Methods in Molecular Biology 1543

Springer Protocols

Sara Napoli *Editor*

Promoter Associated RNA

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, AL10 9AB, UK

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

Promoter Associated RNA

Methods and Protocols

Edited by

Sara Napoli

Experimental Therapeutics Group, Institute of Oncology Research, Bellinzona, Switzerland

Humana Press

Editor
Sara Napoli
Experimental Therapeutics Group
Institute of Oncology Research
Bellinzona, Switzerland

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-6714-8 ISBN 978-1-4939-6716-2 (eBook)
DOI 10.1007/978-1-4939-6716-2

Library of Congress Control Number: 2016958614

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

A large portion of the human genome encodes for long noncoding RNAs (lncRNAs), transcripts longer than 200 nucleotides and without an open reading frame, which play widespread roles in gene regulation and other cellular processes. They can be transcribed from intergenic regions, but many of them are associated with annotated protein-coding genes, instead.

Noncoding and coding transcripts at a given locus can overlap, and this phenomenon is called transcriptional forest. It includes sense-antisense transcripts coming from transcription of both strands, sense noncoding RNA overlapping mRNA from the sense strand but not encoding for protein, or totally intronic RNAs.

Other long ncRNAs linked to coding genes are associated with enhancers (eRNAs) or promoters (paRNAs). They are generally low expressed but involved in important cellular mechanisms, like the dynamics of nuclear architecture, chromatin remodeling, and transcriptional regulation. They exert their function *in cis* affecting neighbor gene expression.

Both eRNAs and paRNAs can interact with DNA by several mechanisms, forming triplex with dsDNA, displacing a single strand of DNA to form RNA:DNA hybrid (so-called R loops), interacting with nascent RNA or with DNA-binding proteins in a sequence-specific manner. As many other long ncRNAs, they can interact with proteins important for transcriptional regulation. In those RNP complexes, they often are the scaffold which brings two or more proteins together and enable their co-localization and interaction. Another interesting role is the assignment of specificity to the binding of one or more proteins to a given genomic locus. Their regulated expression, for instance, can recruit certain protein complexes to an allele but not to the other.

Many of their interactors are epigenetic regulators. Epigenetic processes, such as DNA methylation and histone post-translational modifications which influence chromatin remodeling, contribute to the pathogenesis of many diseases, in particular human cancers, and impact on disease progression, treatment responses, and clinical outcome.

Promoter-associated RNA of key disease genes may represent a natural switch to exploit in order to manipulate dysregulation of relevant coding transcripts. PaRNAs expression, under certain conditions, can enable or not the binding of specific proteins to a promoter and selectively modulate the transcription of adjacent gene. According to this, they can be considered valuable regulatory elements to investigate further.

The main goal of this book is to summarize methods of molecular biology, biochemistry, and bioinformatics, useful to explore the expression and functions of promoter-associated RNA, which, among the classes mentioned till now, are still less characterized.

The book is subdivided into four parts. In the first part, genome-wide approaches are described to identify functionally relevant elements in noncoding regions and to detect transcription in correspondence of promoters. Importantly, bioinformatics chapters are included to help the reader to use publicly available data as a source of information about noncoding transcriptome.

In the second part of this book, techniques useful to deeply characterize paRNA structural features are described. Accurate investigation of physical features of a noncoding RNA, such as extension, secondary structure, and binding affinity for RNA-binding proteins, can help in understanding its putative function. Promoter-associated RNAs can be

considered a new class of molecules which play a key role in transcriptional regulation. In the third part is described how selecting good therapeutic target among paRNAs relevant in diseases and how impeding their function by RNA interference. Strategies to investigate transcriptional gene silencing mechanism are described. Further, some methods are reported to study R-loop structures, RNA:DNA hybrid structures used by noncoding RNAs to regulate gene expression, changing local chromatin environment. The last part of the book is dedicated to paRNA therapeutic potential. This part also describes how siRNAs directed against paRNAs can be applied in vivo to modulate transcription of important genes controlled by paRNAs.

I hope this book will help the reader to appreciate the potential of paRNAs as a new class of regulatory molecules to further investigate and value as tools for fine transcriptional tuning.

Bellinzona, Switzerland *Sara Napoli*

Contents

Contributors

Reuven Agami • *Division of Biological Stress Response, The Netherlands Cancer Institute, Amsterdam, The Netherlands; Erasmus MC, Rotterdam University, Rotterdam, The Netherlands*

Ophélie Arnaud • *Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa, Japan*

Raquel Boque-Sastre • *Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Catalonia, Spain*

Anouar Boucheham • *IBISC, UEVE/Genopole, Evry, France; College of NTIC, Constantine University 2, Constantine, Algeria*

Parthaprasad Chattopadhyay • *Department of Biochemistry, All India Institute of Medical Sciences, New Delhi, India*

Gianluca Civenni • *Laboratory of Experimental Therapeutics, IOR, Institute of Oncology Research, Bellinzona, Switzerland*

Drena Dobbs • *Genetics, Development and Cell Biology Department, Iowa State University, Ames, IA, USA*

Manuel Esteller • *Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Catalonia, Spain; Departament de Ciències Fisiològiques II, Escola de Medicina, Universitat de Barcelona, Barcelona, Catalonia, Spain; Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain*

Richard J. Fish • *Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland*

Alexandre Fort • *Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland*

Sonia Guil • *Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Catalonia, Spain*

Nobuhiko Hamazaki • *Department of Stem Cell Biology and Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan*

Matthias Harbers • *Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa, Japan; RIKEN Omics Science Center (OSC), Yokohama, Japan*

Takuya Imamura • *Department of Stem Cell Biology and Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan*

Ranveer Singh Jayani • *Howard Hughes Medical Institute, Department of Medicine, School of Medicine, University of California, San Diego, La Jolla, CA, USA*

Sachi Kato • *Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa, Japan; RIKEN Omics Science Center (OSC), Yokohama, Japan*

Gozde Korkmaz • *Division of Biological Stress Response, The Netherlands Cancer Institute, Amsterdam, The Netherlands*

Rui Lopes • *Division of Biological Stress Response, The Netherlands Cancer Institute, Amsterdam, The Netherlands*
Marco Magistri • *Center for Therapeutic Innovation, Department of Psychiatry and Behavioral Sciences, University of Miami, Miller School of Medicine, Miami, FL, USA*
Carla M. Mann • *Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA, USA*
Georgi K. Marinov • *Department of Biology, Indiana University, Bloomington, IN, USA*
Jean-Étienne Morlighem • *RIKEN Omics Science Center (OSC), Yokohama, Japan; Laboratory of Biochemistry and Biotechnology, Institute for Marine Sciences, Federal University of Ceara, Fortaleza, CE, Brazil*
Usha K. Muppirala • *Genome Informatics Facility, Iowa State University, Ames, IA, USA*
Kinichi Nakashima • *Department of Stem Cell Biology and Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan*
Sara Napoli • *Laboratory of Experimental Therapeutics, IOR, Institute of Oncology Research, Bellinzona, Switzerland*
Dimple Notani • *Howard Hughes Medical Institute, Department of Medicine, School of Medicine, University of California, San Diego, La Jolla, CA, USA; National Center for Biological Sciences, Bangalore, Karnataka, India*
Chiara Pastori • *Sylvester Comprehensive Cancer Center, Department of Human Genetics, Miller School of Medicine, University of Miami, Miami, FL, USA*
Veronica Julia Peschansky • *Center for Therapeutic Innovation and Department of Psychiatry and Behavioral Sciences, Miller School of Medicine, University of Miami, Miami, FL, USA*
Charles Plessy • *Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan; RIKEN Omics Science Center (OSC), Yokohama, Kanagawa, Japan*
Stéphane Poulain • *Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa, Japan*
Debi P. Sarkar • *Department of Biochemistry, University of Delhi, New Delhi, India*
Amanjot Singh • *Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA*
Marta Soler • *Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Catalonia, Spain*
Makoto Suzuki • *RIKEN Omics Science Center (OSC), Yokohama, Japan; DNAFORM, Inc., Yokohama, Kanagawa, Japan*
Fariza Tahi • *IBISC, UEVE/Genopole, Evry, France; IPS2, University of Paris-Saclay, Gif-sur-Yvette, France*
Van Du T. Tran • *Vital-IT group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*
Dmitry Velmeshev • *Department of Psychiatry and Behavioral Sciences, Center for Therapeutic Innovation, Miller School of Medicine, University of Miami, Miami, FL, USA*
Mohammad Khalid Zakaria • *Department of Biochemistry, All India Institute of Medical Sciences, New Delhi, India; International Centre for Genetic Engineering and Biotechnology, Trieste, Italy*

Part I

Identification of Promoter-Associated RNAs

Chapter 1

ChIP-seq for the Identification of Functional Elements in the Human Genome

Georgi K. Marinov

Abstract

Functional elements in the genome express their function through physical association with particular proteins: transcription factors, components of the transcription machinery, specific histone modifications, and others. The genome-wide characterization of the protein-DNA interaction landscape of these proteins is thus a key approach toward the identification of candidate genomic regulatory regions. ChIP-seq (**Ch**romatin **I**mmuno**p**recipitation coupled with high-throughput **seq**uencing) has emerged as the primary experimental methods for carrying out this task. Here, the ChIP-seq protocol is described together with some of the most important considerations for applying it in practice.

Key words Regulatory element, Transcription factor, Histone modification, Chromatin immunoprecipitation, High-throughput sequencing

1 Introduction

Chromatin immunoprecipitation [1–3] has become the primary method for characterizing protein-DNA interactions. It is usually based on the chemical cross-linking of protein molecules to DNA and their subsequent pull-down together with their associated DNA fragments using specific immune reagents. The changes in the technology used for the subsequent analysis of these fragments have been a major factor in the historical evolution of the assay. Initially, ChIP was coupled with qPCR to characterize occupancy at specific genomic regions [4]. Later, the development of microarray technology and ChIP-chip/ChIP-on-chip [5–9] allowed the parallel assaying of large numbers of genomic regions or even whole genomes. The more recent coupling of ChIP with high-throughput sequencing in the form of ChIP-seq [10–13] finally allowed the comprehensive genome-wide characterization of genomic occupancy landscapes in large genomes, with very high resolution.

ChIP-seq datasets are very information-rich and enable the identification of candidate functional regulatory elements on a

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_1,

genome-wide scale, especially when multiple datasets are integrated, and have been successfully used for this purpose in a variety of systems by the ENCODE [14, 15], mouseENCODE [16], modENCODE [17–19], Roadmap Epigenomics [20] and other large-scale consortia, as well as by individual laboratories. Candidate regulatory regions can be identified based on the physical association of particular proteins or histone modifications with DNA when their function is expressed [21]. Examples include the typical association of active gene promoters with trimethylated histone H3 (H3K4me3) [22], H3K4me1, H3K27ac, and the histone acetyltransferase p300, which are useful markers for transcriptional enhancers [23–28], the enrichment of H3K36me3 and H3K79 methylation over actively transcribed gene bodies [29, 30], CTCF and other insulator proteins [31, 32] marking insulator elements, and others. Finally, the integrative analysis of multiple genome-wide profiles for histone marks and other proteins can be used to generate more finely parsed catalogs of chromatin states and their maps in a wide variety of tissues and cell types [33–35]. Here, a version of the ChIP-seq protocol that has been successfully used to generate a large number of high-quality datasets, in particular within the ENCODE Consortium, is described, and some of the key considerations for optimal experimental design and execution are discussed.

2 Materials

Prepare all buffers and solutions by filtering through a 0.22 μm syringe filter after having mixed all components.

2.1 Crosslinking and Sonication

1. 37 % Formaldehyde (*see* **Note 1**).
2. 2.5 M Glycine.
3. PBS 1×.
4. Farnham Lysis Buffer (FLB): 5 mM HEPES (pH 8.0), 85 mM KCl, Nonidet-P40/IGEPAL CA-630 0.5 % (v/v), Protease Inhibitor Cocktail, (add immediately prior to use; 1 mini tablet for 10–15 mL of FLB, one complete tablet for 50 mL).
5. RIPA Buffer: PBS 1×, Nonidet-P40/IGEPAL CA-630 1 % (v/v), Sodium Deoxycholate 0.5 % (w/v), SDS 0.1 % (w/v), Protease Inhibitor Cocktail (add immediately prior to use).

2.2 ChIP

1. BSA solution (made fresh; can be kept at 4 °C for up to a week): 5 mg/mL BSA solution (Bovine Serum Albumin).
2. PBS 1×.
3. Magnetic beads: Dynabeads Sheep Anti-Rabbit IgG or Sheep Anti-Mouse IgG (LifeTechnologies), or equivalent product.

4. Antibodies against the protein/histone modification of interest (*see* **Note 2**).
5. LiCl IP Wash Buffer: 10 mM Tris–HCl (pH 7.5), 500 mM LiCl, Nonidet-P40/IGEPAL CA-630 1 % (v/v), Sodium Deoxycholate 0.5 % (w/v).
6. TE buffer 1×: 10 mM Tris–HCl (pH 8.0), 1 mM EDTA.
7. IP Elution Buffer: SDS 1 % (w/v), 0.1 M $NaHCO_3$.
8. Proteinase K.
9. Phenol/chloroform/isoamyl alcohol (25:24:1).

2.3 Library Building and Sequencing

1. T4 DNA Ligase and T4 DNA Ligase Buffer.
2. T4 DNA Polymerase.
3. dNTP mix.
4. T4 Polynucleotide Kinase.
5. Klenow DNA polymerase.
6. dATP.
7. 10× NEBuffer2 (NEB M0212).
8. Klenow fragment (3'->5'exo-).
9. ddH_2O.
10. Adapter oligo mix (appropriate for the sequencing platform used).
11. PCR Primer Mix (appropriate for the sequencing platform used).
12. Phusion DNA Polymerase Mix.
13. Agencourt Ampure XP beads (Beckman-Coulter).

2.4 General Materials and Equipment

1. Cell scraper (if working with adherent cells).
2. 15 or 50 mL centrifuge tubes.
3. 0.22 μm syringe filters nine. 200 μL PCR tubes.
4. Probe sonicator or equivalent (for example, Bioruptor; *see* **Note 3**).
5. Cold room (4 °C).
6. Incubator/water bath (65 °C), or a Thermomixer R.
7. Magnetic stand.
8. Tube rotator.
9. Tabletop centrifuge.
10. Tabletop vortex.
11. Thermal cycler.
12. Electrophoresis unit.
13. High-voltage power supply.
14. QIAquick PCR Cleanup Kit (QIAquick PCR cleanup columns, Buffer PM, Buffer PE (with EtOH added), Buffer EB), or equivalent.

15. 1.5 mL microcentrifuge tubes, preferably low protein and DNA binding (*see* **Note 4**).
16. Qubit fluorometer or equivalent (Life Technologies).
17. QuBit dsDNA HS Assay Kit (Life Technologies).
18. BioAnalyzer (Agilent) or equivalent.
19. 200 proof EtOH.

3 Methods

The general outline of a ChIP-seq experiment is shown in Fig. 1. The source biological material (cells cultures or tissues) is treated with a cross-linking agent (usually formaldehyde) to chemically "fix" protein-DNA interactions (*see* **Note 1**). The fixed cells are then lysed and chromatin is sheared so that fragments of length ~200 bp are generated (*see* **Note 5**). The sheared chromatin is then subjected to immunoprecipitation with an antibody against the protein or histone modification of interest, enriching for the cross-linked DNA fragments bound to it in vivo. After immunoprecipitation, cross-links are reversed at high temperature and DNA is isolated and converted into a sequencing library, by end-repair, dA addition, ligation of adapters, and PCR. The library is then sequenced. In parallel, a control library is generated, either from sheared chromatin ("Input" control), or from chromatin subjected to immunoprecipitation with an unrelated antibody ("IgG" control) (*see* **Note 6**).

3.1 Cross-Linking

Grow cells according to proper cell culture practices for the cell line studied, and up to needed density/cell numbers. Adherent cells should be grown in plates for easier harvesting. The protocol described here is robust for cell numbers on the order of 2×10^7 but the number can be lowered depending on the immunoprecipitation target, with the usual caveats (*see* **Note 7** for further discussion). Information on the fixation of tissue material can be found in [36].

3.1.1 Suspension Cells

1. Inside a chemical hood, add the appropriate amount of formaldehyde so that its final concentration is 1 %(v/v). Mix well and incubate for 15 min at RT (preferably on a lab shaker).
2. Quench the reaction by adding the appropriate amount of 2.5 M glycine so that the final concentration is 0.125 M. Mix well and incubate at RT for 5 min.
3. Transfer cells to 50 mL centrifuge tubes (or 250 mL tubes, if appropriate) and pellet by centrifuging at 4 °C for 5 min. Discard the supernatant.
4. Transfer pellets on ice and resuspend in cold PBS 1×. If necessary, pool and/or split cells from different tubes so that each tube contains the desired number of cells for a ChIP reaction. Pellet by centrifuging at 4 °C for 5 min. Discard the supernatant.

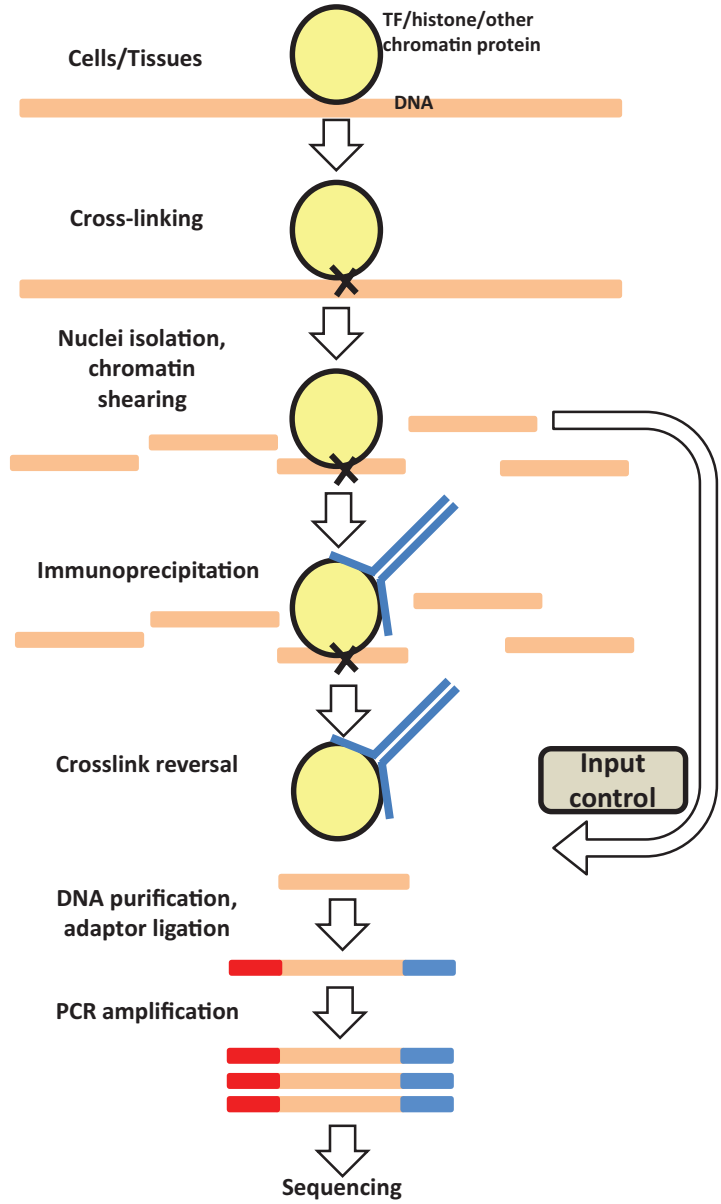


Fig. 1 Overview of a chromatin immunoprecipitation experiment. Cells or tissues are subjected to cross-linking of chromatin-associated proteins to DNA. Nuclei are then isolated and chromatin is sheared. The protein of interest, together with the DNA fragments associated with it, is then pulled down by immunoprecipitation. Cross-links are reversed, DNA is purified, and a sequencing library is built by repairing ends, ligating adapters and PCR. Note that a "sonicated input" control is depicted here, but other control samples can also be generated and used, such as IgG (*see* text for discussion)

5. The pellet can be immediately frozen in liquid nitrogen and stored at −80 °C, or resuspended in 1 mL FLB, transferred to 1.5 mL microcentrifuge tubes, and then frozen.

3.1.2 Adherent Cells

1. Add the appropriate amount of formaldehyde so that its final concentration is 1 % (v/v) to each plate. Place on a tabletop shaker inside a chemical hood and incubate for 15 min at RT with gentle shaking.
2. Quench the reaction by adding the appropriate amount of 2.5 M glycine so that the final concentration is 0.125 M. Incubate at RT for 5 min.
3. Wash cells twice with PBS 1×.
4. Collect cells from plates using a cell scraper and a small volume of cold PBS 1×, and transfer into centrifuge tubes on ice. If multiple plates are being processed, pool the cells from all plates into a single tube.
5. Transfer cells to 50 mL centrifuge tubes and pellet by centrifuging at 4 °C for 5 min.
6. Transfer pellets on ice and resuspend in cold PBS 1×. If necessary, pool and/or split cells from different tubes so that each tube contains the desired number of cells for a ChIP reaction. Pellet by centrifuging at 4 °C for 5 min. Discard the supernatant.
7. The pellet can be immediately frozen in liquid nitrogen and stored at −80 °C, or resuspended in 1 mL FLB, transferred to 1.5 mL microcentrifuge tubes, and then frozen.

This point in the protocol is usually a convenient place to stop as fixed chromatin stored at −80 °C is stable almost indefinitely.

3.2 Chromatin Shearing

There are multiple options for shearing chromatin (*see* **Note 3**). With any one of them, it is necessary to optimize conditions so that the desired extent of shearing (*see* Fig. 2) is achieved by testing a variety of treatment intensities and then evaluating the fragment distribution on a gel or the BioAnalyzer. Thus, the ideal conditions may differ from what is described below, which is the typical protocol we have used with a 1/8″ tip probe sonicator (Misonix).

It is possible to sonicate large batches of cells, ensure they are well sheared, and then keep the sonicated lysate frozen for later use. Alternatively, if shearing conditions are stable and reliable, sonication can be carried out on the second day of the protocol and the lysate used directly in the immunoprecipitation step

1. Resuspend the frozen cell pellet (~2×10^7 cells) in 1 mL cold FLB (with Protease Inhibitor Cocktail added) in a 1.5 mL microcentrifuge tube. If the cells have been frozen in FLB, thaw them one ice instead.
2. Pellet cells by centrifuging at 2000 rpm/500 × *g* at 4 °C for 5 min. Discard supernatant.
3. Resuspend in cold FLB and pellet again as in **step 2**.

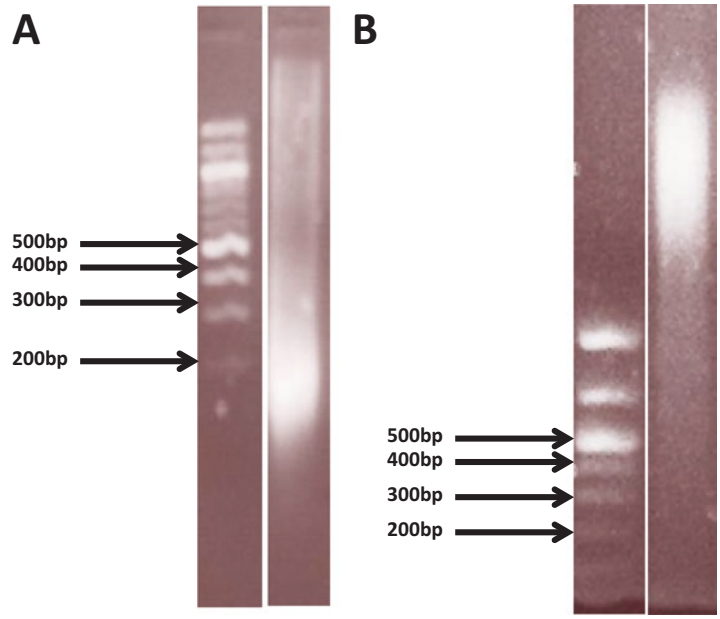


Fig. 2 Optimization of chromatin shearing. The desired range in which the bulk of the sheared fragments should be is in the neighborhood of 200 bp (**a**). Unsuccessful shearing shown in (**b**) for comparison

4. Resuspend in 1 mL cold RIPA buffer (with Protease Inhibitor Cocktail added). Incubate on ice for 10 min.
5. Sonicate in the cold room, by immersing the sonicator tip in the tube so that it reaches halfway inside its conical portion (make sure the tip does not touch the sides of the tube), with the tube immersed in cold (−20 °C) EtOH (use a bottle full of cold EtOH and a rubber plug with a hole in it, in which the tube can be securely fixed; put a magnetic stir bar in the bottle and place it on the top of a magnetic stirrer to keep the liquid circulating), using two rounds of 12 cycles of 30 s sonication, 60 s rest (a total of 2 × 6, or 12 min of sonication), and power level of 4.0. Make sure the tubes do not overheat and that no foaming occurs.
6. Centrifuge at 14,000 rpm/18,000 × *g* at 4 °C for 15 min to pellet cell debris.
7. Save 10 % of the supernatant (~100 μL) as "input DNA." Keep the input DNA at the same temperature as the ChIP samples throughout the rest of the protocol.

