \frac{Nb^2}{3}. \quad (37)$$

The central limit theorem predicts that the probability distribution of a component (for example, R_x) of the end-to-end vector is a Gaussian function

$$P(R_x) = \sqrt{\frac{3}{2\pi N b^2}} e^{-3R_x^2/\left(2Nb^2\right)}, \quad (38)$$

where we used $\langle R_x \rangle = 0$ and $\left\langle R_x^2 \right\rangle = Nb^3/3$. Because of the fact that the three directions, x, y, and z, are independent, the probability with which the end-to-end vector of the chain is R has the form.

$$P(R) = P(R_x)P\left(R_y\right)P(R_z) = \left(\frac{3}{2\pi N b^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Nb^2}}. \quad (39)$$

The number of conformations of the polymer chain with the fixed end-to-end vector R is proportional to the probability $P(R)$, $W = \Omega^N P(R)$ (Ω is the number of directions that each of the vectors, b_1, b_2, $\cdots$, b_N, can take). The conformational entropy has the form

$$S = -\frac{3}{2} k_B \frac{R^2}{Nb^2}, \quad (40)$$

where we neglect the terms $N \log \Omega + \frac{3}{2} \log \frac{3}{2\pi N b^2}$, which do not depend on R. By substituting Eq. (40) into Eq. (1), the conformational free energy is derived as the form

$$F = \frac{3}{2} k_B T \frac{R^2}{Nb^2}. \quad (41)$$

Note that the conformational free energy, Eq. (41), is proportional to the square of the end-to-end vector R, where it is analogous to the free energy stored in a spring is proportional to the square of the extension. The quantity corresponding to the stiffness of the spring is $3k_BT/(Nb^2)$. Microscopically, the free energy stored by extending a metallic spring is due to the energetic cost by the displacement of atoms. The stiffness of the spring usually decreases as the temperature increases. In contrast, a polymer chain behaves as a spring because of its conformational entropy (entropic spring). The stiffness of the polymer chain indeed increases as the temperature increases. You can test it by putting a weight to a rubber band in a water bath and then observe the changes of the extension of the rubber band when you increase the temperature. Note that Eq. (40) is the conformational entropy of a polymer (the entropy due to the conformational fluctuation of a polymer chain), which is different from the first term of Eq. (11), which represents the translational entropy of polymers (the entropy due to the diffusion of the center of mass of polymer chains).

3. For cases in which the second virial coefficient is small, $a_2 \approx 0$, the free energy of the shell has the form

$$\frac{F_{sh}}{k_BT} = \int_{r_c}^{r_{ex}} \frac{4\pi r^2}{R^3(r)} \left[\frac{3}{2} \frac{R^2(r)}{b^2 g(r)} + v \frac{g^2(r)}{R^3(r)} + w \frac{g^3(r)}{R^6(r)} \right]. \tag{42}$$

The first term of Eq. (42) is the stretching free energy of the A blocks in the shell, *see* Eq. (41) in **Note 2** to make sense of this term. The second term of Eq. (42) is the free energy due to the two-body excluded volume interactions between A monomers. The third term of Eq. (42) is the three-body excluded volume interactions between A monomers. v $(=b^3a_2)$ is the excluded volume that accounts for the two-body interactions between A monomers. w $(=b \wedge 3)$ is the excluded volume that accounts for the three-body interactions between A monomers. $R(r)$ is the size of blobs at the distance r from the center of the micelle and thus the form $4\pi r^2 = n R^2(r)$, where n is the number of block copolymers in the micelle. $g(r)$ is the number of A monomers in each blob and is determined by the relationship

$$R(r) = bg^{1/2}(r). \tag{43}$$

4. The surface free energy F_{sur}/n per block copolymer is a decreasing function of the number n of block copolymers in a micelle, whereas the free energy of the core and the shell, F_{cr}/n and F_{shl}/n, per block copolymer is an increasing function of n.

The minimum of the free energy F_n/n per block copolymer is determined by the balance of these contributions.

For cases in which the free energy of the core dominates the free energy of the shell, $F_{\mathrm{cr}} \gg F_{\mathrm{shl}}$ (core-limited regime), the number n of block copolymers at the minimum of the free energy per block copolymer has an asymptotic form

$$n \approx \frac{\gamma b^2}{k_{\mathrm{B}} T} N_{\mathrm{B}}. \tag{44}$$

Here and after, we denote the equality with which we neglect the numerical factor of order unity (so-called scaling equality) by $\approx$. In this asymptotic limit, the number n of block copolymers with which the free energy per block copolymer becomes minimum does not depend on the number N_{A} of monomers in A blocks, *see* Fig. 4c for cases in which the number N_{A} of monomers in A blocks is small.