3.3 ChIP

The protocol described here is robust and reliable, but it takes more than 3 days to complete, mainly due to overnight incubations. There is evidence that it can be shortened substantially by reducing incubation times and carrying them out at RT instead of 4 °C (*see*, for example, [37]).

3.3.1 Day 1

1. Add 200 μL well-resuspended magnetic beads (anti-mouse or anti-rabbit, depending on the primary antibody used) to protein low-binding 1.5 mL microcentrifuge tubes.
2. Wash the beads three times in 1 mL BSA, by resuspending beads in BSA, spinning briefly to collect any liquid in the tube cap and incubating on the magnet stand for ~5 min.
3. Resuspend the beads in 1 mL BSA.
4. Add the primary antibody (typically 5 μg if the antibody is monoclonal and 10 μg if it is polyclonal).
5. Incubate overnight on a tube rotator at 4 °C.

3.3.2 Day 2

1. Place beads on a magnet for 5 min and remove the supernatant.
2. Wash magnetic beads with BSA three times as described above. Resuspend beads in 100 μL BSA after the final wash.
3. Add ~900 μL of sonicated chromatin to the beads 10. Incubate overnight on a tube rotator at 4 °C.

3.3.3 Day 3

1. Place beads on a magnet for 5 min and remove the supernatant.
2. Wash five times with LiCl buffer by resuspending in 1 mL LiCl buffer, incubating on a tube rotator at 4 °C for 10 min, spinning briefly to collect any liquid in the tube cap, incubating on the magnet for 5 min, and removing the supernatant.
3. Rinse pellet with 1 mL TE buffer.
4. Resuspend in 200 μL IP Elution Buffer at RT.
5. Incubate both ChIP and Input samples at 65 °C to dissociate antibodies from beads. Carry out this step in a Thermomixer R (2 min off, 15 s on, 1000 rpm), or in an incubator by inverting the tubes manually every 10–15 min.
6. Centrifuge at 14,000 rpm for 3 min at RT. Transfer the supernatant to a fresh 1.5 mL DNA low-binding microcentrifuge tube.
7. Add 100 μg Proteinase K to ChIP samples and 500 μg Proteinase K to Input samples.
8. Incubate for 12–16 h at 65 °C to reverse cross-links.

3.3.4 Day 4

1. Warm EB buffer at 55 °C.
2. Add an equal volume of 25:24:1 phenol/chloroform/isoamyl alcohol to Input and ChIP samples. Vortex for 20 s and centrifuge for 3 min at 14,000 rpm at RT.
3. Transfer the top (aqueous) phase to a fresh 1.5 mL DNA low-binding tube.

4. Add 3× volume of Qiagen Buffer PM and mix well.
5. Load onto a Qiagen spin column, incubate for 2 min, and centrifuge for 2 min at 14,000 rpm. 24. Discard liquid. Add 750 µL Buffer PE, incubate for 2 min, and centrifuge for 2 min at 14,000 rpm.
6. Discard liquid, then centrifuge for 2 min at 14,000 rpm to dry.
7. Elute with warm EB Buffer (50 µL) by adding it to the column, letting it soak for 1 min and then spinning for 2 min at 14,000 rpm.
8. Measure DNA concentration using the QuBit dsDNA HS Assay Kit following the manufacturer's instructions.
9. Store at 4 °C short-term, at −20 °C long-term.

3.4 Library Building and Sequencing

There are multiple commercially available kits specifically designed for constructing ChIP-seq libraries, such as the TruSeq ChIP Sample Prep Kit from Illumina (IP-202-1012), or the NEB NextChIP-Seq Library Prep Reagent Set (E6200). It is also possible to adapt genomic DNA preparation kits, with the key consideration being matching the concentration of adapter oligos to the much lower relative to a typical genomic DNA sample concentration of ChIP DNA (failure to do so tends to result in a very high fraction of reads consisting of primer dimers). This is also a consideration for making sonicated input libraries as unlike ChIP samples, which usually contain very little DNA, only a small portion (500 ng) of the input DNA should be used for library making. The library-building protocol [37] described here uses reagents commercially available outside of kits, with the exception of adapters and primers, which are to be adjusted depending on the sequencing platform and barcoding strategy used.

3.4.1 End Repair

1. Prepare end repair reaction (50 µL total volume) in a 200 µL PCR tube:

   41.5 µL sample DNA.

   5.0 µL 10× T4 DNA Ligase buffer.

   0.5 µL 10 mM dNTP mix.

   1.0 µL T4 DNA Polymerase.

   1.0 µL T4 Polynucleotide Kinase.

   1.0 µL Klenow DNA Polymerase.
2. Centrifuge briefly and incubate at 20 °C for 30 min in a thermal cycler.
3. Purify DNA using the QIAquick PCR Cleanup Kit as described above. Elute with 32 µL warm EB buffer.

3.4.2 dA Addition

1. Prepare dA addition reaction (50 μL total volume) in a 200 μL PCR tube:

   32 μL end-repaired sample DNA.

   10 μL 1 mM dATP.

   5 μL 10× NEBuffer2.

   3 μL Klenow fragment (3'→5' exo-).
2. Centrifuge briefly and incubate at 37 °C for 30 min in a thermal cycler.
3. Purify DNA using the QIAquick PCR Cleanup Kit as described above. Elute with 42 μL warm EB buffer.

3.4.3 Adapter Ligation

1. Prepare Adapter ligation reaction (50 μL total volume) in a 200 μL PCR tube:

   42 μL sample DNA from dA addition step.

   5 μL T4 DNA Ligase Buffer.

   0.5 μL Adapter oligo mix.

   0.5 μL ddH_2O.

   2 μL T4 DNA Ligase.
2. Centrifuge briefly and incubate at 20 °C for 30 min in a thermal cycler. Purify and size-select DNA with Agencourt Ampure XP beads (90 μL, or 1.8× the volume of the reaction) following manufacturer's instructions in 1.5 mL DNA low-binding tubes.
3. Mix thoroughly by pipetting up and down and incubate at RT for 5 min.
4. Place on a magnet for 2 min to separate beads from the solution. Remove the supernatant.
5. Wash twice with freshly made 70 % EtOH by incubating for 30 s at RT while on the magnet and aspirating the supernatant.
6. Remove residual EtOH by letting the tube dry for ≤5 min while being careful not to dry the beads too much.
7. Remove from magnet and add 32 μL of EB buffer. Mix well.
8. Separate from beads on the magnet and transfer to a fresh 200 μL PCR tube.

3.4.4 PCR Amplification

1. Prepare PCR mix in a 200 μL PCR tube (72 μL final volume):

   32 μL adapter-ligated sample DNA.

   36 μL Phusion DNA Polymerase Mix.

   4 μL PCR primer mix.
2. Spin down briefly and incubate in a thermal cycler using the following program:

98 °C for 30 s.

15 cycles of:

98 °C for 10 s.

65 °C for 30 s.

72 °C for 30 s.

72 °C for 5 min.

On hold at 4 °C.

3. Purify the library using Agencourt Ampure XP beads as described above (using 130 μL beads, or 1.8× the volume of the reaction). Elute in 32 μL EB buffer.

3.4.5 Library QC and Sequencing

1. Measure DNA concentration using the QuBit dsDNA HS Assay Kit following the manufacturer's instructions.
2. Evaluate the library fragment length distribution on the BioAnalyzer.

The library should show clear enrichment of inserts in the neighborhood of 200 bp and should have a sufficiently high concentration to be sequenced (usually >>2 ng/μL).

Sequencing as single-end 50-mer reads (1 × 50 mers) is usually sufficient and cost-effective for most purposes, though paired-end sequencing provides some analytical advantages [38, 39], especially when longer reads (for example, 2 × 100 mers) are generated. The optimal sequencing depth depends on the target. With the continuously increasing throughput of sequencing instruments, the economically affordable depth of sequencing keeps increasing, although eventually a saturation point is reached at which the library complexity is exhausted and further sequencing is of little benefit [40–42]. For transcription factors in mammalian genomes, 20–30 million reads per ChIP is a good target [40]; deeper sequencing is optimal for histone marks, with broad-range marks [43] benefiting particularly from increased sequencing depth [44]. Controls should be sequenced deeper than ChIP datasets.

4 Notes

1. The protocol described here uses the standard fixation condition of 1 % formaldehyde for 15 min at RT. However, many variations of fixation conditions have been tested, some of them providing better success in particular situations. These include combinations of:
   - Increasing the formaldehyde percentage. For example, fixation for ChIP in *Caenorhabditis elegans* is often performed using 2 % formaldehyde [18, 45].

- Increasing the incubation time [37, 45].
- Cross-linking at 37 °C.
- Using dual cross-linking strategies, involving the sequential use of a long-arm crosslinker and formaldehyde. Commonly used long-arm crosslinkers include ethylene glycol bis(succinimidyl succinate) (EGS), [46–48], disuccinimidylglutarate (DSG) [49], N-hydroxy-succinimide (NHS) [50], and others. Combinations of glutaraldehyde and formaldehyde can also be used.

The effect of these variations is more aggressive crosslinking, and they are typically applied to stabilize transient protein-DNA interactions or indirect interactions with DNA mediated by other proteins, by generating cross-links between proteins. For example, cross-linking 37 °C for 30 min significantly improves ChIP-seq results for the p300 transcriptional coactivator compared to the regular protocol [37], and dual cross-linking with EGS is widely used in chromatin conformation capture assays such as Hi-C and ChIA-PET [51, 52] to improve the capture of protein-protein interactions. A concern with aggressive cross-linking is that it can potentially lead to exacerbation of the known sources of artifactual signal in ChIP-seq, such as the preferential shearing of open-chromatin regions [53–55] resulting in higher background enrichment over regulatory elements [41]. These issues have not been investigated in depth. Cross-linking at 37 °C does not seem to negatively affect the quality of ChIP-seq data [37], but anecdotal evidence suggests that this can happen with some of the more aggressive crosslinking conditions, such as EGS [41]. Alternatively, cross-linking-free ChIP methods have also been described [56, 57], but those are primarily effective for assaying chromatin marks.

2. Here, the chromatin immunoprecipitation protocol is described, with the assumption that well-characterized and highly specific antibodies are used. However, the quality of antibodies is one of the major challenges in practice; in particular when no high-quality monoclonal reagents are available (there can be significant variability between the efficiency and specificity of different polyclonal lots). Ensuring the quality of the immune reagents is thus a very important part of ChIP-seq, as discussed elsewhere [40, 58].
3. A variety of methods for shearing chromatin are available. In addition to the probe sonicator, the physical methods for shearing include the Bioruptor (Diagenode) and the Covaris instruments. Enzymatic methods such as MNAse [59] or DNAse digestion have also been applied. A common consideration for physical shearing methods is maintaining the samples at low

temperature during sonication as very high temperatures can develop locally due to the energy input from the sonicator, which is undesirable as it can cause damage to the target proteins or even reversal of cross-links. The advantages of instruments such as the BioRuptor include the higher throughput and parallelization they can in principle offer and that there is no need for careful manual adjustments, such as the precise positioning of sonication tips in tubes. However, the throughput of probe sonicators can also be increased significantly—the protocol described here uses a 1/8" tip and 2×10^7 cells in a 1.5 mL microcentrifuge tube, but sonication can be efficiently carried out on higher numbers of cells in larger volumes with bigger sonication tips (for example, 1/4″). Note that sonication tips wear out with use (the erosion of the tip is usually clearly visible) and when that happens, the efficiency of sonication decreases significantly. Such tips have to be either replaced or polished. This also means that it is not advisable to blindly trust the sonication outcome even if very reliable sonication conditions have been established in the past, as the performance of the sonication tools can deteriorate with time; the fragment distribution should be characterized for each sonication batch to ensure optimal results. Finally, note that sonication of samples crosslinked under more aggressive conditions (*see* **Note 1**) is usually more difficult.

4. Low-binding tubes are preferable as the concentration of the targeted protein-DNA complexes and of the immunoprecipitated DNA can be fairly low.
5. Very large fragments might decrease the resolution of the ChIP-seq assay as only their ends are sequenced. Thus shorter fragments are preferable. On the other hand, it is also not desirable for a large fraction of fragments to be shorter than the planned sequencing read length.
6. There are two different types of controls that can be used for ChIP-seq: the sonicated input ("Input") control described here, and DNA immunoprecipitated with an irrelevant antibody, such as the secondary antibody used for the ChIP ("IgG"). There are tradeoffs between the two: IgG controls are thought to provide a better representation of, and therefore, control for the various biases inherent to the ChIP process, but can also yield libraries of lower molecular complexity [41], which is problematic on its own.
7. The protocol described here has been extensively tested and works robustly with the prescribed number of cells (2×10^7). As many cells of great biological interest are in scarce supply, there is significant incentive to reduce the number of cells used as input for ChIP-seq, and a number of low-input ChIP-Seq protocols have been developed [60–64], with varying degrees

of success. The input cell number can likely be decreased, but not to too low values without significantly impacting library complexity and the ability to exhaustively identify the target's interactomes. Notably, different targets exhibit differential sensitivity to the amount of input. As a rule, ChIP-seq targeting histone modifications is more robust to reduction in the number of input cells than ChIP-seq against transcription factors. This is most likely due to the strong, stable, and almost constant association of nucleosomes with DNA, which contrasts with the more transient interaction between DNA and transcription factors, meaning that a modified nucleosome is more likely to be crosslinked and captured in the immunoprecipitation process.

Acknowledgments

The author wishes to thank members of the Barbara Wold and Richard Myers labs and of the ENCODE Consortium for many helpful discussions, and Gilberto DeSalvo and Matthew D. Smalley for critical reading of the manuscript.

References

1. Gilmour DS, Lis JT (1984) Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. Proc Natl Acad Sci U S A 81:4275–4279
2. Gilmour DS, Lis JT (1985) In vivo interactions of RNA polymerase II with genes of *Drosophila melanogaster*. Mol Cell Biol 5:2009–2018
3. Solomon MJ, Larsen PL, Varshavsky A (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell 53: 937–947
4. Hecht A, Strahl-Bolsinger S, Grunstein M (1996) Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. Nature 383(6595):92–96
5. Ren B, Robert F, Wyrick JJ et al (2000) Genome-wide location and function of DNA binding proteins. Science 290:2306–2309
6. Iyer VR, Horak CE, Scafe CS et al (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409:533–538
7. Horak CE, Snyder M (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. Methods Enzymol 350: 469–483
8. Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28:327–334
9. Weinmann AS, Yan PS, Oberley MJ et al (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 16:235–244
10. Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
11. Mikkelsen TS, Ku M, Jaffe DB et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560
12. Robertson G, Hirst M, Bainbridge M et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657
13. Johnson DS, Mortazavi A, Myers RM et al (2007) Genome-wide mapping of in vivo

protein-DNA interactions. Science 316: 1497–1502

14. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
15. Gerstein MB, Kundaje A, Hariharan M et al (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100
16. Mouse ENCODE Consortium (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355–364
17. modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797
18. Gerstein MB, Lu ZJ, Van Nostrand EL et al (2010) Integrative analysis of the *Caenorhabditi selegans* genome by the modENCODE project. Science 330:1775–1787
19. Negre N, Brown CD, Ma L et al (2011) A cis-regulatory map of the *Drosophila* genome. Nature 471:527–531
20. Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. Nature 518:317–330
21. Kellis M, Hardison RC, Wold BJ et al (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 111:6131–6138
22. Guenther MG, Levine SS, Boyer LA et al (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130:77–88
23. May D, Blow MJ, Kaplan T et al (2011) Large-scale discovery of enhancers from human heart tissue. Nat Genet 44:89–93
24. Rada-Iglesias A, Bajpai R, Swigut T et al (2010) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470:279–283
25. Visel A, Blow MJ, Li Z et al (2009) ChIPseq accurately predicts tissue-specific activity of enhancers. Nature 457:854–858
26. Visel A, Taher L, Girgis H et al (2013) A high-resolution enhancer atlas of the developing telencephalon. Cell 152:895–908
27. Heintzman ND, Hon GC, Hawkins RD et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459:108–112
28. Heintzman ND, Stuart RK, Hon G et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311–318
29. Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T (2005) Spatial distribution of di- and trimethyl lysine 36 of histone H3 at active genes. J Biol Chem 280:17732–17736
30. Nguyen AT, Zhang Y (2011) The diverse functions of Dot1 and H3K79 methylation. Genes Dev 25:1345–1358
31. Phillips-Cremins JE, Corces VG (2013) Chromatin insulators: linking genome organization to cellular function. Mol Cell 50:461–474
32. Kyrchanova O, Georgiev P (2014) Chromatin insulators and long-distance interactions in *Drosophila*. FEBS Lett 588:8–14
33. Mortazavi A, Pepke S, Jansen C et al (2013) Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. Genome Res 23: 2136–2148
34. Hoffman MM, Ernst J, Wilder SP et al (2013) Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res 41:827–841
35. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9:215–216
36. Savic D, Gertz J, Jain P, Cooper GM, Myers RM (2013) Mapping genome-wide transcription factor binding sites in frozen tissues. Epigenetics Chromatin 6:30
37. Gasper WC, Marinov GK, Pauli-Behn F et al (2014) Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: identifying ChIP-quality p300 monoclonal antibodies. Sci Rep 4:5152
38. Chen Y, Negre N, Li Q et al (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 9:609–614
39. Wang C, Xu J, Zhang D et al (2010) An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics 11:81
40. Landt SG, Marinov GK, Kundaje A et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831
41. Marinov GK, Kundaje A, Park PJ et al (2014) Large-scale quality analysis of published ChIP-seq data. G3 (Bethesda) 4:209–223
42. Daley T, Smith AD (2013) Predicting the molecular complexity of sequencing libraries. Nat Methods 10:325–327
43. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6:S22–S32

44. Jung YL, Luquette LJ, Ho JW et al (2014) Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res 42:e74
45. Niu W, Lu ZJ, Zhong M et al (2011) Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. Genome Res 21:245–254
46. Zeng PY, Vakoc CR, Chen ZC et al (2006) In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. Biotechniques 41:694
47. Blum R, Vethantham V, Bowman C et al (2012) Genome-wide identification of enhancers in skeletal muscle: the role of MyoD1. Genes Dev 26:2763–2779
48. Law MJ, Lower KM, Voon HP et al (2010) ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell 143:367–378
49. Tian B, Yang J, Brasier AR (2012) Two-step cross-linking for analysis of protein-chromatin interactions. Methods Mol Biol 809:105–120
50. Nowak DE, Tian B, Brasier AR (2005) Two-step cross-linking method for identification of NF-κB gene network by chromatin immunoprecipitation. Biotechniques 39:715–725
51. Lin YC, Benner C, Mansson R et al (2012) Global changes in the nuclear positioning of genes and intra- and inter-domain genomic interactions that orchestrate B cell fate. Nat Immunol 13:1196–1204
52. Li G, Ruan X, Auerbach RK et al (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148:84–98
53. Auerbach RK, Euskirchen G, Rozowsky J et al (2009) Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci U S A 106:14926–14931
54. Park D, Lee Y, Bhupindersingh G, Iyer VR (2013) WidespreadmisinterpretableChIP-seq bias in yeast. PLoS One 8:e83506
55. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Natl Acad Sci U S A 110:18602–18607
56. Kasinathan S, Orsi GA, Zentner GE et al (2014) High-resolution mapping of transcription factor binding sites on native chromatin. Nat Methods 11:203–209
57. Tseng Z, Wu T, Liu Y et al (2014) Using native chromatin immunoprecipitation to interrogate histone variant protein deposition in embryonic stem cells. Methods Mol Biol 1176:11–22
58. Egelhofer TA, Minoda A, Klugman S et al (2011) An assessment of histone-modification antibody quality. Nat Struct Mol Biol 18: 91–93
59. Wal M, Pugh BF (2012) Genome-wide mapping of nucleosome positions in yeast using high-resolution MNaseChIP-Seq. Methods Enzymol 513:233–250
60. Adli M, Zhu J, Bernstein BE (2010) Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat Methods 7:615–618
61. Brind'Amour J, Liu S, Hudson M et al (2015) An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat Commun 6:6033
62. Jakobsen JS, Bagger FO, Hasemann MS et al (2015) Amplification of pico-scale DNA mediated by bacterial carrier DNA for small-cell-number transcription factor ChIP-seq. BMC Genomics 16:46
63. Gilfillan GD, Hughes T, Sheng Y et al (2012) Limitations and possibilities of low cell number ChIP-seq. BMC Genomics 13:645
64. Shankaranarayanan P, Mendoza-Parra MA, Walia M et al (2011) Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. Nat Methods 8:565–567

Chapter 2

Identification of Candidate Functional Elements in the Genome from ChIP-seq Data

Georgi K. Marinov

Abstract

ChIP-seq datasets provide a wealth of information for the identification of candidate regulatory elements in the genome. For this potential to be fully realized, methods for evaluating data quality and for distinguishing reproducible signal from technical and biological noise are necessary. Here, the computational methods for addressing these challenges developed by the ENCODE Consortium are described and the key considerations for analyzing and interpreting ChIP-seq data are discussed.

Key words Regulatory elements, Transcription factors, Histone modifications, Chromatin immunoprecipitation, High-throughput sequencing

1 Introduction

The high resolution, genome-wide coverage, and overall information richness characteristic to ChIP-seq datasets have made the assay the primary experimental tool for profiling protein-DNA interactions since its initial introduction in the second half of the 2000s [1–4]. The field has since entered maturity. A diverse array of analytical tools have been published, and much experience in working with ChIP-seq data has accumulated from the very large number of datasets that have been generated [5, 6], in particular in association with the activities of the ENCODE, modENCODE, and mouse ENCODE consortia [7–12]. These studies have helped clarify the most important characteristics of the data that need to be evaluated to ensure that biologically reliable results are obtained, and robust analytical pipelines for identifying reproducible regions of enrichment have been devised as a result. ChIP-seq data analysis can be broadly divided into two (not independent of each other) steps—data quality evaluation and identification of enriched regions (outlined in Fig. 1). The importance of quality assessment derives from the observation that there can be considerable variation in the quality of ChIP-seq datasets, sometimes even between

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_2,

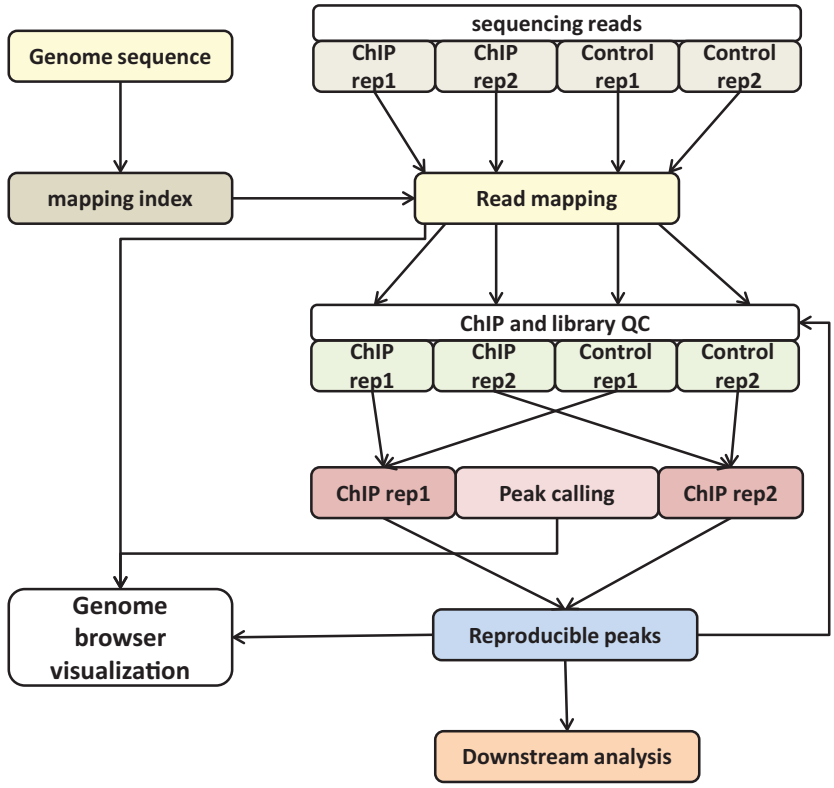


Fig. 1 Overview of the ChIP-seq analysis workflow. Reads from replicate ChIP and matching Control experiments are aligned against the genome and subjected to library and ChIP quality evaluation. Peak calling is then carried out followed by identification of consensus peaks using reproducibility analysis. The reproducibility analysis also serves as another QC step, if very low reproducibility is detected. Peaks and alignments can then be subjected to downstream analysis and/or visualized on a genome browser

what is essentially technical replicates of the same experiment. Poor-quality datasets negatively affect the ability to reliably identify occupancy regions (due to the presence of large numbers of false negatives and/or false positives), which can in turn confound biological interpretations of the data, in particular when multiple datasets are jointly analyzed (discussed in detail in [5, 6]). Low-quality ChIP-seq datasets are characterized by some combination of very low sequencing depth, low molecular complexity of the library, and low degree of enrichment of target-DNA complexes during the ChIP reaction. The latter manifests itself as the absence of strong localized read clustering (it should be noted that the presence of such clustering in control datasets, for which localized enrichment is not expected, is potentially problematic too [6]).

The main challenge when identifying regions of enrichment is distinguishing true signal from noise. It is commonly observed that regions of enrichment in ChIP-seq datasets follow an exponential-like distribution, with a small number of strong and a large number of weak sites, and no clear separation between the

very weak sites and noise. Dozens of peak calling algorithms have been published, each of them with its own specific, often tunable, approach toward thresholding. Significant differences can often be observed between their output, usually at the low end of the signal intensity spectrum (top sites tend to be reliably found by all algorithms but the inclusion of weaker sites in peak call lists is highly dependent on the particulars of the algorithms and thresholds applied). The threshold-independent IDR (Irreproducible Discovery Rate) analysis [13], which relies on the comparison of biological replicates, has recently emerged as the most robust way of identifying reproducibly occupied regions.

These computational procedures are described and discussed in this chapter.

2 Materials

The analyses described are designed to run on standard Linux systems through the UNIX command line. The maximal memory usage depends on the size of the datasets but is usually less than 15GB for typical depths of sequencing. Some of the programs used are multithreaded and will therefore complete faster if run on multiple cores.

2.1 Genomic Sequence and Annotation Files

1. A FASTA file containing the hg19 version of the human genome can be downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz. The more recent hg20 version can be found at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg38.fa.gz. The UCSC Genome Browser also has genome files for many other species. Other rich genomic resources are ENSEMBL (http://ensemblgenomes.org/) and the NCBI website (http://www.ncbi.nlm.nih.gov/assembly/).
2. "Blacklist" regions (discussed recently in [14]). "Excludable" files containing catalogs of common artifacts in hg19 coordinates can be downloaded from http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/.
3. The GENCODE annotation in GTF format: http://www.gencodegenes.org/.

2.2 Software Packages

1. Bowtie [15] (http://bowtie-bio.sourceforge.net/index.shtml) or Bowtie2 [16] (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) (*see* **Note 1** for discussion of alignment).
2. samtools [17]: http://www.htslib.org/.
3. Preseq [18]: http://smithlabresearch.org/software/preseq/.

4. FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
5. SRA Toolkit: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
6. SPP [19]: https://code.google.com/p/phantompeakqualtools/.
7. IDR [13] analysis code: https://sites.google.com/site/anshulkundaje/projects/idr.
8. UCSC Genome Browser [20, 21] utilities: http://hgdownload.cse.ucsc.edu/admin/exe/.
9. Additional scripts: https://github.com/georgimarinov/GeorgiScripts. Contains the python scripts used in the examples shown below; some of the scripts depend on having pysam installed.

3 Methods

The general outline of the pipeline is presented in Fig. 1. As an initial and one-time step, a genome index file is prepared to be used during read mapping. Reads for ChIP and Control datasets are then mapped to the genome, and the library and read clustering characteristics of each dataset are evaluated, by calculating the number and fraction of mapped reads, estimating the library complexity, and running cross-correlation analysis [5, 6]. The datasets are then subjected to reproducibility analysis (Fig. 2), which consists of:

1. Calling peaks with very relaxed settings (*see* **Notes 2** and **3**).
2. Running IDR on the individual set of peaks. This gives the number of reproducible peaks between replicates *Nt*.
3. Pooling the reads for the two ChIP replicates and separately for the two controls.
4. Creating two sets of pooled pseudoreplicates with approximately equal number of randomly sampled reads from the pooled sets of ChIP and Control reads.
5. Calling peaks with relaxed settings on each pooled pseudoreplicate.
6. Running IDR on the pooled pseudoreplicates. The number of pseudoreplicate self-consistent peaks *Np* is obtained from this step.
7. Creating two sets of individual pseudoreplicates with approximately equal number of randomly sampled reads from each ChIP and Control replicate.
8. Calling peaks with relaxed settings on each individual pseudoreplicate.

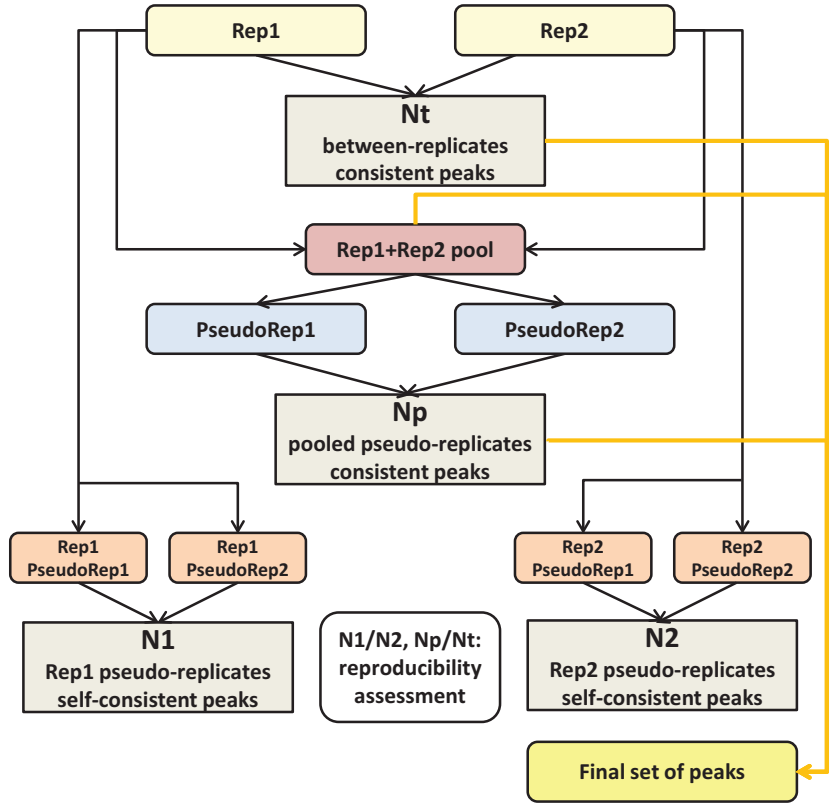


Fig. 2 Summary of the IDR-based reproducibility analysis pipeline. IDR is used to identify reproducible peaks between individual replicates (*Nt*), between pooled pseudoreplicates (*Np*), generated by pooling the two replicates and randomly splitting them in two, and between pseudoreplicates generated from each individual replicate (*N1* and *N2*). These values are used to assess reproducibility and to derive the final set of peaks (from the ranked peak calls derived from the pooled ChIP and Control datasets)

9. Running IDR on the two sets of individual pseudoreplicates. This provides the numbers of individual replicate self-consistent peaks *N1* and *N2*.
10. Checking for abnormalities in reproducibility. Specifically, when $N1/N2 > 2$, and/or $Np = Nt > 2$, the reproducibility between the two replicates is considered low [5] (*see* **Note 4**).
11. Peak calling with relaxed settings on the pooled sets of reads.
12. Using the number of reproducible peaks between replicates *Nt* and pooled pseudoreplicates *Np* to derive the final set of peaks by taking the top *max(Nt, Np)* peaks from the peak calls on the pooled sets of reads.

If more than two replicates are available, they can be analyzed in all possible pairs and the results merged at the level of the maximum *Nt* (*see* **Note 5**).

A step-by-step example using ENCODE data for the TAF1 protein (a component of the transcription preinitiation complex [22], and thus expected to be associated with promoter elements) in H1 human embryonic stem cells (H1-hESCs) is used as an illustration of the process throughout this chapter.

3.1 Preparation of Genomic Files

1. Download and uncompress genomic sequence files:

```
mkdir genomes; cd genomes; mkdir hg19; cd hg19;
mkdir sequence; cd sequence
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz
tar -xzvf chromFa.tar.gz
```

2. Create a fasta file without the "alt"/"hap" chromosomes (if starting from a larger fasta file, the fastaSubset.pyscript can be used for this purpose). The "random" contigs can also be removed (*see* **Note 6** for further discussion on this topic). Create a separate file with the Y chromosome removed if samples of known female origin are to be analyzed (the H1-hESC cells used here are male).

3. Create bowtie indexes (male version shown):

```
mkdir genomes/hg19/bowtie-indexes
cd genomes/hg19/bowtie-indexes
ln ../sequence/hg19-male.fa
bowtie-build -f hg19-male.fa hg19-male
```

   With Bowtie2:

```
mkdir genomes/hg19/bowtie2-indexes
cd genomes/hg19/bowtie2-indexes
ln ../sequence/hg19-male.fa
bowtie2-build -f hg19-male.fa hg19-male
```

4. Create chromosome size info (chrom.sizes) files (*see* **Note 7**):

```
python makeChromSizesFromFasta.py hg19-male.fa hg19-male.chrom.sizes
```

3.2 Read Mapping

1. Download reads. In this case, the following files:

```
wgEncodeHaibTfbsH1hescTaf1V0416102RawData-Rep1.fastq.gz
wgEncodeHaibTfbsH1hescTaf1V0416102RawData-Rep2.fastq.gz
wgEncodeHaibTfbsH1hescRxlchV0422111RawData-Rep1.fastq.gz
wgEncodeHaibTfbsH1hescRxlchV0422111RawData-Rep2.fastq.gz
```

   are downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/. The first

two files are the TAF1 ChIP samples, the last two are the Input Control samples.

2. Count raw reads:

```
gunzip -c SAMPLE.fastq.gz | wc -l
```

Divide by 4 to get the number of reads (each read is represented by four lines in a FASTQ file).

3. Read quality filtering (optional, *see* **Note 8**).

4. Decompress reads, trim the sequences to the desired read length (if necessary), align, convert to BAM, and sort the BAM file. This can be carried out in one step as follows:

```
gunzip -c wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.fastq.gz |
python trimfastq.py - 36 -stdout | bowtie
genomes/hg19/bowtie-indexes/hg19-male -p 16
-v 2 -k 2 -m 1 -t --best --strata -q --sam -
| samtools view -bT genomes/hg19/sequence/
hg19.fa - | samtools sort - wgEncodeHaibTfb-
sH1hescTaf1V0416102RawDataRep1.unique
```

This retains uniquely mapping reads with up to two mismatches relative to the reference.

With Bowtie2:

```
gunzip -c wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.fastq.gz | py-
thon trimfastq.py - 36 -stdout | bowtie2 -x
genomes/hg19/bowtie2-indexes/hg19-male -p
16 -t -q -U - | egrep -v 'XS:i:0' | | sam-
tools view -bT genomes/hg19/sequence/hg19.
fa - | samtools sort - wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.unique
```

This command filters out all alignments with a second equally good alignment (i.e., the equivalent of "multireads"; specified in the XS:itag).

If reads are downloaded from the Short Read Archive, they can be similarly directly streamed. The following command will print reads to standard output.

```
sratoolkit.2.4.0-1-ubuntu64/bin/fastq-
dump.2.4.0 -Z reads.sra
```

If paired-end reads are to be mapped with Bowtie1, this can be done from compressed *.gzor *bz2files as follows:

```
python PEFastqToTabDelimited.py end1.
fastq.gz end2.fastq.gz | bowtie genomes/
hg19/bowtie-indexes/hg19-male -p 16 -v 2 -k
2 -m 1 -t --best --strata -q -X 1000 --sam
--12 - | samtools view -bT genomes/hg19/se-
quence/hg19.fa - | samtools sort - sample.
PE.unique
```

With Bowtie2:

```
bowtie2 -x genomes/hg19/bowtie2-indexes/
hg19-male -1 end1.fastq.gz -2 end2.fastq.
gz -p 16 -t -X 1000 --no-mixed --no-
discordant - | egrep -v 'XS:i:0' | samtools
view -bT genomes/hg19/sequence/hg19.fa - |
samtools sort - sample.PE.unique
```

5. Index bam files with samtools:
```
samtools index wgEncodeHaibTfbsH1hescCtcf-
sc5916V0416102RawDataRep1.unique.bam
```
6. Evaluate sequencing quality with FastQC (*see* **Note 9**)
```
mkdir fastqc-wgEncodeHaibTfbsH1hescTaf1V041
6102RawDataRep1;
fastqc wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.fastq.gz -o
fastqc-wgEncodeHaibTfbsH1hescTaf1V0416102Ra
wDataRep1
```

The same procedure is carried out in parallel for all ChIP and Control samples. The resulting BAM files can be used as input for the next steps (*see* **Note 10**).

3.3 Library and ChIP/Control Quality Assessment

Library quality is evaluated here by calculating the apparent library complexity (NRF, or Non-Redundant Fraction of reads, [5, 6]), defined as:

$$NRF = \frac{\text{Number distinct unique reads}}{\text{Total number unique reads}}$$

More recently, ways to estimate the absolute molecular complexity of sequencing libraries have been developed, one example being the Preseqpackage [18]. Such knowledge is highly valuable but the NRF-based guidelines established previously (NRF∈[0:8; 1] => "high complexity," NRF < 0:5 => "low complexity" [5, 6]) are still useful for the typical sequencing depth of a ChIP-seq dataset (5×10^7reads; Fig. 3b, c).

The extent of read clustering, i.e., ChIP enrichment if the library is from a ChIP experiment, is evaluated using cross-correlation analysis [19] as described in detail previously [5, 6] based on the NSC and RSC coefficients and the cross-correlation plots (Fig. 4). The NSC and RSC coefficients are defined as follows:

$$NSC = \frac{CC(\text{fragment length})}{\min(CC)}$$

$$RSC = \frac{CC(\text{fragment length}) - \min(CC)}{CC(\text{read length}) - \min(CC)}$$

where CC refers to the cross-correlation function.

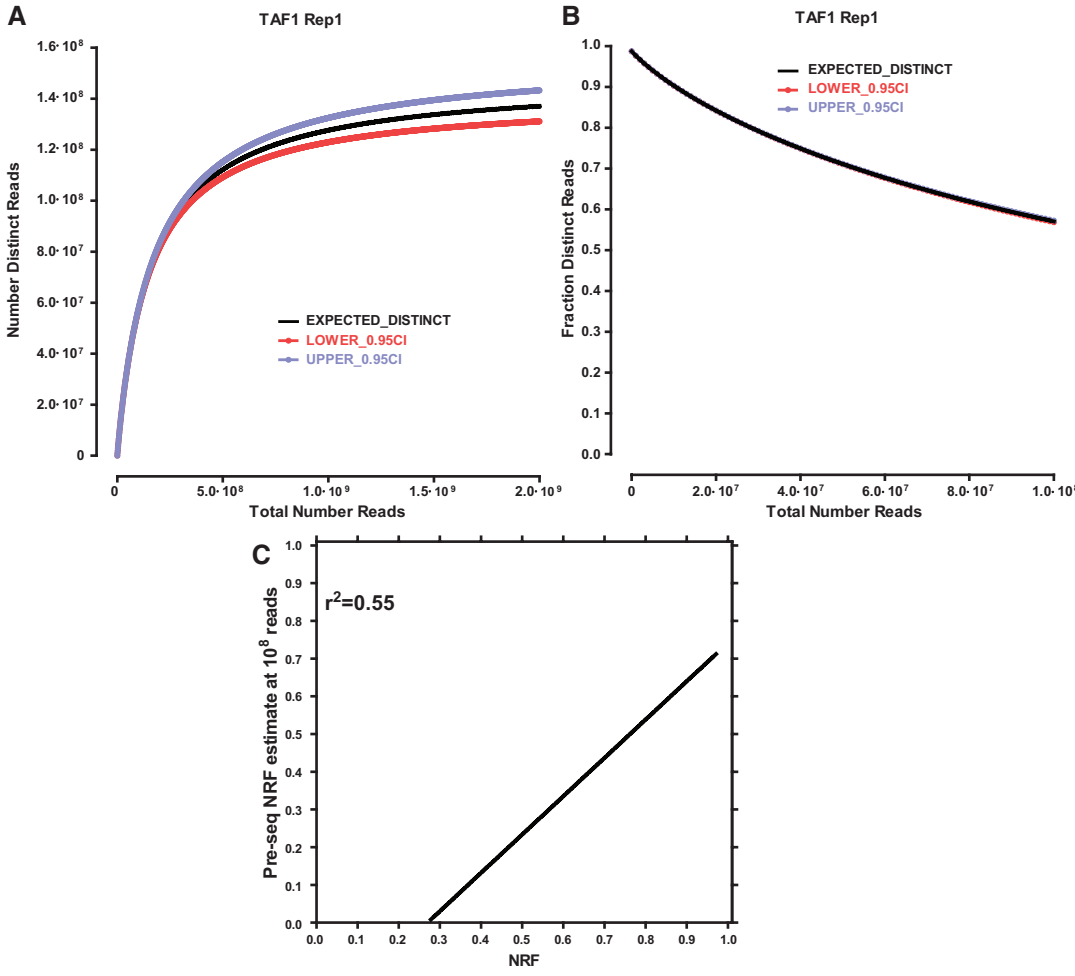
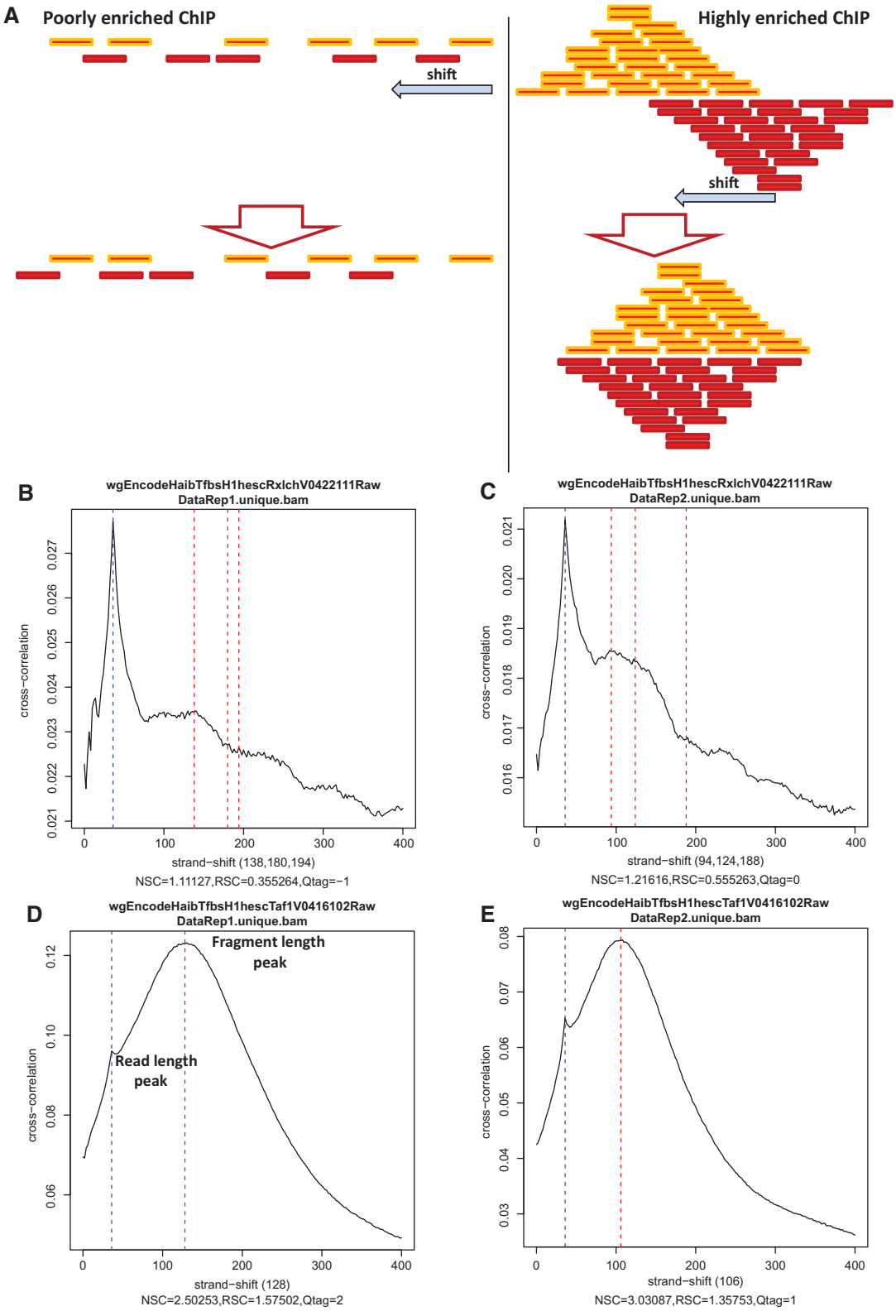


Fig. 3 Assessment of molecular complexity of libraries. (**a**) Preseq estimates for the total absolute molecular complexity of the HAIB TAF1 Rep1 dataset. (**b**) Preseq-estimated NRF fraction as a function of sequencing depth. (**c**) Correlation between empirical NRF values for all sequenced reads and Preseq-estimated NRF values at 1×10^8 reads for a subset of human ENCODE datasets (histone marks from the Broad Institute group)

See **Notes 11** and **12** for further discussion.

1. Calculate mapping statistics and apparent library complexity:
```
python SAMstats.py wgEncodeHaibTfbsH1hesc-
Ctcfsc5916V0416102RawDataRep1.unique.bam
SAMstats-H1hescCtcfscRep1 -bam genomes/hg19/
hg19.chrom.sizes samtools
```
2. Estimate the absolute library complexity using Preseq:
```
preseq-1.0.2.Linux_x86_64/preseq lc_extrap
-B wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
DataRep1.unique.bam > wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.lc_extrap
```


A
Poorly enriched ChIP
shift
Highly enriched ChIP
shift
B
wgEncodeHaibTfbsH1hescRxlchV0422111Raw
DataRep1.unique.bam
cross−correlation
strand−shift (138,180,194)
NSC=1.11127,RSC=0.355264,Qtag=−1
C
wgEncodeHaibTfbsH1hescRxlchV0422111Raw
DataRep2.unique.bam
cross−correlation
strand−shift (94,124,188)
NSC=1.21616,RSC=0.555263,Qtag=0
D
wgEncodeHaibTfbsH1hescTaf1V0416102Raw
DataRep1.unique.bam
Fragment length
peak
Read length
peak
cross−correlation
strand−shift (128)
NSC=2.50253,RSC=1.57502,Qtag=2
E
wgEncodeHaibTfbsH1hescTaf1V0416102Raw
DataRep2.unique.bam
cross−correlation
strand−shift (106)
NSC=3.03087,RSC=1.35753,Qtag=1

3. Generate cross-correlation profiles and calculate NSC and RSC scores using SPP:

```
Rscript spp_package/run_spp.R
-c=wgEncodeHaibTfbsH1hescTaf1V0416102RawData
Rep1.unique.bam -p=16 -savp -rf -s=-0:2:400
-out=wgEncodeHaibTfbsH1hescTaf1V0416102RawD
ataRep1.QC
```

Make sure to manually examine the resulting plots to identify possible oddities in the cross-correlation profiles or instances of incorrect assignment of the fragment peak (*see* **Note 12**). *See* **Note 13** for discussion on quality score cutoffs and their application.

Figures 3a, b and 4 and Table 1 show the mapping statistics, QC metrics, and the cross-correlation plots for the datasets discussed here.