In the other limit, $F_{\mathrm{cr}} \ll F_{\mathrm{shl}}$ (shell-limited regime), the number n of block copolymers with which the free energy per block copolymer becomes minimum is the solution of

$$\log\left(1+\frac{h}{r_{\mathrm{c}}}\right) - \frac{4}{9}\frac{\frac{h}{r_{\mathrm{c}}}}{1+\frac{h}{r_{\mathrm{c}}}} \approx \frac{\gamma b^2}{k_{\mathrm{B}} T} N_{\mathrm{B}}^{2/3} n^{-5/6} \tag{45}$$

with

$$\frac{h}{r_{\mathrm{c}}} \approx \frac{N_{\mathrm{A}}}{N_{\mathrm{B}}^{5/9}} \left(\frac{v}{b^3}\right)^{1/3} n^{-2/9}. \tag{46}$$

The solution of Eq. (45) is approximately equal to Eq. (44) at the crossover between the core-limited and shell-limited regimes, see the magenta broken line in Fig. 7. The number N_{A}^* of monomers in A blocks at the crossover thus has approximate forms

$$\frac{N_{\mathrm{A}}^* k_{\mathrm{B}} T}{\gamma b^2} \approx \left(\frac{v}{b^3}\right)^{-\frac{1}{3}} \left(\frac{N_{\mathrm{B}} k_{\mathrm{B}} T}{\gamma b^2}\right)^{\frac{7}{9}} \mathrm{e}^{\left(\frac{N_{\mathrm{B}} k_{\mathrm{B}} T}{\gamma b^2}\right)^{-\frac{1}{6}}} \tag{47}$$

for $\frac{h}{r_{\mathrm{c}}} \gg 1$ and

$$\frac{N_{\mathrm{A}}^* k_{\mathrm{B}} T}{\gamma b^2} \approx \left(\frac{v}{b^3}\right)^{-\frac{1}{3}} \left(\frac{N_{\mathrm{B}} k_{\mathrm{B}} T}{\gamma b^2}\right)^{\frac{11}{18}} \tag{48}$$

for $\frac{h}{r_{\mathrm{c}}} \ll 1$. Eqs. (47) and (48) are covered by an interpolation formula

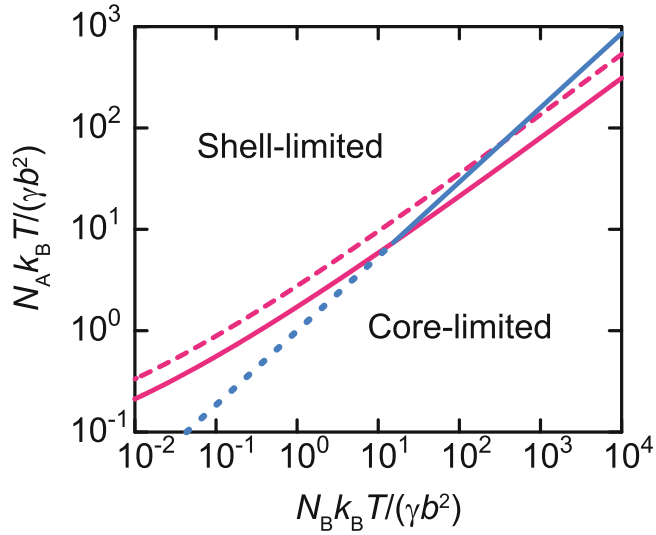


Fig. 7 The crossover between the shell-limited and the core-limited regimes shown with respect to the numbers, N_A and N_B, of monomers in A and B blocks (magenta solid line), as predicted by using the interpolation formula Eq. (36). The numerical calculation by using Eq. (33) is shown by the magenta broken line. The cyan line divides the regimes of the crew-cut regime ($h/r_c \ll 1$) and the starlike regime ($h/r_c \gg 1$), *see* Eq. (39). We used $v/b^3 = 1$ for the calculation. The numerical factors are neglected in these calculations (*see* **Note 2**)

$$\frac{N_A^* k_B T}{\gamma b^2} \approx \left(\frac{v}{b^3}\right)^{-\frac{1}{3}} \left(\frac{N_B k_B T}{\gamma b^2}\right)^{\frac{7}{9}} \left[e^{\left(\frac{N_B k_B T}{\gamma b^2}\right)^{-\frac{1}{6}}} - 1 \right], \tag{49}$$

see the magenta solid line in Fig. 7.