3.4 Peak Calling and Identification of Reproducible Peaks

Here, peak calling is carried out with very relaxed settings (requesting the top 3×10^5 peaks) with the SPP peak caller. (*See* **Notes 3** and **14** for discussion of peak calling algorithms. Detailed technical information is also available at https://sites.google.com/site/anshulkundaje/projects/idr). The commands in the pipeline can be quickly generated automatically for large numbers of datasets using the IDRSPPcommands.pyscript.

3.4.1 IDR Analysis Pipeline

1. Create output directories for each replicate:

```
mkdir SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0
416102RawDataRep1
the pooled sample:
```

Table 1
Read mapping and dataset quality statistics for TAF1 example (Bowtie1 alignments)

Dataset	NRF	Number post-IDR peaks	post-IDR FRiP	NSC	RSC	QC	Read Length	Unique reads	Raw reads
Control-Rep1	0.96		1.111		0.355	−1	36	40,869,857	55,951,481
Control-Rep2	0.98		1.216		0.555	0	36	22,632,758	27,548,706
TAF1-Rep1	0.89	20,001	0.206	2.503	1.575	2	36	14,023,010	20,170,636
TAF1-Rep2	0.88	20,001	0.149	3.031	1.358	1	36	13,217,524	20,470,131

Fig. 4 Assessment of ChIP enrichment using cross-correlation analysis. (**a**) Overview of cross-correlation. The cross-correlation curve is generated by shifting reads on the two strands by an increasing number of base pairs relative to each other and calculating the correlation between the read profiles on the two strands. This results in two peaks—one at the read length ("phantom" peak) and one at the fragment length. The absolute height of the fragment peak and its height relative to the "phantom" peak provide information about the extent and shape of read clustering in the dataset. (**b** and **c**) Control datasets are not expected to display prominent fragment length peaks and are expected to have low NSC and RSC scores, in contrast to the high-quality ChIP-seq datasets (**d** and **e**)

```
mkdir SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0
416102RawDataRep1
the pooled pseudoreplicates:
SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0416102
RawDataPseudoRep1
SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0416102
RawDataPseudoRep2
and the individual pseudoreplicates:
mkdir SPP-300K-wgEnco
deHaibTfbsH1hescTaf1V0416102RawDataRep1-
PseudoRep1
mkdir SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0
416102RawDataRep2-PseudoRep1
mkdir SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0
416102RawDataRep1-PseudoRep2
mkdir SPP-300K-wgEncodeHaibTfbsH1hescTaf1V0
416102RawDataRep2-PseudoRep2
```

2. Run SPP on individual replicates:

```
Rscript run_spp.R -c=wgEncodeHaibTfbsH1hesc
Taf1V0416102RawDataRep1.unique.bam -i=wgEnc
odeHaibTfbsH1hescRxlchV0422111RawDataRep1.
unique.bam -p=16
-npeak=300000 -savr -savp –rf -odir=SPP-
300K-wgEncodeHaibTfbsH1hescTaf1V0416102RawD
ataRep1
```

3. Merge the BAM files for the individual ChIP and Control replicates:

```
samtools merge wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.bam
wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
DataRep1.unique.bam wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep2.unique.bam
```

4. Sort the merged BAM files:

```
samtools sort wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.bam wgEnco-
deHaibTfbsH1hescTaf1V0416102RawData.pooled.
sorted
```

5. Index the sorted BAM file:

```
samtools index wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.sorted.bam
```

6. Generate pseudoreplicates for the pooled datasets (ChIP and Control):

```
python BAMPseudoReps.py wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.sorted.bam
```

7. Generate pseudoreplicates for each individual replicate (ChIP and Control):

```
python BAMPseudoReps.py wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.unique.bam
python BAMPseudoReps.py wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep2.unique.bam
```

8. Call peaks on the pooled dataset as previously shown.
9. Call peaks on the pooled pseudoreplicates as previously shown.
10. Call peaks on individual pseudoreplicates as previously shown.
11. Uncompress peak calls:
    ```
    gunzip SPP-300K*/*gz
    ```
12. Copy IDR files and genome tables to the current working directory:
    ```
    cp idrCode/*.* .;
    cp idrCode/genome_tables/genome_table.hu-
    man.hg19.txt genome_table.txt
    ```
13. Run IDR on individual replicates:
    ```
    Rscript batch-consistency-analysis.r SPP-
    300KwgEncodeHaibTfbsH1hescTaf1V0416102RawDat
    aRep1/wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
    DataRep1.unique_VS_wgEncodeHaibTfbsH1he-
    scRxlchV0422111RawDataRep1.unique.region-
    Peak SPP-300K-wgEncodeHaibTfbsH1hescTaf
    1V0416102RawDataRep2/wgEncodeHaibTfbsH-
    1hescTaf1V0416102RawDataRep2.unique_VS_wgEn-
    codeHaibTfbsH1hescRxlchV0422111RawDataRep2.
    unique.regionPeak -1 IDR-SPP-wgEncodeHaibTf
    bsH1hescTaf1V0416102RawData 0 F signal.value
    ```
14. Run IDR on pooled pseudoreplicates as above.
15. Run IDR on individual pseudoreplicates as above.
16. Create IDR consistency plots for each replicate, pooled pseudoreplicate, and individual pseudoreplicate run as follows:
    ```
    Rscript batch-consistency-plot.r 1 IDR-SPP-
    wgEncodeHaibTfbsH1hescTaf1V0416102RawData
    IDR-SPP-wgEncodeHaibTfbsH1hescTaf1V0416102R
    awData
    ```
17. Examine the *npeaks-aboveIDR.txtand *overlapped-peaks.txt files to determine the *Nt*, *Np*, *N1*, *N2* values. Here, an IDR threshold of 0.02 is used for between-replicates self-consistency and a0.005 threshold for pseudoreplicate self-consistency (*see* **Note 15**). High *N1/N2* (where *N1* > *N2*) and *Np/Nt* values (for example, >2 [5]) indicate poor reproducibility.
    Nt:
    ```
    awk '$11 <= 0.02 {print $0}' IDR-SPP-wgE
    ncodeHaibTfbsH1hescTaf1V0416102RawData-
    overlapped-peaks.txt | wc -l
    ```

Np:

```
awk '$11 <= 0.005 {print $0}' IDR-SPP-
Pooled-pseudoreps-wgEncodeHaibTfbsH1hescTaf1
V0416102RawData-overlapped-peaks.txt | wc -l
```

N1:

```
awk '$11 <= 0.005 {print $0}' IDR-SPP-
Individual-Pseudoreps-wgEncodeHaibTfbsH1hes
cTaf1V0416102RawDataRep1-overlapped-peaks.
txt | wc -l
```

N2:

```
awk '$11 <= 0.005 {print $0}' IDR-SPP-
Individual-Pseudoreps-wgEncodeHaibTfbsH1hes
cTaf1V0416102RawDataRep2-overlapped-peaks.
txt | wc –l
```

18. Pick the top *max (Nt; Np)* peaks from the peak calls generated from the pooled ChIP and Control datasets:

```
cat SPP-300K-wgEncodeHaibTfbsH1hescTaf1V
0416102RawData.pooled/wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.sorted_VS_
wgEncodeHaibTfbsH1hescRxlchV0422111RawData.
pooled.sorted.regionPeak | sort -k7nr,7nr |
head -n 20001 | cat > SPP-300K-wgEncodeHaib
TfbsH1hescTaf1V0416102RawData.pooled/wgEnco-
deHaibTfbsH1hescTaf1V0416102RawData.pooled.
IDR0.02.regionPeak
```

Figure 5 displays the results of the IDR pipeline for the datasets discussed here.

3.4.2 Removing Known Artifacts

The resulting set of peaks is filtered against the set of known regions of artifactual enrichment (so-called blacklists):

1. Concatenate the two sets of blacklisted regions:

```
cat wgEncodeDacMapabilityConsensusExclud-
able.bed wgEncodeDukeMapabilityRegionsEx-
cludable.bed > wgEncodeBlacklists.bed
```

2. Intersect the post-IDR peaks with the blacklisted regions:

```
python regionIntersection.py SPP-300K-wgEnco
deHaibTfbsH1hescTaf1V0416102RawData.pooled/
wgEncodeHaibTfbsH1hescTaf1V0416102RawData.
pooled.IDR0.02.regionPeak 0 wgEncodeBlack-
lists.bed 0 SPP-300K-wgEncodeHaibTfbsH1hescT
af1V0416102RawData.pooled/wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.IDR0.02-vs-
Blacklist
```

The *outsersect1file contains the wanted peaks.

3.4.3 Calculating FRiP Scores

The FRiP metric is defined as the fraction of peaks falling within called regions and can be a useful complement to the

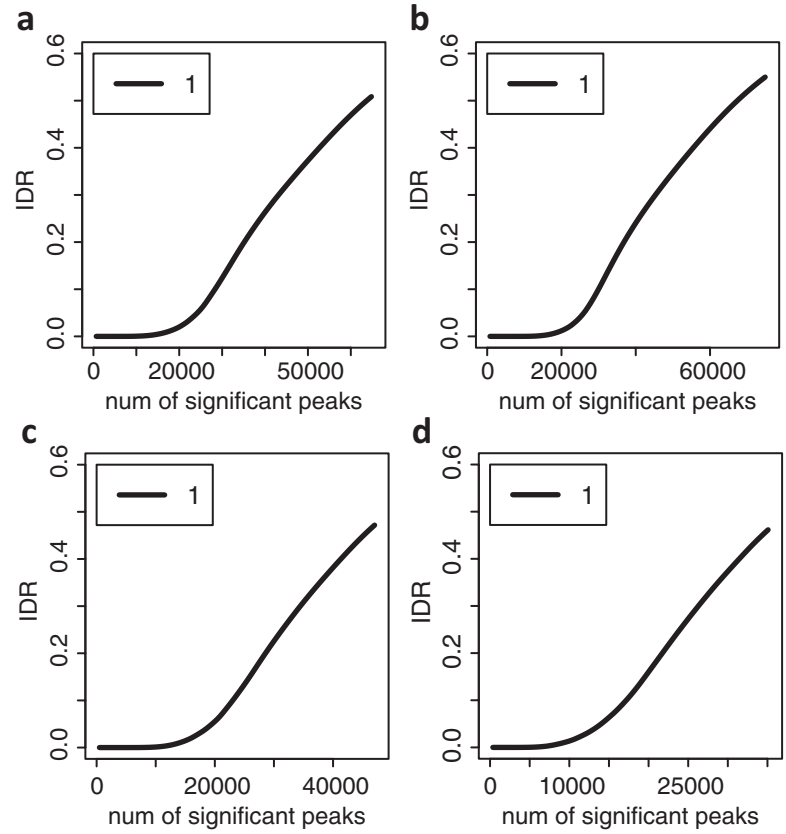


Fig. 5 Identifying reproducible peaks using IDR analysis. (**a**) Reproducible peaks between replicates (*Nt*); (**b**) Self-consistent peaks between pooled pseudoreplicates (*Np*); (**c** and **d**) Self-consistent peaks between individual replicates (*N1* and *N2*, respectively). Here, at IDR = 0.02, *Nt* = 20,001 (*Nt* = 17,367 at IDR = 0.01), and at IDR = 0.005, *Np* = 17,277, *N1* = 12,431, *N2* = 7973, *N1/N2* = 1.55 and *Nt/Np* = 1.15

cross-correlation metric even though it usually correlates well with it [5]. Here, it is calculated on the post-IDR set of peaks and for each replicate separately.

Calculate RPMs for each region and sum them:

```
python bedRPKMfromBAM.py SPP-300K-wgEncode
HaibTfbsH1hescTaf1V0416102RawData.pooled/
wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
Data.pooled.IDR0.02-vs-Blacklist-
outersection1.sorted 0 wgEncodeHaibTfb-
sH1hescTaf1V0416102RawDataRep1.unique.
bam hg19.chrom.sizes wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.IDR0.02.
Rep1.RPM -RPM -printSum –uniqueBAM
```

The sum of RPM (Reads Per Million) values over all peaks is printed at the end of the output file. The FRiP value is equal to it divided by 1×10^6.

3.5 Data Visualization

The post-IDR and post-blacklist filtering peak calls are a generally reliable starting point for subsequent analysis, but it has to be kept in mind that individually each of them is still only a candidate regulatory region and orthogonal evidence is needed to confirm its status as such and clarify its role. The precise nature of the additional evidence varies depending on the type of regulatory element. The different strains of evidence include evolutionary conservation patterns, the presence of additional types of biochemical activity around the region, and others, with the gold standard being the direct experimental testing of biological function [23]. The detailed examination of data on a genome browser is a key component of this process. Here, the UCSC Genome Browser and binary bigWig and bigBed files [21] are used for this purpose, but the plain text files generated by the pipeline are general and can be uploaded to any genome browser. The advantage of bigWig and bigBed files is that they allow fast access to the current view on the browser without having to upload the whole track; they can be placed in a location visible to the internet and loaded on the browser as custom-track, bigBed/bigwig files, to be read directly from that location when needed. Note that if a large number of datasets are to be displayed, track data hubs [24] may be a more appropriate and convenient approach for organizing them, but they are beyond the scope of this chapter.

3.5.1 Genome Brower Data Tracks

1. Generate RPM-normalized total and strand-specific coverage tracks (*see* **Notes 16** and **17**):

```
python makewigglefromBAM-NH.py track_title
wgEncodeHaibTfbsH1hescTaf1V0416102RawDa-
taRep1.unique.bam genomes/hg19/hg19.chrom.
sizes wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
DataRep1.unique.wig -uniqueBAM -RPM –notitle
python makewigglefromBAM-NH.py track_title
wgEncodeHaibTfbsH1hescTaf1V0416102RawDa-
taRep1.unique.bam genomes/hg19/hg19.chrom.
sizes wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
DataRep1.unique.plus.wig -stranded +
-uniqueBAM -RPM -notitle
python makewigglefromBAM-NH.py track_title
wgEncodeHaibTfbsH1hescTaf1V0416102RawDa-
taRep1.unique.bam genomes/hg19/hg19.chrom.
sizes wgEncodeHaibTfbsH1hescTaf1V0416102Raw-
DataRep1.unique.minus.wig -stranded -
-uniqueBAM -RPM -notitle
```

2. Convert bedGraph/wig files to bigWig:

```
UCSC-tools/wigToBigWig wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.unique.plus.wig
genomes/hg19/hg19.chrom.sizes wgEncodeHai-
```

```
bTfbsH1hescTaf1V0416102RawDataRep1.unique.
plus.bigWig
```

3. Create bigBed files containing the final post-IDR peak call set. The input file needs to be first sorted by coordinates:

```
sort -k1,2V SPP-300K-wgEncodeHaibTfbsH1hesc
Taf1V0416102RawData.pooled/wgEncodeHaibTfb-
sH1hescTaf1V0416102RawData.pooled.IDR0.02-
vs-Blacklist-outersection1 > SPP-300K-wgEnco
deHaibTfbsH1hescTaf1V0416102RawData.pooled/
wgEncodeHaibTfbsH1hescTaf1V0416102RawData.
pooled.IDR0.02-vs-Blacklist-outersection1.
sorted
UCSC-tools/bedToBigBed -type=bed6+4 SPP-
300K-wgEncodeHaibTfbsH1hescTaf1V0416
102RawData.pooled/wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.IDR0.02-
vs-Blacklist-outersection1.sorted genomes/
hg19/hg19.chrom.sizes SPP-300K-wgEncodeH
aibTfbsH1hescTaf1V0416102RawData.pooled/
wgEncodeHaibTfbsH1hescTaf1V0416102RawData.
pooled.IDR0.02-vs-Blacklist-outersection1.
bigBed
```

4. Display the tracks on the UCSC Genome Browser through the custom track-loading interface as follows.

   bigBed files:

```
track type=bigBed name=IDR-SPP-HAIB-H1hesc-
Taf1 description=IDR-SPP-HAIB-H1hesc-
Taf1 visibility=full color=255,102,102
bigDataUrl=$URL/wgEncodeHaibTfbsH-
1hescTaf1V0416102RawData.pooled.IDR0.02-
vs-Blacklist-outersection1.bigBed
```

   bigWig files:

```
track type=bigWig name=HAIB-H1hesc-Taf1-
Rep1.minus description=HAIB-H1hesc-Taf1-
Rep1.minus visibility=full color=0,128,255
bigDataUrl=$URL/wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.unique.minus.
bigWig
track type=bigWig name=HAIB-H1hesc-Taf1-
Rep1.plus description=HAIB-H1hesc-Taf1-
Rep1.plus visibility=full color=255,51,51
bigDataUrl=$URL/wgEncodeHaibTfbsH-
1hescTaf1V0416102RawDataRep1.unique.plus.
bigWig
```

The color parameter can be varied (in RGB coordinates) to adjust the color display.

3.5.2 Mappability Tracks

Mappability tracks help visualize where in the genomes reads can and cannot be uniquely mapped, and thus better understand observed patterns of read coverage, especially in repetitive regions. Various approaches of differing sophistication have been proposed for evaluating mappability [25–27]. Here, a simple remapping of tilings of the genome back to itself is described in order to generate mappability tracks for the Bowtie aligner; this approach is generally sufficient for visually figuring out the repeat structure in a given region of the genome.

1. Generate synthetic reads (in this case 36 bases long) tiling the genome at every base pair for each chromosome (each chromosome is processed individually to avoid working with extremely large BAM files):
   ```
   python mappability-make_reads.py chr1.fa
   36 - | bzip2 > chr1.36mers.fa.bz2
   ```
2. Map the synthetic reads for each chromosome against the genome with the same settings used for ChIP-seq data:
   ```
   bzip2 -cd chr1.36mers.fa.bz2 | bowtie genomes/
   hg19/bowtie-indexes/hg19-male -p 16 -v 2 -k 2
   -m 1 -t --best --strata -f --sam - | samtools
   view -bT genomes/hg19/sequence/hg19.fa - |
   samtools sort - chr1.36mers.hg19-male.unique
   ```
3. Index the resulting BAM files:
   ```
   samtools index chr1.36mers.hg19-male.unique.
   bam
   ```
4. Generate mappability tracks for each chromosome (*see* **Note 18**)
   ```
   python makewigglefromBAM-NH.py track_title
   chr1.36mers.hg19-male.unique.bam /genomes/
   hg19/hg19.chrom.sizes chr1.36mers.hg19-male.
   unique.wig -notitle -uniqueBAM
   ```
   This may result in a very small number of alignments to other chromosomes, remove them as follows:
   ```
   grep -P 'chr1\t' chr1.36mers.hg19-male.
   unique.wig > chr1.36mers.hg19-male.unique.
   filtered.wig
   ```
5. Cat the individual chromosomes and convert to bigWig.
   ```
   cat chr*.36mers.hg19-male.unique.filtered.
   wig > hg19-male.36mers.unique.mappability.
   hg19-male.wig UCSC-tools/wigToBigWig hg19-
   male.36mers.unique.mappability.hg19-male.
   wig genomes/hg19/hg19.chrom.sizes hg19-
   male.36mers.unique.mappability.hg19-male.
   bigWig
   ```

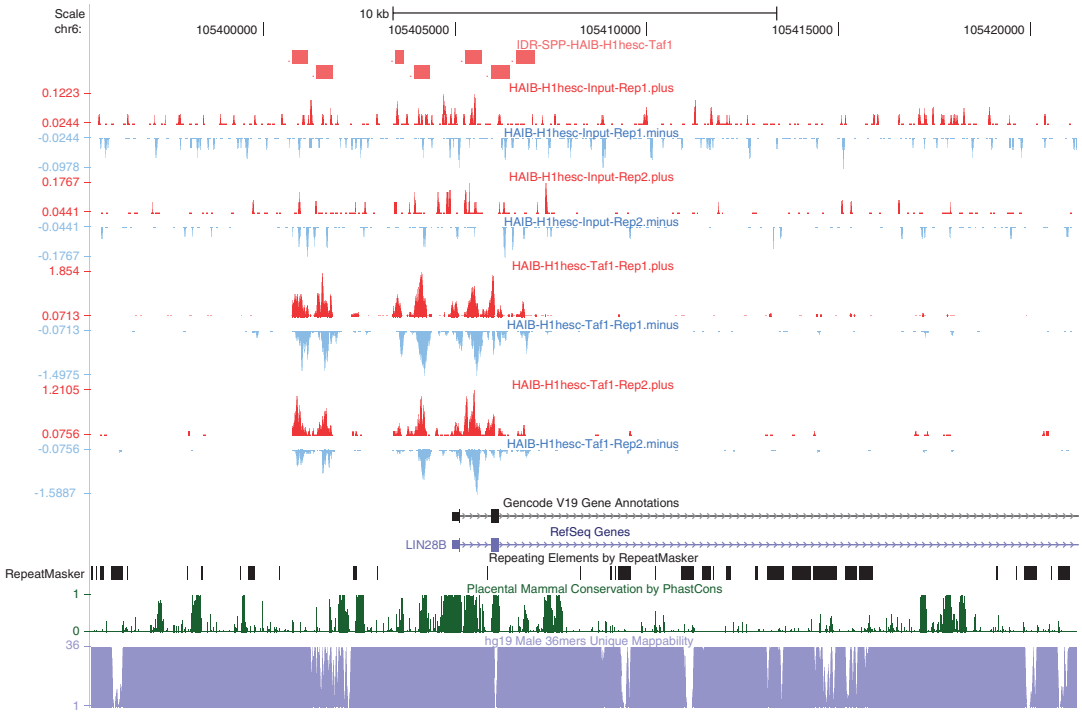


Fig. 6 Example of candidate novel regulatory elements. Shown is the 5′ region of the *LIN28B* gene (coding for an important regulator of stem cell renewal, which acts by repressing the *let-7* miRNA [28]). Multiple TAF1 peaks are observed upstream of the annotated transcription start site, representing possible alternative promoter elements

6. Display on the browser as shown before.

Figure 6 shows the region surrounding the annotated promoter of the LIN28B gene, with the post-IDR peak calls, the forward and reverse strand signal profiles for the two H1-hESC TAF1 ChIP and Control replicates, and the mappability track. Several possible alternative promoters, not present in either the refSeq or the GENCODE V19 annotations, are observed for this gene.

4 Notes

1. The choice of aligner is not absolutely crucial when ChIP-seq data is analyzed with the goal of identifying occupancy sites (it is of more significant importance when sequence polymorphisms in the data need to analyzed). The pipeline presented here is based on Bowtie1, but Bowtie2 [16] (also shown), BWA [29], as well as other aligners are also often used with mostly equally good results. The requirements for the alignments are that only unique alignments are retained (properly dealing with multimappers is a complicated subject beyond the

scope of this chapter) and soft-clipped alignments are not allowed (the latter is not an absolute requirement, but many of the steps described do assume a constant read length and it is preferable that the alignments conform to that assumption).

2. The underlying principle of IDR analysis is the separation of the two sets of features that are being compared into reproducible and irreproducible noise components. In order for this to be accomplished, there has to be a significant noise component present in the data. This requirement is opposed to the goal of the default settings of peak calling algorithms, which is to produce a set of peaks as devoid of noise as possible. It is therefore necessary to carry out peak calling with greatly relaxed relative to the default settings so that the noise component is present.
3. In this chapter, peak calling using SPP [19] is described; however, a number of other peak callers have also been successfully used with IDR (MACS2 [30], PeakSeq [31], GEM [32], and others). The applicability of peak callers to IDR analysis is determined by whether they can be run in the "overcall" mode required for the IDR algorithm to be able to separate the noise from reproducible components, and by how they rank individual peaks (for example, it is preferable that peak callers do not generate very large numbers of ties in rankings). The detailed parameters and potential output issues need to be optimized and characterized for each peak caller individually.
4. The reasons for poor reproducibility can be varied. A typical situation involves a pair of a strong and weak ChIP-seq datasets, in which the IDR analysis is dominated by the weak replicate. Alternatively, the experiments may not have been carried out under properly matched biological conditions, there might have been differences in the lots of antibodies used to carry out the ChIPs, etc. Such cases are cause for concern, should be investigated in detail, and the experiments repeated, if practically possible.
5. If more than two replicates are available, the final set of peaks can be derived by finding the maximum number of reproducible peaks *Nt* between all pairs of replicates and by using it to threshold on the peaks called on the pooled sets of reads.
6. The current human and mouse genome assemblies contain a number of "random" chromosomes, which represent sequences known to exist in the genome but not yet properly placed within the chromosomes. These are often not considered during alignment (for example, this was the policy of the ENCODE Consortium [7]). The hg19 and hg20 versions of the human genome also contain a number of "alt"/"hap" chromosomes, which represent alternative sequences for certain regions of the

genome that are commonly found in the population. Both the "random" and the "alt"/"hap" chromosomes are often places where reads can be aligned to. This includes both reads that would otherwise not align at all and reads that in their absence would erroneously align elsewhere (*see* [33] for a detailed discussion of the current status and significance of "alt"/"hap" segments). However, if the "alt"/"hap" sequences are included in the index, they also have the effect of erroneously making otherwise uniquely mappable reads multimappers, eliminating them from subsequent analysis. This is not the case with the random contigs, thus the optimal policy is to include the "random" chromosome and exclude the "alt"/"hap" chromosomes. When the sex of the source of the biological material assayed is known, it is preferable to align against the corresponding male or female version of the genome.

7. The chrom.sizes files are used to indicate the identity and end points of chromosomes to a number of programs. They are in the following format:
   `chr<tab> chromosome_size`
8. Removal and/or trimming of low-quality reads/bases from reads is essential for a number of genomic applications. It is not strictly necessary for ChIP-seq if the goal is to simply call peaks and the overall sequencing quality is as good as it usually is. It becomes important if additional information, such as allelic bias in occupancy, is to be extracted. In addition, it is preferable that the read length is kept constant, *see* also **Note 1**.
9. If most of the reads mapped (typically more than 80% of ChIP-seq reads align successfully to a mammalian genome, including the ones whose alignments are suppressed due to high multiplicity), then most likely there are no issues with read quality. While FastQC is very useful for spotting issues with reads in general, its application is particularly helpful in the cases when few reads align, due to issues such as general drop-offs in read quality after certain sequencing cycle, the presence of adapter and barcode sequences, and others, information about which is provided by FastQC. Identification of these problems can help rescue datasets by trimming reads accordingly prior to alignment (the trimfastq.py provides various utilities for trimming reads from either the 5′ or 3′ end).
10. It is a standard practice for a number of genomic applications to collapse apparent duplicate reads/fragments. Such groups of reads/fragments may represent potential PCR duplicates, but they can also be distinct but identical in sequence molecular fragments (the latter is more likely with short single-end reads). Such "dedup"-ing is sometimes performed on ChIP-seq datasets. However, as first, it is not necessary for the peak

calling procedures described here, and second, it defeats the purpose of the molecular complexity characterization steps if it is applied before them; here its application is advised against.

11. Cross-correlation analysis is based on the clustering of unpaired reads on opposite strands around sites in the genome enriched in the dataset [5, 6]. The assumption that the reads are unpaired is critical—if cross-correlation is run on a paired-end dataset, results will be meaningless as there will always be a paired read on the other strand at a distance in the neighborhood of the average fragment length. Another consideration to keep in mind is that if long read lengths are generated (e.g., 100 mers), the average fragment length can be very close to, or even shorter than the read length, in which case there will be no separation of the "phantom" peak and the fragment length peak. For these reasons, it is recommended that this step is carried out on single-end 36 bp or 50 bp alignments. Finally, the cross-correlation metrics have been calibrated for genomes with size and repeat structure similar to those of mammals. The baseline profile (i.e., what should be observed in input samples) can look very different in organisms whose genomes significantly deviate from these characteristics, and the metrics are not directly applicable in exactly the same form in such cases.
12. Cross-correlation analysis is carried out on both ChIP and Control samples. In ChIP samples, prominent fragment peaks and high values of the NSC and RSC metrics are desired; however, there are some caveats to be aware of as they can in fact be artifactual. For example, sometimes very high such peaks are observed in both the ChIP and the Control, suggesting that they originate from sources other than real ChIP enrichment. Such cases illustrate the need to carry out cross-correlation analysis of both ChIP and Control samples—in Control samples, prominent cross-correlation peaks are not expected, and if they are observed, this can be a sign that the enrichment patterns in the ChIP sample might be generated by a mixture of real ChIP signal and another source of enrichment, the latter being artifactual [6].
13. In prior studies [5, 6] cutoffs based on the RSC metric have been applied to divide datasets into high-, intermediate, and low-quality groups, for example: RSC $\in (0; 0.25)$ => QC = −2, RSC $\in (0.25; 0.5)$ => QC = −1, RSC $\in (0.5; 1)$ => QC = 0, RSC $\in (1;1:5)$ => QC = +1, RSC > 1.5 => QC = +2, with −2 corresponding to minimal read clustering and 2 to a highly clustered library (this is also the discretization used here). These are useful guidelines but are not on their own absolutely trustworthy metrics to be blindly used, first, because they represent discretizations of inherently continuous variables, and second, because various factors can lead to very high or very low scores

in an otherwise poor or high-quality dataset, respectively (*see* the issues discussed in **Notes 12** and **11**). For such reasons, there is no substitute for the manual examination of cross-correlation plots when evaluating libraries.

14. There are three different types of ChIP-seq datasets: "point-source," "broad-range," and mixed. Point-source datasets are best exemplified by sequence-specific factors, which bind to very precisely defined and short in length (in most cases 6–10 bp) sequences in the genome, generating the classic asymmetric read distribution pattern around binding sites. Most peak callers are tuned to find regions of enrichment with point-source characteristics. In contrast, the computational definition of "broad-range" regions of enrichment is more challenging. Such ChIP-seq datasets are classically observed when targeting histone modifications associated with transcriptional elongation (H3K36me3, H3K79me2) and with broad repression domains (H3K27me3, H3K9me3). Mixed-source datasets contain both regions with point-source and broad-range characteristics (the main representative of this group being RNA polymerase II). The pipeline described here is focused on the characterization of point-source enrichment regions. Fortunately, this characterizes most well-defined candidate regulatory regions (histone marks associated with active enhancers and promoters often approximate the point-source patterns, insulator proteins bind in a very localized manner, and sequence-specific transcription factor binding is the hallmark of enhancer and promoter elements).
15. With the SPP settings used here and mammalian-sized genomes, an IDR = 0.02 cutoff works fine, but the threshold can be increased or decreased if needed depending on the peak caller used (for example, if a smaller number of peaks are used as input to the IDR analysis) and the nature of the dataset. A more stringent threshold is applied for pseudoreplicates as the pooling and splitting process naturally leads to higher levels of reproducibility.
16. It is preferable to separate the forward and reverse strands when displaying ChIP-seq data as the patterns on the two strands provide important information about the nature of the observed enrichment patterns. Regions of artifactual enrichment tend to lack the strand asymmetry that characterizes classical occupancy sites, a difference that can only be readily identified when the signal profiles on the two strands are examined in parallel.
17. RPM normalization is important for the direct comparison of different datasets. Non-normalized tracks with the *y* axis corresponding to the total number of reads cannot be directly

compared to each other if the sequencing depth of the two datasets is significantly different.

18. Note that the mappability track generated as shown here has a range of scores $s \in [0;RL]$, where RL is the read length.

Acknowledgments

The author wishes to thank Anshul Kundaje, members of the Barbara Wold and Richard Myers labs and of the ENCODE Consortium for many helpful discussions, and Gilberto DeSalvo and Matthew D. Smalley for critical reading of the manuscript.

References

1. Shyh-Chang N, Daley GQ (2013) Lin28: primal regulator of growth and metabolism in stem cells. Cell Stem Cell 12:395–406
2. Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
3. Johnson DS, Mortazavi A, Myers RM et al (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502
4. Mikkelsen TS, Ku M, Jaffe DB et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560
5. Robertson G, Hirst M, Bainbridge M et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657
6. Landt SG, Marinov GK, Kundaje A et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831
7. Marinov GK, Kundaje A, Park PJ et al (2014) Large-scale quality analysis of published ChIP-seq data. G3 (Bethesda) 4:209–223
8. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
9. Gerstein MB, Kundaje A, Hariharan M et al (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100
10. Gerstein MB, Lu ZJ, Van Nostrand EL et al (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330:1775–1787
11. modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797
12. The Mouse ENCODE Consortium (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355–364
13. Negre N, Brown CD, Ma L et al (2011) A cis-regulatory map of the *Drosophila* genome. Nature 471:527–531
14. Li Q, Brown J, Huang H et al (2011) Measuring reproducibility of high-throughput experiments. Ann Appl Stat 5:1752–1779
15. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I (2014) Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet 5:75
16. 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
17. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
18. H L, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079
19. Daley T, Smith AD (2013) Predicting the molecular complexity of sequencing libraries. Nat Methods 10:325–327
20. Feng J, Liu T, Qin B et al (2012) Identifying ChIP-seq enrichment using MACS. Nat Protoc 7:1728–1740
21. Kuhn RM, Haussler D, Kent WJ (2013) The UCSC genome browser and associated tools. *Brief Bioinform* 14:144–161
22. Kent WJ, Zweig AS, Barber G et al (2010) BigWig and BigBed: enabling browsing of

large distributed datasets. Bioinformatics 26:2204–2207

23. Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol 41:105–178
24. Kellis M, Hardison RC, Wold BJ et al (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* 111:6131–6138
25. Raney BJ, Dreszer TR, Barber GP et al (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics 30: 1003–1005
26. Koehler R, Issac H, Cloonan N, Grimmond SM (2011) The uniqueome: a mappability resource for short-tag sequencing. Bioinformatics 27:272–274
27. Lee H, Schatz MC (2012) Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28:2097–2105
28. Derrien T, Estell´e J, Marco Sola S et al (2012) Fast computation and applications of genome mappability. PLoS One 7:e30377
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
30. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351–1359
31. Rozowsky J, Euskirchen G, Auerbach R et al (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27:66–75
32. Guo Y, Mahony S, Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol 8:e1002638
33. Church DM, Schneider VA, Steinberg KM et al (2015) Extending reference assembly models. Genome Biol 16:13

Chapter 3

GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression

Rui Lopes*, Reuven Agami, and Gozde Korkmaz*

Abstract

The advent of next-generation sequencing (NGS) technologies has revolutionized the way we do research on gene expression. High-throughput transcriptomics became possible with the development of microarray technology, but its widespread application only occurred after the emergence of massive parallel sequencing. Especially, RNA sequencing (RNA-seq) has greatly increased our knowledge about the genome and led to the identification and annotation of novel classes of RNAs in different species. However, RNA-seq measures the steady-state level of a given RNA, which is the equilibrium between transcription, processing, and degradation. In recent years, a number of dedicated RNA-seq technologies were developed to measure specifically transcription events. Global run-on sequencing (GRO-seq) is the most widely used method to measure nascent RNA, and in recent years, it has been applied successfully to study the function and mechanism of action of noncoding RNAs. Here, we describe a detailed protocol of GRO-seq that can be readily applied to investigate different aspects of RNA biology in human cells.

Key words Sequencing, GRO-seq, Nascent transcription, Noncoding RNA, Enhancer, Running Head: GRO-seq

1 Introduction

RNA analysis was once limited to study individual molecules by Northern blot or real-time quantitative PCR. In the past decade, we witnessed a revolution in the RNA world caused by the rise of NGS. In particular, RNA-seq contributed decisively to make global gene expression analysis as a routine practice in many labs [1]. This technique does not only allow to qualitatively and quantitatively investigate messenger RNAs (mRNAs), but also novel noncoding RNA species (e.g., microRNAs, small interfering RNAs, and long noncoding RNAs) that are recognized nowadays as important players in the biology of the cell [2].

*These authors contributed equally to this work.

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_3,

The cellular levels of a given RNA molecule are determined by the interplay of transcription, processing, and degradation. Consequently, both transcriptional and posttranscriptional changes affect RNA levels as measured by RNA-seq. In recent years, a number of RNA-seq-based technologies were developed that can measure specifically nascent RNA transcription from actively engaged polymerases. Among them, the most widely used are global run-on sequencing (GRO-seq) [3], native elongating transcript sequencing (NET-seq) [4], and precision nuclear run-on sequencing (PRO-seq) [5]. Here, we focus on GRO-seq, an assay that allows mapping and quantification of transcriptionally engaged RNA polymerases and provides a snapshot of genome-wide transcription. GRO-seq measurements are very sensitive and largely independent of RNA stability effects, making it particularly suitable to study noncoding RNA species that are lowly expressed and/or have high decay rate.

Recently, we and others have used GRO-seq to study different classes of noncoding RNAs such as promoter-associated RNAs [3], enhancer-associated RNAs [6], and long noncoding RNAs [7]. Below, we describe a detailed protocol of GRO-seq (Fig. 1), based on the original protocol from Core et al., Science (2008), which can be readily applied to study different aspects of RNA biology in human cells [8–10].

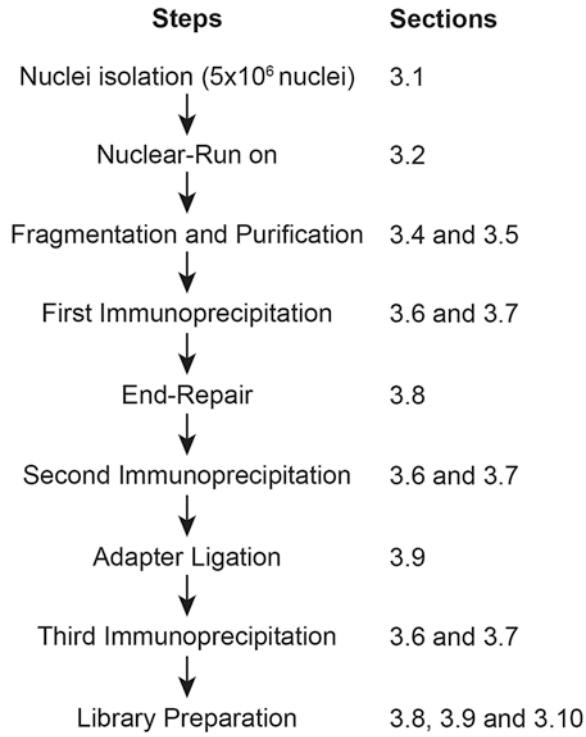


Fig. 1 Summarized scheme of the GRO-seq protocol steps (**left panel**) and corresponding sections (**right panel**) described in this article

2 Materials

All solutions for the protocol must be freshly prepared. In order to reduce the risk of RNase activity, it is necessary to use nuclease-free water for the preparation of all solutions, as well as the addition of RNase inhibitors to all RNA reactions.

2.1 Isolation of Nuclei

Buffers should be ice-cold during the protocol.

1. UltraPure DNase/RNase-Free Distilled Water.
2. PBS (DPBS, no calcium, no magnesium).
3. 1 M $MgCl_2$.
4. 0.5 M EDTA, pH 8.0.
5. UltraPure 1 M Tris–HCI, pH 8.0.
6. UltraPure 1 M Tris–HCI Buffer, pH 7.5.
7. Tris/HCl pH 7.8: mix 1:1 volume/volume ratio of Tris/HCl, pH 7.5 and Tris/HCl, pH 8.
8. RNase Inhibitor.
9. Swelling Buffer: 10 mM Tris/HCl pH 7.5, 2 mM $MgCl_2$ 3 mM $CaCl_2$. Add 2 U/ml RNase Inhibitor immediately before use.
10. Igepal CA-630.
11. Glycerol.
12. Lysis Buffer: 10 mM Tris–HCl pH 7.5, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5 % (v/v) IGEPAL CA-630, 10 % (v/v) glycerol+ 2 U/ml RNase Inhibitor.
13. Freezing buffer: 40 % (v/v) glycerol, 5 mM $MgCl_2$, EDTA pH 8.0, 50 mM Tris/HCl pH 7.8, add 2 U/ml of SUPERase-In immediately before use.
14. Trypan blue solution.
15. Hemocytometer.
16. Liquid nitrogen.

2.2 NRO Reaction

1. 5-bromouridine 5′-triphosphate (BrUTP).
2. Sarkosyl 20 % (*N*-lauryl sarcosine sodium sulfate).
3. ATP Solution (10 mM).
4. CTP Solution (10 mM).
5. GTP Solution (10 mM).
6. KCl (2 M).

2.3 RNA Extraction

1. TRIzol LS (Invitrogen).
2. Chloroform.

3. GlycoBlue Coprecipitant (15 mg/mL) (Ambion).
4. 2-Propanol.

2.4 Fragmentation of NRO-RNA

1. RNA Fragmentation Reagents (Ambion).

2.5 Purification of Fragmented NRO-RNA Through p-30 RNase-free Spin Column

1. Micro Bio-Spin Columns with Bio-Gel P-30 in Tris Buffer (Bio-Rad).
2. TE buffer, pH 8.0.

2.6 Blocking the BrdU Beads for Immunoprecipitation

Add 2 μl of 20 U/μL RNase Inhibitor for each 10 ml of the buffers.

1. BrdU (IIB5) AC (Santa Cruz Biotech, catalogue number sc-32323 AC) (agarose conjugated BrdU antibody for IP studies).
2. UltraPure SSPE, 20×.
3. Tween 20.
4. UltraPure BSA (50 mg/ml).
5. PVP (Polyvinylpyrrolidone solution).
6. Binding Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % Tween-20, 37.5 mM NaCl.
7. Blocking Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % Tween-20, 37.5 mM NaCl, 0.1 % PVP (final concentration), 1 μg/ml BSA.

2.7 Immuno precipitation of NRO-RNA

Add 2 μl of 20 U/μL RNase Inhibitor for each 10 ml of the buffers.

1. PVP 40.000.
2. DTT.
3. 10 % SDS.
4. Low Salt Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % (v/v) Tween-20.
5. High Salt Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % (v/v) Tween-20, 100 mM NaCl.
6. TET Buffer: 1× TE buffer, 0.5 % (v/v) Tween-20.
7. Elution Buffer: 0.15 M NaCl, 0.05 M Tris pH 7.5, 0.001 M EDTA pH 8.0, 0.1 % SDS. Add 0.02 M DTT immediately before use.

2.8 End Repair of NRO-RNA (TAP/PNK Treatment)

1. Tobacco Acid Pyrophosphatase (TAP).
2. T4 Polynucleotide Kinase (PNK).

2.9 GRO-seq Library Preparation

1. TruSeq Small RNA Library Preparation Kit (Illumina).
2. Agencourt AMPure XP (Beckman Coulter).
3. RNA 6000 Ladder (Ambion).
4. Agilent DNA 1000 or Agilent High Sensitivity DNA chip.

3 Methods

3.1 Isolation of Nuclei

Perform all steps on ice or at +4 °C. Volumes below are described for 15 cm^2 plate.

1. Wash cells three times with ice-cold PBS.
2. Aspirate the PBS.
3. Add 10 ml of ice-cold Swelling Buffer to the plate, and incubate for 5 min on ice (*see* **Note 1**).
4. Scrape cells into the solution and transfer the solution into 15 ml tubes. Pellet cells by centrifuging at 400 × *g* (400 × *g*, GH3.8 rotor) for 10 min at +4 °C.
5. Aspirate supernatant (SN) and resuspend cells in 500 μl of Swelling Buffer/10 % glycerol solution containing 4 U/ml RNASe inhibitor by gentle pipetting.
6. Vortex cells slowly (≈800 rpm, so that liquid rises about 1–2 cm) and drop wise add 500 μl of Swelling Buffer/10 % glycerol/1 % IGEPAL CA-630 solution containing 4 U/ml RNASe inhibitor.
7. Incubate cells on ice for 5 min.
8. Bring volume to 10 ml with Lysis Buffer and centrifuge at 600 × *g* (600 × *g*) for 5 min at +4 °C.
9. Aspirate SN.
10. Wash nuclei with 10 ml of Lysis Buffer (*see* **Note 2**) and centrifuge at 600 × *g* (1550 rpm) for 5 min at +4 °C.
11. Aspirate the SN and resuspend the pellet in 1 ml of Freezing Buffer.
12. Mix 10 μl of Freezing buffer containing isolated nuclei with 190 μl Trypan Blue (2:5 diluted in freezing buffer, so that the nuclei will not swell) and count the cells by using hemocytometer.
13. Pellet nuclei at 900 × *g* (900 × *g*) for 6 min at +4 °C and aspirate the SN without disturbing the pellet.
14. Resuspend the pellet to have 5 × 10^6 nuclei per 100 μl with Freezing Buffer. Aliquots every 100 μl into individual 1.5 ml tube.
15. Snap-freeze nuclei in dry ice/methanol or liquid nitrogen or proceed directly to the nuclear run-on reaction (NRO-rxn) (*see* **Note 3**).

Table 1
Content of the NRO master mix (2×) for NRO Reaction

Reagent	Stock[M]	Final[mM]	500 µl
Tris–Cl	1	10	5
$MgCl_2$	1	5	2.5
DTT	0.1	1	5
KCl	2	300	75
ATP	0.01	0.5	25
GTP	0.01	0.5	25
CTP	0.0001	0.002	10
^{32}P-CTP			50
BrUTP	0.01	0.5	25
RNASe inhibitor	20 U/µl	0.4	10
2 % Sarkosyl	2	1	250
Water			17.5
			482.50
			500

3.2 NRO Reaction

1. Prepare the NRO master mix (2×) (Table 1) containing 1.165 µM final concentration of CTP (*see* **Note 4**).
2. Preheat the NRO-mix to +30 °C.
3. Mix 100 µl of pre-warmed NRO-mix with 100 µl nuclei (1:1 ratio) and place the reaction into the heat block (at +30 °C) to perform run-on for 5 min with shaking at the 2nd and 4th min (600–800 rpm) (*see* **Note 5**).
4. Immediately add 900–1000 µl of TRIzol LS and vortex solution for 30 s. This will stop the NRO-rxn (*see* **Note 6**).
5. Incubate for 5 min at RT.
6. It is advisable to keep the samples at −20 °C for short-term storage or at −80 °C for longer storage. OPTIONAL STOPPING POINT.

3.3 RNA Extraction

1. Add 240 µl of Chloroform to the 900–1000 µl of TRIzol LS and shake by hand for 15 s.
2. Incubate for 2–3 min at RT and centrifuge at 12,000 × *g* for 15 min at +4 °C.
3. Take out the colorless upper layer containing RNA (middle phase contains DNA, red phase contains proteins) and add 2 µl of glycogen and same volume of isopropanol as sample and inverse tube ten times (*see* **Note 7**).

4. Incubate for 10 min at RT and centrifuge at 14,000 × *g* for 15 min at +4 °C.
5. Remove supernatant and wash the pellet twice with 75 % EtOH (prepared with DNase, RNase free water), vortex, and centrifuge at 7500 × *g* for 5 min at +4 °C.
6. Remove supernatant, spin and remove the remaining of supernatant. Air-dry the pellet for 1 min at RT and be careful not to over-dry the pellet.
7. Suspend the pellet in 10 μl of DNase, RNase free water containing RNASe inhibitor enzyme (1 U/μl).
8. Incubate the solution for 5 min at 65 °C.

3.4 Fragmentation of NRO-RNA

1. Add 2 μl of fragmentation reagents up to 20 μl NRO-RNA solution and incubate for 10 min at +70 °C (*see* **Note 8**).
2. Add 2 μl of Stop solution and put on ice.

3.5 Purification of Fragmented NRO-RNA Through p-30 RNase-free Spin Column

1. In order to prepare column, invert it several times to resuspend the matrix. Spin at 1000 × *g* for 2 min to remove the flow through. Put column into a new 1.5 ml tube. Do not let the column dry out, prepare and use immediately.
2. Add 500 μl of TE buffer and centrifuge for 1 min at 1000 × *g*. Discard the flow through. Repeat this step for two to three times to change the buffer content.
3. Add sample (between 20 and 100 μl) into the column and centrifuge for 4 min at 1000 × *g*. Keep the flow through which contains fragmented NRO-RNA.
4. Check the radioactivity level with the Geiger counter. The flow through should have more counts than the matrix. If the matrix has more counts, it means that there is unincorporated ^{32}P-CTP. Thus, NRO-rxn was not successful.
5. Sample can be frozen at −80 °C. OPTIONAL STOPPING POINT.

3.6 Blocking the BrdU Beads for Immunoprecipitation

1. Equilibrate BrdU beads in binding buffer by washing them two times in 500 μl for 5 min with rotation (8 rpm). Spin down beads at 1000 × *g* for 1–2 min. Place on ice for 1 min before removing SN (*see* **Note 9**).
2. Block the beads in four to five times volume of blocking buffer for 1–2 h at RT. Add an extra 2 μl RNASe inhibitor for every ml of blocking buffer during this step.
3. Spin down the beads (1000 × *g* for 2 min) and discard the supernatant.
4. Wash beads twice in 500 μl binding buffer.