For the cases of $h/r_c \ll 1$ (crew-cut regime), the solution of Eq. (45) has an approximate form

$$n \approx \left(\frac{\gamma b^2}{k_B T}\right)^{\frac{18}{11}} \left(\frac{v}{b^3}\right)^{-\frac{6}{11}} N_A^{-\frac{18}{11}} N_B^2. \tag{50}$$

For the cases of $h/r_c \gg 1$ (starlike regime), the solution of Eq. (45) has an approximate form

$$n \approx \left(\frac{\gamma b^2}{k_B T}\right)^{6/5} N_B^{4/5} \left[\log\left(\frac{N_A}{N_B^{5/9}} \left(\frac{v}{b^3}\right)^{1/3} \right) \right]^{-6/5}, \tag{51}$$

but, Eq. (35) may not be a good approximation because we neglected the logarithmic correction. The crossover between the two regimes is at

$$\frac{N_A^{**} k_B T}{\gamma b^2} \approx \left(\frac{N_B k_B T}{\gamma b^2}\right)^{\frac{11}{15}} \left(\frac{v}{b^3}\right)^{-\frac{1}{3}}, \tag{52}$$

see the cyan line in Fig. 7.

5. Detail experimental procedures of EM analysis, super-resolution microscopic observation (SIM: structured illumination microscopy) of RNA-FISH and immunofluorescence, and CRISPR-mediated deletion of lncRNA, have been reported [34–36].

Acknowledgments

This research was supported by KAKENHI grants from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan [to T. Yamazaki (19K06479, 19H05250, 21H00253), T. Yamamoto (18K03558, 19H05259, 20H05934, 21K03479, 21H00241)], JST, PRESTO Grant Number JPMJPR18KA (to T. Yamamoto), the Mochida Memorial Foundation for Medical and Pharmaceutical Research (to T. Yamazaki), the Naito Foundation (to T. Yamazaki), and the Takeda Science Foundation (to T. Yamazaki). T. Yamamoto acknowledges S. A. Safran (Weizmann Institute of Science) and Takahiro Sakaue (Aoyama Gakuin University) for their critical reading of the manuscript.

References

1. Banani SF et al (2017) Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol 18(5):285–298
2. Schmidt HB, Gorlich D (2016) Transport selectivity of nuclear pores, phase separation, and Membraneless organelles. Trends Biochem Sci 41(1):46–61
3. Fujioka Y et al (2020) Phase separation organizes the site of autophagosome formation. Nature 578(7794):301–305
4. Spannl S et al (2019) Biomolecular condensates in neurodegeneration and cancer. Traffic 20(12):890–911
5. Shin Y, Brangwynne CP (2017) Liquid phase condensation in cell physiology and disease. Science 357(6357):eaaf4382
6. Mathieu C, Pappu RV, Taylor JP (2020) Beyond aggregation: pathological phase transitions in neurodegenerative disease. Science 370(6512):56–60
7. Zbinden A et al (2020) Phase separation and neurodegenerative diseases: a disturbance in the force. Dev Cell 55(1):45–68
8. Chujo T, Yamazaki T, Hirose T (2016) Architectural RNAs (arcRNAs): a class of long noncoding RNAs that function as the scaffold of nuclear bodies. Biochim Biophys Acta 1859(1):139–146
9. Maharana S et al (2018) RNA buffers the phase separation behavior of prion-like RNA binding proteins. Science 360(6391):918–921
10. Van Treeck B, Parker R (2018) Emerging roles for intermolecular RNA-RNA interactions in RNP assemblies. Cell 174(4):791–802
11. Yamazaki T, Nakagawa S, Hirose T (2020) Architectural RNAs for Membraneless nuclear body formation. Cold Spring Harb Symp Quant Biol 84:227–237
12. Ninomiya K, Hirose T (2020) Short tandem repeat-enriched architectural RNAs in nuclear bodies: functions and associated diseases. Noncoding RNA 6(1):6
13. Chujo T et al (2017) Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncoding RNAs. EMBO J 36(10):1447–1462
14. Yap K et al (2018) A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. Mol Cell 72(3): 525–540.e13
15. Safran SA (2003) Statistical thermodynamics of surfaces, interfaces, and membranes. Westview Press, CO., Routledge
16. Doi M (2013) Soft matter physics. Oxford Univ. Press, Oxford