5. Spin down the beads (1000 × *g* for 2 min) and discard the supernatant.
6. Resuspend beads in 400 µl binding buffer/reaction.

3.7 Immuno precipitation (IP) of NRO-RNA (First IP)

1. Bring purified and fragmented NRO-RNA volume to 100 µl and add EDTA to reach a final concentration of 5 mM (add 1 µl of 0.5 M EDTA).
2. Heat the sample to +65–70 °C for 5 min and then place on ice for 2 min (*see* **Note 10**).
3. Add 400 µl of binding buffer and 50–60 µl of bead slurry into a 1.5 ml tube. Allow binding for +30–60 min (*see* **Note 11**).
4. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
5. Wash one time in 500 µl of binding buffer for 5 min on rotating stand (8 rpm).
6. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
7. Wash one time in 500 µl of low salt buffer for 5 min on rotating stand (8 rpm).
8. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
9. Wash one time in 500 µl of high salt buffer for 3 min on rotating stand (8 rpm).
10. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
11. Wash two times in 500 µl of TET buffer for 5 min on rotating stand (8 rpm).
12. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
13. Elute two times with 125 µl and one time with 250 µl of elution buffer that is heated to +42 °C. Place the tube in a heat block (+42 °C) for 10 min with constant shaking at 500 rpm and every few minutes increase the shaking to 900 rpm.
14. Spin down the beads and transfer the solution (eluate) containing the NRO-RNA to a new tube.
15. Add 900–1000 µl of TRIzol LS and follow Subheading 3.3.

3.8 End Repair of NRO-RNA (TAP/PNK Treatment)

1. Heat RNA to +65–70 °C for 5 min, put on ice for 2 min.
2. Set up the following reaction in a 1.5 ml tube: 3 µl of 10× TAP buffer, 5 µl of nuclease-free water, 1 µl of RNASe inhibitor, 1.5 µl of TAP (10 U/ml).
3. Mix by pipetting and incubate the reaction at +37 °C for 1.5 h.
4. Add 1 µl of PNK, and 1 µl of 300 mM $MgCl_2$ to reach 10 mM final concentration, mix by pipetting, and incubate the reaction another 15 min.
5. Add 20 µl of PNK buffer, and 126 µl of nuclease-free water, 1 µl of RNAse inhibitor (20 U/µL), and another 1 µl of PNK.

6. Mix by pipetting and incubate the reaction for 15 min.
7. Add 20 μl of 10 mM ATP and another 1 μl of PNK, mix by pipetting, and incubate the reaction for 30 min.
8. Stop the reaction by adding 77 μl of nuclease-free water, 18 μl of 5 M NaCl, and 5 μl of 500 mM EDTA.
9. Add 900–1000 μl of TRIzol LS and follow Subheading 3.3.

3.9 Second IP, Adapter Ligation, and Third IP

1. Repeat Subheadings 3.6 and 3.7.
2. Follow the manufacturer's protocol for adapter ligation according to the desired sequencing platform. For example, TruSeq Small RNA kit should be used for Illumina sequencing system. As an alternative, custom protocols for adapter ligation have been previously described [11].
3. After 3′ and 5′ adapter ligation, repeat Subheadings 3.6 and 3.7.

3.10 Reverse Transcription (RT) and Cleaning-up with Magnetic Beads

1. Continue with the manufacturer's protocol for RT-PCR followed by Agencourt AMPure XP cleaning protocol. Use 1:2 (v/v) ratio for mixing sample and beads.
2. Elute in 25 μl for PCR Amplification.

3.11 PCR Amplification of the GRO-seq Library and Cleaning-up with Magnetic Beads

1. Continue with the manufacturer's protocol for PCR (max 12 cycles) followed by Agencourt AMPure XP cleaning protocol. Use 1:1 (v/v) ratio for mixing beads and sample.
2. Repeat magnetic beads cleaning step.
3. Elute in 15 μl.

3.12 Quantification of GRO-seq Library

1. Run 1 μl of the GRO-seq library on a Bioanalyzer using an Agilent DNA 1000 or Agilent High Sensitivity DNA chip. The fragments should be dispersed between 200 and 400 bp (Fig. 2).

3.13 Sequencing of GRO-seq Library and Data Analysis

1. Amplicons can be sequenced in a HiSeq 2500 (Illumina) platform using a standard 65 bp single-read reaction.
2. Sequenced reads can be aligned to the human genome (hg19) using bowtie2 [12] tool.
3. HOMER software can be applied to detect transcriptional units (TUs) [13]. The expression level of each TU can be calculated in each sample by using HTseq package [14].
4. The identification and classification of promoter-associated RNAs requires further bioinformatics analysis to associate their expression with known transcriptional start sites.

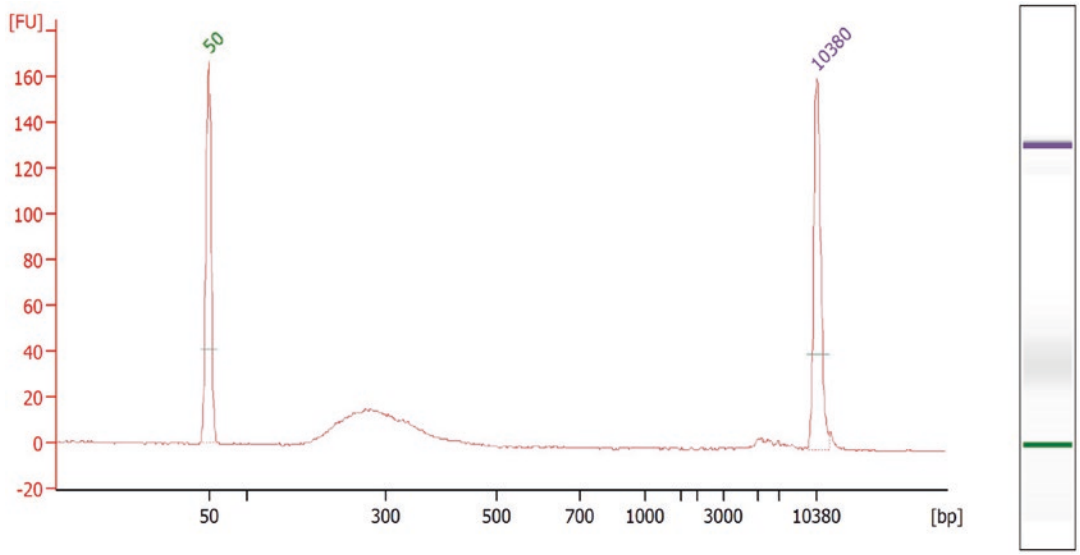


Fig. 2 Bioanalyzer result of a good-quality GRO-seq library is shown as an example. The fragments should be dispersed between 200 and 400 bp

4 Notes

1. It is important to equally expose the surface to swelling buffer. Rock the plates occasionally to prevent drying.
2. Resuspend nuclei by flicking the tube before adding the Lysis Buffer, and then invert tube several times.
3. In order to obtain better yields, it is advisable to use two separate reactions for NRO-rxn (two times of 5×10^6 nuclei).
4. Sarkosyl level and final concentration of CTP need to be adjusted according to the experimental system.
5. Having sarkosyl in NRO-mix causes the mixture to become very viscous. In order to obtain a homogenous mixture, cut the edge off a normal pipette tip and at least mix 15–20 times before placing the mixture at +30 °C.
6. If needed, combine two of the same reactions.
7. It is not recommended to use any salt containing buffers at this step. If you need to use salt containing buffer to obtain clear RNA pellet, you should wash the pellet at least two times with 70 % EtOH to remove the residual salt.
8. Upon fragmentation, it is expected to obtain a distribution of RNA fragments between 100 and 150 bp. The duration of the fragmentation reaction requires optimization and a bioanalyzer profile of different time points can provide the requisite information.

9. Handle the beads carefully because they do not form a rigid pellet. Allow 30–50 μl of SN to remain in the tube, in order to avoid losing your sample.
10. EDTA is needed to chelate divalent ions that catalyze cleavage of the RNA with heat.
11. You can check if the binding is at or near completion by spinning the beads down and removing the SN. Check the beads and supernatant with the Geiger counter. If the SN has 5–10× fewer counts than the beads, then the binding is likely complete.

Acknowledgment

We would like to specially thank Leighton Core for kindly sharing the GRO-seq protocol with us. RL is supported by the Fundação para a Ciência e Tecnologia de Portugal (SFRH/BD/74476/2010; POPH/FSE).

Reference

1. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. doi:10.1038/nrg2484
2. TR C, JA S (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157(1):77–94. doi:10.1016/j.cell.2014.03.008
3. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322(5909):1845–1848. doi:10.1126/science.1162228
4. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469(7330):368–373. doi:10.1038/nature09652
5. Kwak H, Fuda NJ, Core LJ, Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339(6122):950–953. doi:10.1126/science.1229386
6. Leveille N, Melo CA, Rooijers K et al (2015) Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nat Commun 6:6520. doi:10.1038/ncomms7520
7. Sun M, Gadad SS, Kim DS, Kraus WL (2015) Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. Mol Cell 59(4):698–711. doi:10.1016/j.molcel.2015.06.023
8. Korkmaz, G. et al. Functional genetic screens for enhancer elements in the human genome using CRISPR–Cas9. Nat. Biotechnol. 34, 192–198 (2016). doi:10.1038/nbt.3450
9. Hah N, Murakami S, Nagari A et al (2013) Enhancer transcripts mark active estrogen receptor binding sites. Genome Res 23(8):1210–1223. doi:10.1101/gr.152306.112
10. Li W, Notani D, Ma Q et al (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498(7455):516–520. doi:10.1038/nature12210
11. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324(5924):218–223. doi:10.1126/science.1168978
12. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. doi:10.1038/nmeth.1923
13. 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(4):576–589. doi:10.1016/j.molcel.2010.05.004
14. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166–169. doi:10.1093/bioinformatics/btu638

Chapter 4

NanoCAGE: A Method for the Analysis of Coding and Noncoding 5′-Capped Transcriptomes

Stéphane Poulain, Sachi Kato, Ophélie Arnaud, Jean-Étienne Morlighem, Makoto Suzuki, Charles Plessy, and Matthias Harbers

Abstract

Transcripts in all eukaryotes are characterized by the 5′-end specific cap structure in mRNAs. Cap Analysis Gene Expression or CAGE makes use of these caps to specifically obtain cDNA fragments from the 5′-end of RNA and sequences those at high throughput for transcript identification and genome-wide mapping of transcription start sites for coding and noncoding genes. Here, we provide an improved version of our nanoCAGE protocol that has been developed for preparing CAGE libraries from as little as 50 ng of total RNA within three standard working days. Key steps in library preparation have been improved over our previously published protocol to obtain libraries having a good 5′-end selection and a more equal size distribution for higher sequencing efficiency on Illumina MiSeq and HiSeq sequencers. We recommend nanoCAGE as the method of choice for transcriptome profiling projects even from limited amounts of RNA, and as the best approach for genome-wide mapping of transcription start sites within promoter regions.

Key words Cap analysis gene expression, CAGE, nanoCAGE, CAGEscan, RNA, Transcription Start Sites, TSS, Expression profiling, Template switching, Tagmentation, Multiplexing, Unique molecular identifiers, UMI

1 Introduction

Transcriptional regulation is an essential step for the utilization of genetic information, and stands at the basis of all biological processes. The underlying gene expression in any living cell or organism is tightly controlled, and thus expression profiles are commonly used to characterize biological samples on a molecular level. Accordingly, transcript profiling has become a standard approach in molecular biology to identify and quantify specific transcripts. Originally, mostly hybridization-based methods (microarrays) were used for studying genome-wide gene expression levels. However, researchers have increasingly shifted to sequencing-based methods like RNA-Seq (for "RNA Sequencing") or CAGE (for "Cap Analysis Gene Expression"), which provide

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_4,

comprehensive views on all transcripts within a sample without prior probe selection. By now these methods can easily be implemented in the laboratory using reagent kits on the market or following established protocols in the literature.

CAGE provides means for transcript identification and genome-wide identification of transcription start sites ("TSSs"), which makes it the method of choice for obtaining expression profiles and to relate them to regulatory regions in the genome [1, 2]. CAGE makes use of the so-called cap structure, where in eukaryotic cells mRNAs are modified at their 5′-ends by a 7-methylguanylate cap or related structures. The cap has important biological roles in stabilizing the transcript, its nuclear export, and its translation. However, the cap also offers a convenient handle to isolate the 5′-ends of mRNAs for cloning or direct sequencing. Most RNA transcripts contain a cap structure making this a universal approach to identifying coding and noncoding transcripts along with the regulatory regions in the genome controlling their expression. In CAGE, the cap is utilized to produce and sequence cDNAs that cover the full 5′-end of mRNAs. Since 2006 [3], most CAGE libraries are prepared using random primers to assure that cDNAs are extended to the 5′-end of RNAs and also to monitor non-polyadenylated transcripts in our CAGE libraries. To remove the secondary structures in the 5′-UTR that could be blocking the first-strand cDNA extension toward the cap, the reverse transcription reactions are carried out at higher temperature in the presence of reagents such as trehalose [4] and sorbitol [5], which thermostabilize the enzyme, and betain [6] that destabilizes secondary structures in the RNA. Even short sequencing reads of less than 50 bp can be reliably mapped back to the genome to identify the start sites where transcription had been initiated. Since in CAGE only one read per transcript is obtained, transcript levels can be obtained by direct counting of the reads ("digital expression profiling"). Overlapping reads having common sequences are grouped into the same cluster to define transcriptional activities at defined windows in the genome, representing promoters. CAGE provides the basic information needed for studying regulatory networks by combining information on available transcription factors found in the expression profile with information on binding sites in the promoter regions of those genes found up- or down-regulated in the biological sample [7, 8]. With these unique features, CAGE has made major contributions to our understanding of gene expression in human cells and the identification of enhancer elements during the FANTOM project [9], and the characterization of functional elements in the human genome during the ENCODE project [10]. However, with its focus on sequences derived from the 5′-end of mRNAs, CAGE falls short on a broad identification and characterization of different transcript structures, even though alternative use of different TSSs of the same gene can be monitored. Nevertheless, CAGE libraries can be prepared for being

sequenced from both ends, for the purpose of better linking a given TSS to a gene, and possibly to distinguish between certain splice variants having different TSSs (so-called CAGEscan method) [11, 12]. This is an important feature, where most genes have multiple TSSs that can be independently regulated. Hence, the correct assignment of a given TSS is mandatory for understanding tissue- or stage-specific features of promoter regions and their regulation by transcription factors and/or promoter-associated RNAs.

When full-length transcript characterization is needed, CAGE experiments may be combined with RNA-Seq data. During RNA-Seq experiments the starting RNA, or cDNAs derived thereof, are fragmented to obtain short DNA fragments that can be sequenced at a very high throughput (millions if not billions of reads) on a standard high-speed sequencer, e.g., available from Illumina or Life Technologies. In RNA-Seq experiments each transcript is ideally covered by many overlapping short reads that allow for the characterization of individual transcripts and estimation of expression levels [13, 14]. While RNA-Seq data can provide additional information on splice variants, allele-specific gene expression, and expressed genetic variations, a very high sequencing depth is required for minimal coverage of all transcripts. The interpretation of RNA-Seq data remains complicated and computer intensive when following de novo approaches to discover new genes and isoforms. It is difficult to assemble complete transcript sequences from multiple reads and to make accurate estimations on transcript levels, where the number of reads per transcript correlates with transcript length. Assigning reads to transcripts may even be more complicated where different splice variants derived from the same gene share common exons. In addition, it had been observed that for most RNA-Seq experiments there is an uneven distribution of the reads over the entire transcript length, where in particular the 5′-end sequences show lower coverage. As compared to other expression profiling methods, CAGE is the preferred method for transcript identification at lower cost than possible by the use of RNA-Seq methods while providing additional means for analysis of regulatory regions and genome annotation. Since in CAGE only one sequence read per transcript is needed as compared to multiple reads per transcript used in RNA-Seq, CAGE libraries can be completely sequenced even on an Illumina MiSeq sequencer. The nanoCAGE protocol described here provides libraries that can be sequenced from single or both ends depending on the users' needs. Even though sequencing the 5′-ends only is sufficient for transcript identification and TSS mapping, we advise to use CAGEscan for better linking TSS to alternative transcripts, to identify tissue- or stage-specific promoter regions, and to annotate complex regions in the genome. While paired-end sequencing is more expensive on the Illumina HiSeq platforms, it can be done at no extra cost with the standard 50-cycle v2 kit for the Illumina MiSeq, which contains reagents for 75 cycles, including index reads.

Over time various protocols for preparing CAGE libraries have been developed for using different sequencing platforms [15]. They are further distinct by the method used for the cap selection step. Unlike nanoCAGE, other CAGE methods use Cap-Trapping in the 5′-end selection step [16, 17]. While most efficient in 5′-end selection, the Cap Trapper method does not allow for working with very small amounts of RNA. In addition, Cap Trapping is a tedious method requiring many different incubation steps. In brief, after cDNA synthesis the cap structure in the cDNA-RNA hybrids is specifically biotinylated. In the following RNase digestion step, regions of single-stranded RNA are digested and only biotinylated cDNA-RNA hybrids remain, where the cDNA had been extended to the 5′-end. Those cDNA-RNA hybrids are enriched on streptavidin-coated beads and the cDNAs are then isolated for further library preparation. We shifted in nanoCAGE to template switching for 5′-end selection to develop a new protocol for rapid library preparation starting from very small amounts of total RNA [11, 18] that could overcome some of the limitations of the Cap-Trapper method. This protocol allows for the users to routinely prepare nanoCAGE libraries starting from some 50 to 500 ng of total RNA.

Template switching uses the template-free activity of the reverse transcriptase that can add few deoxycytosines (dCs) to the 3′-end of a newly synthesized cDNA strand when reaching the 5′-end of the RNA template. This activity is at least partially depending on the presence of the cap [19, 20, 21, 22], and therefore template switching can enrich the full-length rate of cDNA libraries. During the template switching reaction, a RNA or DNA/RNA hybrid oligonucleotide (the "template switching oligonucleotide") is hybridizing to poly-dC overhangs at the 3′-end of the cDNA, and after "switching" from the RNA template to the oligonucleotide the reverse transcriptase will continue extending the cDNA using the template switching oligonucleotide as its template (Fig. 1). Template switching oligonucleotides usually comprise three riboguanosine bases (rGs) at their 3′-end to efficiently hybridize to the added dC strand. The remaining sequence of the template switching oligonucleotide can be used to introduce a specific priming site for later PCR amplification. Since the 5′-end selection step happens already during the reverse transcription reaction, template switching allows to develop simple and rapid methods for mRNA characterization that require only small amounts of starting RNA.

In the nanoCAGE protocol published in 2011 [18], which is the latest version until the update presented in this chapter, the library preparation simply involved a reverse transcription reaction followed by PCR amplification steps. In libraries made with that protocol, we found a high rRNA content (up to 30 %). In addition, we noticed a wide size distribution of the DNA fragments

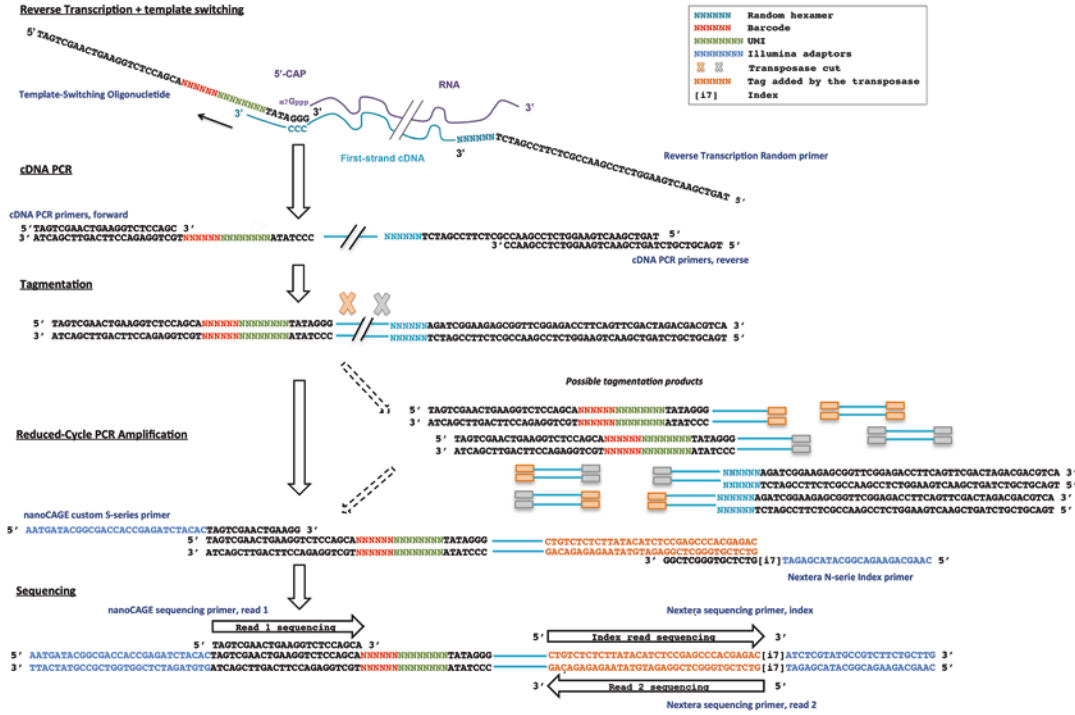


Fig. 1 Schematic representation of the different steps involved in the nanoCAGE protocol

within those CAGE libraries with DNA fragments longer than 1500 bp. Because those DNA fragments were too long to allow for cluster formation by bridge PCR, most likely a large number of fragments could not be sequenced on an Illumina platform. Those points have been addressed in the improved version of the nanoCAGE protocol described here (Fig. 1), where we added the following modifications:

1. Because of the 5′-end selection step, nanoCAGE libraries can be prepared directly from total RNA without need to remove rRNAs. However, we found that removal of rRNAs prior to library preparation can be helpful to remove the rRNA content within nanoCAGE libraries. We obtained satisfactory results by pretreating for example rat and honeybee total RNA samples with a 5′-phosphate-dependent exonuclease ("Terminator enzyme"). This pretreatment has worked very well when starting nanoCAGE library preparation from 500 ng of total RNA. In addition, the exonuclease treatment can improve the selection of 5′-ends, because truncated mRNAs having a 5′-end phosphate group are digested as well. We have also prepared a nanoCAGE library from an rRNA-depleted sample using a Ribominus kit, but we found this approach less suitable for working with very small amounts of RNA.

2. We added a tagmentation step (similarly as in C1 STRT) [23] using the Tn5 transposase for preparing DNA fragments of a well-defined length distribution of about 150–1000 bp that is most suitable for sequencing on Illumina platforms using an Illumina Nextera DNA Sample Prep Kit.
3. We also added unique molecular identifiers (UMIs)[24, 25]to the sequence of the template switching oligonucleotides, to count the number of transcript molecules detected for each gene in each sample. In nanoCAGE, UMIs consist of eight random nucleotides, where in theory only one random sequence should be selected per transcript. Therefore, distinct UMIs disambiguate CAGE tags derived from identical transcripts. Hence, UMIs can be useful to remove PCR duplicates, which have the same UMI.
4. Finally, we provide more information for multiplex sequencing. The protocol allows for analyzing up to 96 RNA samples in a single experiment working in a 96-well plate format. Accordingly, we are providing here for our protocol 96 different barcode sequences. These barcodes can be introduced during the reverse transcription step, where the cDNA molecules obtained from each sample are tagged by a single barcode sequence composed of six nucleotides to specifically mark each RNA sample. The barcoded cDNAs obtained for each sample are further amplified individually or as pools by cDNA PCR, and, equal quantities of the different libraries are then pooled for further tagmentation. During the tagmentation step, Illumina index sequences are added that can be used for additional or alternative multiplexing strategies.

With those improvements, we hope that the new nanoCAGE protocol will be a valuable tool for the research community to perform expression-profiling experiments from small amounts of RNA. Currently, efforts are being made by our group to further improve the protocol presented here for working with even smaller amounts of starting RNA to finally track even the transcriptomes of single cells on a high throughput.

The results from CAGE experiments can provide useful data sets to link regulatory regions to transcriptional activity, attesting for the availability of expressed transcription factors and certainly some groups of promoter-associated RNAs [26, 27]. CAGE is thus a powerful tool for analyzing gene networks and genome annotations. In addition, CAGE may be used to extend expression information from RNA-Seq data sets for better linking transcripts to TSSs. During the FANTOM project, we found that CAGE methods were indispensable for genome-wide studies on regulatory regions. Therefore, CAGE data represent an essential reference for studying promoter regions by other approaches targeting at the identification of promoter-associated proteins and RNAs, or chromatin structures.

2 Materials

2.1 Reagents

1. Agencourt AMPure XP (Beckman Coulter).
2. Agilent RNA 6000 Nano Kit (Agilent Technologies).
3. Agilent High Sensitivity DNA kit (Agilent Technologies).
4. Betaine, 5 M.
5. dNTP mixture, 2.5 mM each.
6. DTT, 0.1 M.
7. Ethanol, 70 %.
8. Illumina Nextera XT DNA Library Preparation Kit.
9. Illumina PhiX Control v3.
10. Illumina sequencing kit (e.g., MiSeq Reagent Kit v2).
11. Kapa HiFi HotSart Ready Mix.
12. Lambda DNA standard, 100 μg/mL in TE.
13. NaOH, 5 M, pH ≥ 13.
14. Qiagen EB buffer.
15. Fluorescence-based dsDNA quantitation Assay.
16. RNase decontamination solution.
17. SuperScript III Reverse Transcriptase (Life Technologies) (*see* **Note 1**).
18. SYBR qPCR master mix for real-time RT-PCR (qPCR), with broad range of amplification size and high sensitivity.
19. TE buffer 20× (200 mM Tris–HCl, 20 mM EDTA, pH 7.5).
20. Terminator 5′-Phosphate-Dependent Exonuclease.
21. Trehalose/sorbitol stock solution: saturate 8.02 g of D(+)-Trehalose dehydrate and 17.8 g of d-Sorbitol in 30 mL of water, and autoclave the mixture at 121 °C for 30 min (*see* **Notes 2** and **3**).
22. UltraPure DNase/RNase Free Distilled Water.

2.2 Equipment

1. Bioanalyzer (Agilent Technologies).
2. Centrifuges, benchtop, adapted for use with 96-well plates, PCR tubes, and microcentrifuge tubes.
3. Illumina sequencing system (e.g., MiSeq or HiSeq Sequencing System).
4. Magnetic stand for bead separation, adapted for PCR tubes and 96-well plates.
5. Micropipettors.
6. Microplate reader.
7. NanoDrop spectrophotometer (Thermo Scientific).

8. Qubit fluorometer (Life Technologies).
9. Real-Time PCR system.
10. Sealant applicator.
11. Thermal cycler.
12. Water purification system.

2.3 Consumables

1. Gloves.
2. Low binding filter tips.
3. Low binding PCR plates, 96-well.
4. Low binding PCR tube strips with caps, 0.2 mL.
5. Low binding microcentrifuge tubes, 1.5 mL.
6. Microplates for fluorescence assays, 384-well.
7. PCR seals.
8. Real-Time PCR plates, 96-well.

2.4 Oligonucleotides

The different oligonucleotides used in nanoCAGE are listed in Table 1 and are further detailed in the following paragraphs.

2.4.1 Reverse Transcription Random Primers

These primers bind RNA molecules at random sites and are used as template by the reverse transcriptase for the synthesis of first-strand cDNAs. The use of oligo-dT reverse transcription primers is not recommended, because cDNA transcripts often do not reach the 5′-ends. In addition, oligo-dT priming does not allow for the detection of non-polyadenylated RNAs.

5′-TAGTCGAACTGAAGGTCTCCGAACCGCTCT TCCGATCTNNNNNN-3′

Legend:

- T A G T C G A A C T G A A G G T C T C C G A A C C G C T CTTCCGATCT: Target sequence for second-strand cDNA synthesis by cDNA PCR
- NNNNNN: random sequence of six nucleotides used as template for first-strand cDNA synthesis by the reverse transcriptase

2.4.2 Template Switching Oligonucleotides

These oligonucleotides contain at their 3′-end a sequence of three guanosine ribonucleotides (rG) that allows for template switching of capped RNA molecules during the reverse transcription reaction. To reduce the cost of synthesis, all of the bases contained in the sequence of the primer are deoxynucleotides except for the guanosine ribonucleotides (rG). These ribonucleotides are essential for preventing strand-invasion artifacts [28] and cannot be replaced by deoxyriboguanosines. A sequence identifier ("barcode," *see* Table 2) is inserted 5′ of the riboguanosines to multiplex DNA libraries produced from different RNA specimens in a single

Table 1
Oligonucleotides used in nanoCAGE

Step	Primer name	Primer sequence (5′→3′)	Stock concentration	Purification grade
First-strand cDNA synthesis	Reverse Transcription Random primer	TAGTCGAACTGAAGGTCTCCGAACCGCTCTTCCGATCTNNNNNN	100 μM	Standard desalting
First-strand cDNA synthesis	Template Switching Oligonucleotide	TAGTCGAACTGAAGGTCTCCAGCA[barcode_sequence]NNNNNNNNTATA(rG)(rG)(rG)	1 mM, each	Standard desalting
Real-Time qPCR and Second-strand cDNA synthesis	cDNA PCR primer, Forward	TAGTCGAACTGAAGGTCTCCAGC	100 μM	Standard desalting
Real-Time qPCR and Second-strand cDNA synthesis	cDNA PCR primer, Reverse	TGACGTCGTCTAGTCGAACTGAAGGTCTCCGAACC	100 μM	Standard desalting
Tagmentation—Reduced-Cycle PCR Amplification	nanoCAGE custom S-series primer	AATGATACGGCGACCACCGAGATCTACACTAGTCGAACTGAAGG	100 μM	Standard desalting
Tagmentation—Reduced-Cycle PCR Amplification	Nextera XT N-series Index primers[a]	CAAGCAGAAGACGGCATACGAGAT[index]GTCTCGTGGGCTCGG		(*see* **Note 4**)
Sequencing	nanoCAGE Sequencing primer, Read 1 (*see* **Note 5**)	TAGTCGAACTGAAGGTCTCCAGCA	100 μM	Standard desalting
Sequencing	Nextera Sequencing primer, Index (*see* **Note 6**)			

[a]Oligonucleotide sequences© 2007–2013 Illumina, Inc. All rights reserved

Table 2
List of the 96 nanoCAGE barcodes that can be inserted in the sequence of the Template Switching Oligonucleotide

Barcode ID	Barcode sequence[a] (5′→3′)	Template switching oligonucleotide sequence (5′→3′)
01	ACACAG	TAGTCGAACTGAAGGTCTCCAGCAACACAGNNNNNNNNTATA(rG)(rG)(rG)
02	ACACGT	TAGTCGAACTGAAGGTCTCCAGCAACACGTNNNNNNNNTATA(rG)(rG)(rG)
03	ACACTC	TAGTCGAACTGAAGGTCTCCAGCAACACTCNNNNNNNNTATA(rG)(rG)(rG)
04	ACAGAT	TAGTCGAACTGAAGGTCTCCAGCAACAGATNNNNNNNNTATA(rG)(rG)(rG)
05	ACAGCA	TAGTCGAACTGAAGGTCTCCAGCAACAGCANNNNNNNNTATA(rG)(rG)(rG)
06	ACAGTG	TAGTCGAACTGAAGGTCTCCAGCAACAGTGNNNNNNNNTATA(rG)(rG)(rG)
07	ACATAC	TAGTCGAACTGAAGGTCTCCAGCAACATACNNNNNNNNTATA(rG)(rG)(rG)
08	ACATCT	TAGTCGAACTGAAGGTCTCCAGCAACATCTNNNNNNNNTATA(rG)(rG)(rG)
09	ACATGA	TAGTCGAACTGAAGGTCTCCAGCAACATGANNNNNNNNTATA(rG)(rG)(rG)
10	AGTACG	TAGTCGAACTGAAGGTCTCCAGCAAGTACGNNNNNNNNTATA(rG)(rG)(rG)
11	AGTAGC	TAGTCGAACTGAAGGTCTCCAGCAAGTAGCNNNNNNNNTATA(rG)(rG)(rG)
12	AGTATA	TAGTCGAACTGAAGGTCTCCAGCAAGTATANNNNNNNNTATA(rG)(rG)(rG)
13	AGTCAG	TAGTCGAACTGAAGGTCTCCAGCAAGTCAGNNNNNNNNTATA(rG)(rG)(rG)
14	AGTCGT	TAGTCGAACTGAAGGTCTCCAGCAAGTCGTNNNNNNNNTATA(rG)(rG)(rG)
15	AGTCTC	TAGTCGAACTGAAGGTCTCCAGCAAGTCTCNNNNNNNNTATA(rG)(rG)(rG)
16	AGTGAT	TAGTCGAACTGAAGGTCTCCAGCAAGTGATNNNNNNNNTATA(rG)(rG)(rG)
17	AGTGCA	TAGTCGAACTGAAGGTCTCCAGCAAGTGCANNNNNNNNTATA(rG)(rG)(rG)
18	AGTGTG	TAGTCGAACTGAAGGTCTCCAGCAAGTGTGNNNNNNNNTATA(rG)(rG)(rG)
19	ATCACG	TAGTCGAACTGAAGGTCTCCAGCAATCACGNNNNNNNNTATA(rG)(rG)(rG)

20	ATCAGC	TAGTCGAACTGAAGGTCTCCAGCAATCAGCNNNNNNNNTATA(rG)(rG)(rG)
21	ATCATA	TAGTCGAACTGAAGGTCTCCAGCAATCATANNNNNNNNTATA(rG)(rG)(rG)
22	ATCGAT	TAGTCGAACTGAAGGTCTCCAGCAATCGATNNNNNNNNTATA(rG)(rG)(rG)
23	ATCGCA	TAGTCGAACTGAAGGTCTCCAGCAATCGCANNNNNNNNTATA(rG)(rG)(rG)
24	ATCGTG	TAGTCGAACTGAAGGTCTCCAGCAATCGTGNNNNNNNNTATA(rG)(rG)(rG)
25	ATCTAC	TAGTCGAACTGAAGGTCTCCAGCAATCTACNNNNNNNNTATA(rG)(rG)(rG)
26	ATCTCT	TAGTCGAACTGAAGGTCTCCAGCAATCTCTNNNNNNNNTATA(rG)(rG)(rG)
27	ATCTGA	TAGTCGAACTGAAGGTCTCCAGCAATCTGANNNNNNNNTATA(rG)(rG)(rG)
28	CACACG	TAGTCGAACTGAAGGTCTCCAGCACACACGNNNNNNNNTATA(rG)(rG)(rG)
29	CACAGC	TAGTCGAACTGAAGGTCTCCAGCACACAGCNNNNNNNNTATA(rG)(rG)(rG)
30	CACATA	TAGTCGAACTGAAGGTCTCCAGCACACATANNNNNNNNTATA(rG)(rG)(rG)
31	CACGAT	TAGTCGAACTGAAGGTCTCCAGCACACGATNNNNNNNNTATA(rG)(rG)(rG)
32	CACGCA	TAGTCGAACTGAAGGTCTCCAGCACACGCANNNNNNNNTATA(rG)(rG)(rG)
33	CACGTG	TAGTCGAACTGAAGGTCTCCAGCACACGTGNNNNNNNNTATA(rG)(rG)(rG)
34	CACTAC	TAGTCGAACTGAAGGTCTCCAGCACACTACNNNNNNNNTATA(rG)(rG)(rG)
35	CACTCT	TAGTCGAACTGAAGGTCTCCAGCACACTCTNNNNNNNNTATA(rG)(rG)(rG)
36	CACTGA	TAGTCGAACTGAAGGTCTCCAGCACACTGANNNNNNNNTATA(rG)(rG)(rG)
37	CGACAG	TAGTCGAACTGAAGGTCTCCAGCACGACAGNNNNNNNNTATA(rG)(rG)(rG)
38	CGACGT	TAGTCGAACTGAAGGTCTCCAGCACGACGTNNNNNNNNTATA(rG)(rG)(rG)
39	CGACTC	TAGTCGAACTGAAGGTCTCCAGCACGACTCNNNNNNNNTATA(rG)(rG)(rG)
40	CGAGAT	TAGTCGAACTGAAGGTCTCCAGCACGAGATNNNNNNNNTATA(rG)(rG)(rG)
41	CGAGCA	TAGTCGAACTGAAGGTCTCCAGCACGAGCANNNNNNNNTATA(rG)(rG)(rG)
42	CGAGTG	TAGTCGAACTGAAGGTCTCCAGCACGAGTGNNNNNNNNTATA(rG)(rG)(rG)

(continued)

Table 2
(continued)

Barcode ID	Barcode sequence[a] (5′→3′)	Template switching oligonucleotide sequence (5′→3′)
43	CGATAC	TAGTCGAACTGAAGGTCTCCAGCACGATACNNNNNNNNTATA(rG)(rG)(rG)
44	CGATCT	TAGTCGAACTGAAGGTCTCCAGCACGATCTNNNNNNNNTATA(rG)(rG)(rG)
45	CGATGA	TAGTCGAACTGAAGGTCTCCAGCACGATGANNNNNNNNTATA(rG)(rG)(rG)
46	CTGACG	TAGTCGAACTGAAGGTCTCCAGCACTGACGNNNNNNNNTATA(rG)(rG)(rG)
47	CTGAGC	TAGTCGAACTGAAGGTCTCCAGCACTGAGCNNNNNNNNTATA(rG)(rG)(rG)
48	CTGATA	TAGTCGAACTGAAGGTCTCCAGCACTGATANNNNNNNNTATA(rG)(rG)(rG)
49	CTGCAG	TAGTCGAACTGAAGGTCTCCAGCACTGCAGNNNNNNNNTATA(rG)(rG)(rG)
50	CTGCGT	TAGTCGAACTGAAGGTCTCCAGCACTGCGTNNNNNNNNTATA(rG)(rG)(rG)
51	CTGCTC	TAGTCGAACTGAAGGTCTCCAGCACTGCTCNNNNNNNNTATA(rG)(rG)(rG)
52	CTGTAC	TAGTCGAACTGAAGGTCTCCAGCACTGTACNNNNNNNNTATA(rG)(rG)(rG)
53	CTGTCT	TAGTCGAACTGAAGGTCTCCAGCACTGTCTNNNNNNNNTATA(rG)(rG)(rG)
54	CTGTGA	TAGTCGAACTGAAGGTCTCCAGCACTGTGANNNNNNNNTATA(rG)(rG)(rG)
55	GAGACG	TAGTCGAACTGAAGGTCTCCAGCAGAGACGNNNNNNNNTATA(rG)(rG)(rG)
56	GAGAGC	TAGTCGAACTGAAGGTCTCCAGCAGAGAGCNNNNNNNNTATA(rG)(rG)(rG)
57	GAGATA	TAGTCGAACTGAAGGTCTCCAGCAGAGATANNNNNNNNTATA(rG)(rG)(rG)
58	GAGCAG	TAGTCGAACTGAAGGTCTCCAGCAGAGCAGNNNNNNNNTATA(rG)(rG)(rG)
59	GAGCGT	TAGTCGAACTGAAGGTCTCCAGCAGAGCGTNNNNNNNNTATA(rG)(rG)(rG)
60	GAGCTC	TAGTCGAACTGAAGGTCTCCAGCAGAGCTCNNNNNNNNTATA(rG)(rG)(rG)
61	GAGTAC	TAGTCGAACTGAAGGTCTCCAGCAGAGTACNNNNNNNNTATA(rG)(rG)(rG)
62	GAGTCT	TAGTCGAACTGAAGGTCTCCAGCAGAGTCTNNNNNNNNTATA(rG)(rG)(rG)

63	GAGTGA	TAGTCGAACTGAAGGTCTCCAGCAGAGTGANNNNNNNNNTATA(rG)(rG)(rG)
64	GCTACG	TAGTCGAACTGAAGGTCTCCAGCAGCTACGNNNNNNNNNTATA(rG)(rG)(rG)
65	GCTAGC	TAGTCGAACTGAAGGTCTCCAGCAGCTAGCNNNNNNNNNTATA(rG)(rG)(rG)
66	GCTATA	TAGTCGAACTGAAGGTCTCCAGCAGCTATANNNNNNNNNTATA(rG)(rG)(rG)
67	GCTCAG	TAGTCGAACTGAAGGTCTCCAGCAGCTCAGNNNNNNNNNTATA(rG)(rG)(rG)
68	GCTCGT	TAGTCGAACTGAAGGTCTCCAGCAGCTCGTNNNNNNNNNTATA(rG)(rG)(rG)
69	GCTCTC	TAGTCGAACTGAAGGTCTCCAGCAGCTCTCNNNNNNNNNTATA(rG)(rG)(rG)
70	GCTGAT	TAGTCGAACTGAAGGTCTCCAGCAGCTGATNNNNNNNNNTATA(rG)(rG)(rG)
71	GCTGCA	TAGTCGAACTGAAGGTCTCCAGCAGCTGCANNNNNNNNNTATA(rG)(rG)(rG)
72	GCTGTG	TAGTCGAACTGAAGGTCTCCAGCAGCTGTGNNNNNNNNNTATA(rG)(rG)(rG)
73	GTACAG	TAGTCGAACTGAAGGTCTCCAGCAGTACAGNNNNNNNNNTATA(rG)(rG)(rG)
74	GTACGT	TAGTCGAACTGAAGGTCTCCAGCAGTACGTNNNNNNNNNTATA(rG)(rG)(rG)
75	GTACTC	TAGTCGAACTGAAGGTCTCCAGCAGTACTCNNNNNNNNNTATA(rG)(rG)(rG)
76	GTAGAT	TAGTCGAACTGAAGGTCTCCAGCAGTAGATNNNNNNNNNTATA(rG)(rG)(rG)
77	GTAGCA	TAGTCGAACTGAAGGTCTCCAGCAGTAGCANNNNNNNNNTATA(rG)(rG)(rG)
78	GTAGTG	TAGTCGAACTGAAGGTCTCCAGCAGTAGTGNNNNNNNNNTATA(rG)(rG)(rG)
79	GTATAC	TAGTCGAACTGAAGGTCTCCAGCAGTATACNNNNNNNNNTATA(rG)(rG)(rG)
80	GTATCT	TAGTCGAACTGAAGGTCTCCAGCAGTATCTNNNNNNNNNTATA(rG)(rG)(rG)
81	GTATGA	TAGTCGAACTGAAGGTCTCCAGCAGTATGANNNNNNNNNTATA(rG)(rG)(rG)
82	TATACG	TAGTCGAACTGAAGGTCTCCAGCATATACGNNNNNNNNNTATA(rG)(rG)(rG)
83	TATAGC	TAGTCGAACTGAAGGTCTCCAGCATATAGCNNNNNNNNNTATA(rG)(rG)(rG)
84	TATATA	TAGTCGAACTGAAGGTCTCCAGCATATATANNNNNNNNNTATA(rG)(rG)(rG)

(continued)

Table 2
(continued)

Barcode ID	Barcode sequence[a] (5′→3′)	Template switching oligonucleotide sequence (5′→3′)
85	TATCAG	TAGTCGAACTGAAGGTCTCCAGCATATCAGNNNNNNNNTATA(rG)(rG)(rG)
86	TATCGT	TAGTCGAACTGAAGGTCTCCAGCATATCGTNNNNNNNNTATA(rG)(rG)(rG)
87	TATCTC	TAGTCGAACTGAAGGTCTCCAGCATATCTCNNNNNNNNTATA(rG)(rG)(rG)
88	TATGAT	TAGTCGAACTGAAGGTCTCCAGCATATGATNNNNNNNNTATA(rG)(rG)(rG)
89	TATGCA	TAGTCGAACTGAAGGTCTCCAGCATATGCANNNNNNNNTATA(rG)(rG)(rG)
90	TATCTC	TAGTCGAACTGAAGGTCTCCAGCATATCTCNNNNNNNNTATA(rG)(rG)(rG)
91	TCGACG	TAGTCGAACTGAAGGTCTCCAGCATCGACGNNNNNNNNTATA(rG)(rG)(rG)
92	TCGAGC	TAGTCGAACTGAAGGTCTCCAGCATCGAGCNNNNNNNNTATA(rG)(rG)(rG)
93	TCGATA	TAGTCGAACTGAAGGTCTCCAGCATCGATANNNNNNNNTATA(rG)(rG)(rG)
94	TCGCAG	TAGTCGAACTGAAGGTCTCCAGCATCGCAGNNNNNNNNTATA(rG)(rG)(rG)
95	TCGCGT	TAGTCGAACTGAAGGTCTCCAGCATCGCGTNNNNNNNNTATA(rG)(rG)(rG)
96	TCGCTC	TAGTCGAACTGAAGGTCTCCAGCATCGCTCNNNNNNNNTATA(rG)(rG)(rG)

[a]The barcode sequences have been designed to avoid palindromes and low-complexity stretches

sequencing sample. The barcode is followed in 3′ by a sequence of eight random bases (Unique Molecular Identifier) that is further used to count the number of transcript molecules detected for a gene in each RNA sample analyzed [25]. The UMI is followed in 3′ by a 4-bases spacer (TATA) to minimize template switching artifacts induced by the presence of the barcode through the strand invasion process [28].

5′-TAGTCGAACTGAAGGTCTCCAGCA[barcode_sequence]NNNNNNNNTATA(rG)(rG)(rG)-3’

Legend:

- TAGTCGAACTGAAGGTCTCCAGCA: Target sequence for second-strand cDNA synthesis during the cDNA PCR.
- [barcode_sequence]: defined sequence of six bases, specific of each Template Switching Oligonucleotide (Table 2). This sequence allows identifying the sequencing tags corresponding to each RNA sample contained in the sequenced nanoCAGE library.
- NNNNNNNN: Unique Molecular Identifier (UMI), random sequence of eight nucleotides designed to count the number of transcripts detected per gene expressed in each RNA sample.
- TATA: spacer, sequence inserted to reduce the template switching artifacts generated by the presence of the barcode sequence, introduced in Tang et al., 2013 [28].
- (rG)(rG)(rG): guanosine ribonucleotides used for the template switching.

2.4.3 cDNA PCR Primers

These primers are used for second-strand synthesis and PCR amplification of the first-stand cDNA molecules produced in the reverse transcription reaction.

Forward primer: 5′-TAGTCGAACTGAAGGTCTCCAGC-3′

Reverse primer: 5′-TGACGTCGTCTAGTCGAACTGAAGGTCTCCGAACC-3′

2.4.4 nanoCAGE Custom S-series Primer

This primer adds an Illumina sequencing adapter to the 5′-end of nanoCAGE fragments during the Reduced-Cycle PCR Amplification step of the tagmentation reaction.

5′-AATGATACGGCGACCACCGAGATCTACACTAGTCGAACTGAAGG-3′

Legend:

- AATGATACGGCGACCACCGAGATCTACAC: Illumina sequencing adapter sequence
- TAGTCGAACTGAAGG: Target sequence for Reduced-Cycle PCR Amplification

2.4.5 Nextera XT N-Series Index Primers (Nextera XT DNA Sample Prep Kit)

These primers add Nextera N-Series Index (*see* Table 3) and an Illumina sequencing adapter sequences to the 3′-ends of nanoCAGE fragments during the Reduced-Cycle PCR Amplification step of the tagmentation reaction.

5 ′ - C A A G C A G A A G A C G G C A T A C G A G A T [index]GTCTCGTGGGCTCGG-3’ *****

***** Oligonucleotide sequences© 2007–2013 Illumina, Inc. All rights reserved.

Legend:

- CAAGCAGAAGACGGCATACGAGAT: Illumina sequencing adapter sequence.
- [index]: sequence of eight bases corresponding to the Nextera N-Series Index (Table 3).
- GTCTCGTGGGCTCGG: Target sequence for Reduced-Cycle PCR Amplification.

2.4.6 nanoCAGE Sequencing Primer, Read 1

This forward primer allows for sequencing the 5′-end of nanoCAGE fragments from 5′ to 3′. The Read 1 contains the barcode, the UMI, the TATAGGG linker, and the sequence of the 5′-end of the cDNA, allowing to map the sequencing read to a reference genome assembly, and thus identify transcripts and transcription start sites.

5′-TAGTCGAACTGAAGGTCTCCAGCA-3′

Legend:

- TAGTCGAACTGAAGGTCTCCAGCA: Target sequence used for sequencing the 5′-end of nanoCAGE fragments.

2.4.7 Nextera Sequencing Primer, Read 2

This reverse primer allows for sequencing the 3′-end of nanoCAGE fragments from 3′ to 5′. The Read 2 contains further information about the sequence of the cDNA identified by the Read 1. As nanoCAGE libraries are produced by tagmentation with the Nextera XT DNA Library Preparation kit, the molecules contained in the pool of cDNA PCR products used as input material are fragmented at random sites by the Tn5 transposase. Therefore, Reads 2 from pairs having an identical Read 1 can be combined to provide a better coverage of the sequence of the transcripts.