17. Doi M (1996) Introduction to polymer physics. Oxford Univ. Press, Oxford
18. de Gennes PG (1979) Scaling concepts in polymer physics. Cornell Univ. Press, New York
19. Rubinstein M, Colby R (2003) Polymer physics. Oxford Univ. Press, New York
20. Yamazaki T, Yamamoto T, Yoshino H, Souquere S, Nakagawa S, Pierron G, Hirose T (2021) Paraspeckles are constructed as block copolymer micelles. EMBO J 40(12): e107270. https://doi.org/10.15252/embj.2020107270
21. Yamamoto T, Yamazaki T, Hirose T (2020) Triblock copolymer micelle model of spherical paraspeckles. bioRxiv. https://doi.org/10.1101/2020.11.01.364190
22. Halperin A, Alexander S (1989) Polymer micelles: their relaxation kinetics. Macromolecules 22:2403–2412
23. Zhulina EB, Adam M, LaRue I, Sheiko SS, Rubinstein M (2005) Diblock copolymer micelles in a dilute solution. Macromolecules 38:5330–5535
24. Yamamoto T, Yamazaki T, Hirose T (2020) Phase separation driven by production of architectural RNA transcripts. Soft Matter 16(19): 4692–4698
25. Sasaki YT et al (2009) MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc Natl Acad Sci U S A 106(8):2525–2530
26. Sunwoo H et al (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res 19(3):347–359
27. Chen LL, Carmichael GG (2009) Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol Cell 35(4):467–478
28. Clemson CM et al (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 33(6):717–726
29. Yamazaki T et al (2018) Functional domains of NEAT1 architectural lncRNA induce Paraspeckle assembly through phase separation. Mol Cell 70(6):1038–1053 e7
30. Mao YS et al (2011) Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. Nat Cell Biol 13(1): 95–101
31. Naganuma T et al (2012) Alternative 3′-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. EMBO J 31(20):4020–4034
32. Souquere S et al (2010) Highly ordered spatial organization of the structural long noncoding NEAT1 RNAs within paraspeckle nuclear bodies. Mol Biol Cell 21(22):4020–4027
33. West JA et al (2016) Structural, super-resolution microscopy analysis of paraspeckle nuclear body organization. J Cell Biol 214(7): 817–830
34. Souquere S, Pierron G (2015) Ultrastructural analysis of nuclear bodies using electron microscopy. Methods Mol Biol 1262:105–118
35. Mito M et al (2016) Simultaneous multicolor detection of RNA and proteins using super-resolution microscopy. Methods 98:158–165
36. Yamazaki T, Hirose T (2021) CRISPR-mediated mutagenesis of long noncoding RNAs. Methods Mol Biol 2254:283–303

Correction to: CRISPR-Mediated Activation of Transposable Elements in Embryonic Stem Cells

Akihiko Sakashita, Masaru Ariura, and Satoshi H. Namekawa

Correction to:
Chapter 11 in: Nicolas F. Parrish, Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*, Methods in Molecular Biology, vol. 2509,
https://doi.org/10.1007/978-1-0716-2380-0_11

Table 6 of chapter 11 was printed incorrectly. The first row was incorrectly split into two rows ("P13 GA cl pLV-U6-" and "gRNA-DsRed_Fw"), which has now been corrected by merging these two rows.

The book has also been updated to reflect this change.

The updated original version of this chapter can be found at
https://doi.org/10.1007/978-1-0716-2380-0_11

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0_23,

Table 6
List of primer sequence used in amplification of gRNA transcriptional unit

Primer ID	Sequence (5'→3')
P13 GA cl pLV-U6-gRNA-DsRed_Fw	caccatctttaattgcttcagaaacttcgaaGAGGGCCTATTTCCCATGATTC
P2; sgRNA1_Rv (+Index1)	ttggaaagctcgtcttagacACGCGCTAAAAACGGACTAGC
P3; sgRNA2_Fw (+index1 anchor)	gtctaagacgagctttccaaGAGGGCCTATTTCCCATGATTC
P4; sgRNA2_Rv (+Index2)	gatacttacagctaccactacACGCGCTAAAAACGGACTAGC
P5; sgRNA3_Fw (+index2 anchor)	gtagtggtagctgtaagtatcGAGGGCCTATTTCCCATGATTC
P6; sgRNA3_Rv (+Index3)	ggttagtcaacaatgtgtccaACGCGCTAAAAACGGACTAGC
P7; sgRNA4_Fw (+index3 anchor)	tggacacattgttgactaaccGAGGGCCTATTTCCCATGATTC
P8; sgRNA_Rv (+index4)	aggttactcgcactgttgaaACGCGCTAAAAACGGACTAGC
P9; sgRNA_Fw (+index4 anchor)	ttcaacagtgcgagtaacctGAGGGCCTATTTCCCATGATTC
P14 GA cl to pLV-U6-gRNA-DsRed_Rv	acatgatggtcattttgtaagattagatggaaatcACGCGCTAAAAACGGACTAGC

INDEX

Nicholas F. Parrish and Yuka W. Iwasaki (eds.), *piRNA: Methods and Protocols*,
Methods in Molecular Biology, vol. 2509, https://doi.org/10.1007/978-1-0716-2380-0,

H

I

L

M

N

O

P

R

S

T

U

V

W