2.4.8 Nextera Sequencing Primer, Index

This forward primer allows for sequencing the index of nanoCAGE fragments from 5′ to 3′. The sequence of the Index Read contains information about the Nextera XT N-series Index (sequence of eight bases) to identify an RNA sample or a group of RNA samples.

Table 3
Detailed sequences of the 12 Nextera XT N-Series Index Primers

Index ID	Index bases in adapter	Nextera XT N-series index primer sequence (5′→3′)	Index bases for entry on Illumina sample sheet
N701[a]	TCGCCTTA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG	TAAGGCGA
N702[a]	CTAGTACG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG	CGTACTAG
N703[a]	TTCTGCCT	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGG	AGGCAGAA
N704	GCTCAGGA	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG	TCCTGAGC
N705[a]	AGGAGTCC	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGG	GGACTCCT
N706[a]	CATGCCTA	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGG	TAGGCATG
N707[a]	GTAGAGAG	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGG	CTCTCTAC
N708	CCTCTCTG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGG	CAGAGAGG
N709[a]	AGCGTAGC	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGG	GCTACGCT
N710[a]	CAGCCTCG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGG	CGAGGCTG
N711[a]	TGCCTCTT	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGG	AAGAGGCA
N712[a]	TCCTCTAC	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGG	GTAGAGGA

[a]Oligonucleotide sequences

3 Methods

Work always under RNase-free conditions. Wear gloves and lab coat throughout the procedure. Keep samples and reagents under RNase-free conditions, as RNA degradation may interfere with nanoCAGE library preparation. Use low binding plastic consumables to improve the recovery of nucleic acids at each step of the protocol.

Before using valuable RNA samples, we recommend investigators to practice making nanoCAGE libraries with a control total RNA to develop their abilities to check cDNA synthesis by Real-Time quantitative PCR, to quantify cDNA PCR products by Quant-iT PicoGreendsDNA Assay, and to interpret Bioanalyzer profiles of the prepared libraries.

For the convenience of the users, we provide a nanoCAGE bench protocol (Fig. 2) (*see* **Note 7**) that can be downloaded at the following URL: https://github.com/Population-Transcriptomics/nanoCAGE-2016/raw/master/nanoCAGE_bench_protocol.pdf

3.1 Method Summary and Time Requirements

The different steps included in the nanoCAGE protocol and their time requirements are presented in Table 4.

3.2 General Guidelines for the Planning of nanoCAGE Experiments

3.2.1 RNA Extraction and Quality- Check

1. Appropriate methods should be used to extract small amounts of RNA (e.g., Invitrogen, TRIzol Reagent). Total RNA can be further purified (e.g., Invitrogen, PureLink RNA Micro Scale Kit) for nanoCAGE library preparation. Nuclease-free water (or a suitable elution buffer provided with the RNA extraction kit) should be used to elute the RNA.
2. Store total RNA samples at −80 °C until used. Try to reduce the number of freeze/thaw cycles as much as possible to prevent the degradation of RNA samples.
3. Preferably test integrity of purified total RNA (if >50 ng are available) using an Agilent Bioanalyzer (Agilent RNA 6000 Nano Kit). We recommend total RNA samples with an RNA integrity number (RIN) value ≥7 [29, 30] (*see* **Note 8**).
4. In addition to the Bioanalyzer analysis, we recommend quantifying the RNA using a Qubit Fluorometer (Invitrogen Qubit RNA BR Assay Kit) or a NanoDrop spectrophotometer. Alternatively, use agarose gel electrophoresis in the presence of formaldehyde [31, 32].
5. When enough RNA material is available as a starting material, we recommend analyzing technical duplicates or triplicates of each sample.

nanoCAGE – DAY 1

RNA samples:

Before start checklist:

- □ Pipettes 1000 µL, 200 µL, 10 µL, 2 µL
- □ Filtered Tips Low-bind
- □ 0.2 mL PCR tubes low-bind + caps
- □ 1.5 mL microcentrifuge tubes low-bind
- □ 96-well PCR plate low-bind
- □ 96-well Real-Time PCR plate
- □ Ultra-Pure Water
- □ Magnet for beads separation
- □ AMPure XP beads, resuspended and at room temperature
- □ Freshly prepared 70% EtOH, at room temperature
- □ Terminator Exonuclease (1 U/µL) / SuperScript III Reverse Transcriptase (200 U/µL) / Takara SYBR Premix Ex Taq (2x), in freezer until needed
- □ SuperScript III First-Strand Buffer (5x) / DTT (0.1 M) / dNTPs (2.5 mM) / Betaine (5 M), on ice
- □ Trehalose/Sorbitol stock solution (0.66 M/3.3 M) / Reverse Transcription Random primers (100 µM) / Template Switching Oligonucleotides (1 mM), on ice
- □ cDNA PCR primers Forward and Reverse (10 µM) / Rox Reference Dye II (50x), on ice
- □ RNA samples, in freezer until needed

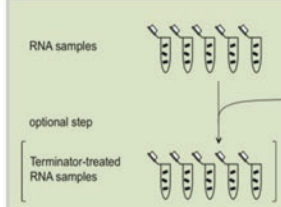
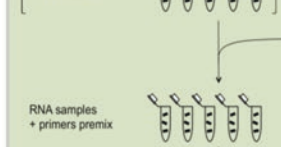
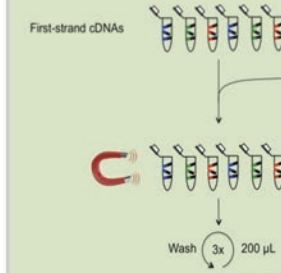
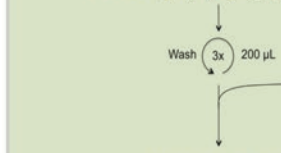
WORKFLOW	INSTRUCTIONS	NOTES/OBSERVATIONS	TIME/DATE
	□ ≥ 50 ng/µL of total RNA per sample or □ ≥ 500 ng/µL of total RNA per sample, if performing the Terminator treatment	- Use in preference high quality total RNA to optimize nanoCAGE library preparation for best sequencing results	
	(optional step) Add per sample: □ 1.3 µL of SuperScript III First-Strand Buffer (5x) □ 0.2 µL of Water □ 0.5 µL of Terminator Exonuclease (1 U/µL) Mix by pipetting + spin down Incubate at 30°C for 1 h		
	For each total RNA sample, add: □ 1 µL of a primers premix stock solution □ 1 µL of total RNA Mix by pipetting 10 times carefully + spin down Incubate at 65°C for 10 min Transfer on ice for at least 2 min Add per sample: □ 2 µL of SuperScript III First-Strand Buffer (5x) □ 1 µL of DTT (0.1 M) □ 2.5 µL of dNTPs (2.5 mM) □ 1.5 µL of Betaine (5 M) □ 1 µL of SuperScript III (200 U/µL) Incubate at 22°C for 10 min, 50°C for 30 min, 75°C for 15 min, and hold at 4°C	- Use a different barcode (primers premix solution) for each RNA sample and for each sample replicate - Prepare the primers premix solutions in advance by mixing 1 µL of a Template Switching Oligonucleotide (1 mM) + 1 µL of Reverse Transcription Random primers (100 µM) + 8 µL of Trehalose/Sorbitol stock solution (0.66 M/3.3 M) – [store at -20°C] - Trehalose/Sorbitol solutions are viscous; mix very carefully by pipetting up and down slowly for at least 10 times - Prepare a master mix for the RT reaction - Prepare a positive control of the RT reaction by using 1 µL of a control RNA sample - Prepare a negative control of the RT reaction by using 1 µL of water	
	□ Add 18 µL of AMPure XP beads at RT □ Mix by pipetting 10x slowly and incubate for 5 min □ Pellet on magnet and pipette off the supernatant □ Keep on magnet, wash 3x with 200 µL fresh 70% EtOH, do not disturb the pellet □ Resuspend pellet in 30 µL water, incubate at RT for 5 min □ Pellet beads on magnet, elute purified first-strand cDNAs and transfer to new PCR tubes low-bind	- Aspirate supernatant carefully - In order to avoid losing the sample, be sure not to aspirate beads in the pipette tip together with the solution	
			
	Mix per qPCR replicate: □ 5 µL of Takara SYBR Premix Ex Taq (2x) □ 0.1 µL of cDNA PCR primer, Forward (10 µM) □ 0.1 µL of cDNA PCR primer, Reverse (10 µM) □ 0.2 µL of Rox Reference Dye II (50x) □ 3.1 µL of Water □ 1.5 µL of purified first-strand cDNA □ Analyze qPCR results □ Calculate the average Ct value for each cDNA sample □ Determine the optimum number (N) of cDNA PCR cycles to be performed for each sample	- Analyze each first-strand cDNA and control samples in triplicates - Prepare a master mix for the real-time qPCR reaction - Prepare a negative control of the real-time qPCR reaction by adding 1.5 µL of water	
			

Fig. 2 nanoCAGE bench protocol

nanoCAGE – DAY 2

Purified First-Strand cDNA samples:

Before start checklist:

- □ Pipettes 1000 µL, 200 µL, 10 µL, 2 µL
- □ Filtered Tips Low-bind
- □ 0.2 mL PCR tubes low-bind + caps
- □ 1.5 mL microcentrifuge tubes low-bind
- □ 96-well PCR plate low-bind
- □ 384-well microplate for fluorescence assays
- □ Ultra-Pure Water
- □ Kapa HiFi HotStart Ready Mix (2x), on ice
- □ cDNA PCR primers, Forward and Reverse (10 µM), on ice
- □ Magnet for beads separation
- □ AMPure XP beads, resuspended and at room temperature
- □ Freshly prepared 70% EtOH, at room temperature
- □ Lambda DNA standard stock solution (1 ng/µL), on ice
- □ Quant-iT Picrogreen Reagent (200x), in fridge until use
- □ TE buffer (1x), at room temperature
- □ Bioanalyzer DNA High Sensitivity DNA kit, at room temperature 30 min before use
- □ Purified first-strand cDNA samples, on ice

WORKFLOW	INSTRUCTIONS	NOTES/OBSERVATIONS	TIME/DATE
First-strand cDNA samples	□ Purified first-strand cDNA samples prepared on Day 1		
cDNA PCR products	Add for each sample: □ 25 µL of Kapa HiFi HotStart Ready Mix (2x) □ 0.5 µL of cDNA PCR primer, Forward (10 µM) □ 0.5 µL of cDNA PCR primer, Reverse (10 µM) □ 4 µL of Water □ 20 µL of first-strand cDNA sample Mix by vortexing + spin down Incubate at 95°C for 3 min, then perform [N] PCR cycles of 98°C for 15 s, 65°C for 10 s, 72°C for 2 min, and incubate at 72°C for 2 min, hold at 4°C	- Prepare a master mix of the cDNA PCR reaction, including extra volumes for the positive and negative controls of the RT reaction - Perform [N] cDNA PCR cycles for each first-strand cDNA sample based on the Ct values obtained from the real-time qPCR experiment performed on Day 1	
Wash 3x 200 µL	□ Add 90 µL of AMPure XP beads at RT □ Mix by pipetting 10x slowly and incubate for 5 min □ Pellet on magnet and pipette off the supernatant □ Keep on magnet, wash 3x with 200 µL fresh 70% EtOH, do not disturb the pellet □ Resuspend pellet in 25 µL water, incubate at RT for 5 min □ Pellet beads on magnet, elute purified cDNA PCR products and transfer to new PCR tubes low-bind	- Aspirate supernatant carefully - In order to avoid losing the sample, be sure not to aspirate beads in the pipette tip together with the solution	
Purified cDNA PCR products			
	□ Use 3 µL of each purified cDNA PCR products samples for quantification by Quant-iT PicoGreen dsDNA assay		
Pool of purified cDNA PCR products	□ Pool purified cDNA PCR products tagged with different barcodes together [Mix at least 2 ng of each purified cDNA PCR products sample at this step]	- It is also possible to mix purified cDNA PCR products tagged with different barcodes by group/experimental condition and further perform one tagmentation reaction per group using one Nextera N-series Index primer per reaction, and then pool the different nanoCAGE libraries obtained together, prior to library quantification and sequencing	
	□ Use 3 µL of the pool of cDNA PCR products for quantification of the mix by Quant-iT PicoGreen dsDNA assay	- The concentration of the mix should be at least 0.1 ng/µL and the volume of at least 2.5 µL in order to perform the tagmentation reaction	
Quantified pool of purified cDNA PCR products	(optional) □ Use 1 µL to visualize the size profile of the pool of cDNA PCR products and confirm the final concentration of the mix on an Agilent Bioanalyzer High Sensitivity DNA chip	- Prepare 5 µL of a 1 ng/µL dilution of the pool of cDNA PCR products and measure the concentration of the dilution in triplicate by applying 3 × 1 µL on the Bioanalyzer chip - Quantify the cDNA products in the 100 bp – 9,000 bp range	

Fig. 2 (continued)

nanoCAGE – DAY 3

Quantified pool(s) of cDNA PCR products:

Before start checklist:

- □ Pipettes 1000 µL, 200 µL, 10 µL, 2 µL
- □ Filtered Tips Low-bind
- □ 0.2 mL PCR tubes low-bind + caps
- □ 1.5 mL microcentrifuge tubes low-bind
- □ 96-well PCR plate low-bind
- □ 384-well microplate for fluorescence assays
- □ Ultra-Pure Water
- □ Magnet for beads separation
- □ AMPure XP beads, resuspended and at room temperature
- □ Freshly prepared 70% EtOH, at room temperature
- □ Lambda DNA standard stock solution (1 ng/µL), on ice
- □ Quant-iT Picogreen Reagent (200x), in fridge until use
- □ TE buffer (1x), at room temperature
- □ Bioanalyzer DNA High Sensitivity DNA kit, at room temperature 30 min before use
- □ Tagment DNA Buffer / NT Buffer, on ice
- □ Amplicon Tagment Mix / Nextera PCR Mastermix, in freezer until needed
- □ 12 Nextera N-series Index primers / nanoCAGE custom S-series primer (10 µM), on ice
- □ NaOH 5 M, pH ≥ 13/ Qiagen EB buffer, at room temperature
- □ Illumina sequencing kit (e.g., MiSeq Reagent Kit v2) / PhiX Sequencing Control v3 / nanoCAGE Sequencing primer Read 1 (100 µM), in freezer until needed
- □ Quantified pool(s) of purified cDNA PCR products, on ice

WORKFLOW	INSTRUCTIONS	NOTES/OBSERVATIONS	TIME/DATE
Quantified pool of purified cDNA PCR products	□ Quantified pool(s) of purified cDNA PCR products prepared on Day 2 □ Prepare ≥ 2.5 µL of a dilution of the pool at 0.1 – 0.2 ng/µL	- Either one pool containing all cDNA PCR products samples or several pools containing cDNA PCR products samples grouped by experimental condition	
nanoCAGE library	Add in PCR tube low-bind: □ 5 µL of Tagment DNA Buffer □ 2.5 µL of the dilution of the pool of cDNA PCR products (0.25 – 0.5 ng) □ 2.5 µL of Amplicon Tagment Mix Incubate at 50°C for 10 min, Hold at 10°C □ Add 2.5 µL of NT buffer as soon as the temperature reaches 10°C Mix by pipetting 10x carefully and incubate 5 min at RT Add: □ 2.5 µL of a Nextera XT N-series Index primer □ 2.5 µL of nanoCAGE custom S-series primer (10 µM) □ 7.5 µL of Nextera PCR Mastermix Mix by pipetting carefully, spin down briefly Incubate at 72°C for 3 min, 95°C for 30 s, then perform 12 PCR cycles of 95°C for 10 s, 55°C for 30 s, 72°C for 1 min, and incubate at 72°C for 5 min, hold at 10°C	- Mix NT buffer carefully in order to avoid the formation of bubbles - If processing several pools of cDNA PCR products, choose one Nextera N-series Index primer per pool when performing the Reduced-Cycle PCR Amplification - If processing a single pool of cDNA PCR products, perform the PCR reaction using a single Nextera N-series Index primer and do not sequence the index. Alternatively, perform the PCR using 2.5 µL of an equimolar mix of the 12 Nextera N-series Index primers (i.e., mix 2 µL of each primer) and sequence the index	
Wash 3x 200 µL	□ Add 22.5 µL of AMPure XP beads at RT □ Mix by pipetting 10x slowly and incubate for 5 min □ Pellet on magnet and pipette off the supernatant □ Keep on magnet, wash 3x with 200 µL fresh 70% EtOH, do not disturb the pellet □ Resuspend pellet in 20 µL water, incubate at RT for 5 min □ Pellet beads on magnet, elute purified cDNA PCR products and transfer to a new PCR tube low-bind	- Aspirate supernatant carefully - In order to avoid losing the sample, be sure not to aspirate beads in the pipette tip together with the solution	
Purified nanoCAGE library	□ Prepare ≥ 20 µL of 1:10 and 1:20 dilutions of the purified nanoCAGE library	- If processing several nanoCAGE libraries prepared from different groups of samples, first quantify each library individually, then prepare a mix by adding equal quantities of each library and quantify dilutions of the pool as indicated	
	□ Use 3 × 1 µL of purified nanoCAGE library and 3 × 1 µL of the 1:10 and 1:20 dilutions to visualize the size profile of the nanoCAGE library and confirm the final concentration on an Agilent Bioanalyzer High Sensitivity DNA chip	- Quantify the nanoCAGE library by selecting nanoCAGE fragments in the 150 bp – 1,500 bp range	
	(optional) □ Use 3 µL of the dilutions 1:10 and 1:20 of the purified nanoCAGE library for quantification by Quant-iT PicoGreen dsDNA assay		
Quantified purified nanoCAGE library	□ Prepare the nanoCAGE library for sequencing according to the protocol and recommendations provided by Illumina for the sequencing system selected	- In general, sequence the nanoCAGE library at a concentration of 6 – 12 pM - Add 10% of PhiX Sequencing Control library to the final nanoCAGE sequencing sample - Output sequencing data as FASTQ files for bioinformatics analysis and demultiplexing based on barcode, index, and UMI informations	
Sequence the nanoCAGE library on an Illumina sequencing system			

Fig. 2 (continued)

Table 4
Summary of the different steps included in the nanoCAGE protocol

Step	Action	Time required
Day 1		
1	Retrieval of reagents and equipment preparation	30 min
2 (optional)	Pretreatment of RNA samples with Terminator 5′-Phosphate-Dependent Exonuclease	1 h
3	First-strand cDNA synthesis	2 h
4	Purification of first-strand cDNAs with Agencourt AMPure XP kit	1 h
5	Determination of optimal PCR cycle number by Real-Time quantitative PCR for second-strand cDNA synthesis	3 h
Day 2		
1	Retrieval of reagents and equipment preparation	30 min
2	Second-strand cDNA synthesis by cDNA PCR	2 h
3	Purification of cDNA PCR products with Agencourt AMPure XP kit	1 h
4	Quantification of cDNA PCR products by fluorimetry with Quant-iT PicoGreendsDNA Assay Kit	1 h
5	Pooling of the cDNA PCR products and quantification of the pool by fluorimetry with Quant-iT PicoGreendsDNA Assay Kit	45 min
6 (optional)	Visualization of cDNA PCR products pool profile and quantification with a Bioanalyzer DNA High Sensitivity chip	1 h
Day 3		
1	Retrieval of reagents and equipment preparation	30 min
2	Tagmentation followed by PCR amplification with the Nextera XT DNA Library Preparation kit	1 h 30 min
3	Purification of the nanoCAGE library with Agencourt AMPure XP kit	30 min
4	Quantification of the nanoCAGE library with a Bioanalyzer DNA High Sensitivity chip	1 h
5 (optional)	Quantification of the nanoCAGE library by fluorimetry with Quant-iT PicoGreendsDNA assay	30 min
6	Sequencing the nanoCAGE library	2 h for sample preparation, overnight or more for sequencing

6. If an excess of RNA sample is available, keep a backup to repeat the experiment. Do not work with the smallest possible RNA amount if additional RNA is available for the experiment. Using very small amounts of RNA can reduce the complexity of the nanoCAGE library.

3.2.2 RNA Concentration This protocol is optimized for as little as 50 ng of total RNA. For best results we therefore recommend using at least 50 ng of total RNA for nanoCAGE library preparation. However, total RNA ranging from 10 ng to 1250 ng has been used in different experiments.

3.2.3 Pretreatment of RNA Samples with an Exonuclease

Prior to nanoCAGE library preparation, total RNA samples can be treated with an exonuclease that digests only 5′-end phosphorylated RNAs. This will digest most rRNAs and most truncated mRNAs having a 5′-end phosphate after cleavage by an endonuclease. Therefore, this digesting step can reduce the number of reads obtained from rRNA while improving the rate of reads obtained from capped 5′-ends of mRNAs [26, 27] (*see* **Note 9**).

3.2.4 Other Considerations on RNA Samples

1. NanoCAGE has worked on degraded RNA samples, to the expense of its capacity of accurately detecting promoters (e.g., RNA partially shipped at room temperature). However, in those experiments nanoCAGE could be used for transcript identification. Therefore, you can try nanoCAGE library preparation in cases where no high-quality RNA can be obtained. This should not imply that we recommend working with poor RNA samples, which should normally be avoided.
2. The nanoCAGE protocol has also been used to prepare libraries starting from bacterial RNA that should be void of any cap structures. In this case, nanoCAGE could detect the promoters in the absence of the cap, where most likely the phosphorylated ends of high-quality RNA transcripts were instead used in the template switching reaction. For instance, we could prepare a nanoCAGE library from 500 ng of FirstChoice E. coli Total RNA.
3. NanoCAGE has been successfully used on non-polyadenylated RNA, on polysomal RNA, on RNA extracted from histological sections that were stained by immunohistochemistry, as well as on insects (Apis mellifera) [33], on plant (Arabidopsis thaliana) [34] and yeast (Schizosaccharomyces pombe) RNA (where rRNA depletion was found to be necessary).
4. Some starting materials may lead to RNA preparations having DNA or carbohydrate contaminations, which have not been tested systematically for problems during the RT reaction. But we know that some contaminants coming from cell culture can disturb the nanoCAGE protocol as for example, we could not prepare libraries from K562 cells induced with hemin (a known reverse transcriptase inhibitor). Similarly, some plant and brain samples can be difficult to process.

3.2.5 Considerations About Experiment Design and Multiplexing

In principle, we advise to work with three biological replicas for each data point (*see* ref. [33] for an example on how to use of nanoCAGE in expression profiling). Although descriptive CAGE experiments have been done with only one library per data point, this is not suitable for comparing expression levels across different samples and doing a statistical analysis.

One sequencing run on Illumina MiSeq should produce more than 10 million reads, while in one run of the Illumina HiSeq, each

lane should provide more than 100 million reads. This protocol provides information so that up to 96 nanoCAGE libraries can be pooled together. Therefore, the number of libraries per Illumina lane should be decided according to the desired sequencing depth per library. For instance, if 5 million tags are needed, then 20 libraries can be pooled per lane. Another possibility is to pool more libraries, and sequence them on multiple lanes. For example, 40 libraries pooled together and sequenced in two lanes should also give more than 5 million reads per library on average. The number of reads per nanoCAGE library to aim for is an essential part of the experimental planning that should be discussed in advance with the researchers who will analyze the data.

This protocol introduces a new strategy for the multiplexing of nanoCAGE libraries, using Illumina indexes introduced during tagmentation. The original multiplexing strategy with barcodes is still used [28] and both can be combined, for hundreds or even thousands of samples. The difference between multiplexing with barcodes or indexes is shown in Table 5.

These multiplexing strategies can be combined depending on the nature of the experiment to be performed. For example:

1. Use one barcode per sample, and pool immediately after the reverse transcription. This strategy saves reagents. Alternatively, pool all the cDNAs that will need the same number of cycles for the cDNA PCR.

Table 5
Difference between multiplexing strategies with barcodes and indexes

	Barcodes	Indexes
Timing of the multiplexing	Early (reverse transcription)	Late (tagmentation or library PCR, *see* Subheadings 3.12 and 3.13) [18]
Introduced by	Template Switching Oligonucleotide	Nextera XT N-Series Index primer in the Reduced–Cycle PCR Amplification step of the tagmentation reaction (*see* Subheading 3.12), or Reverse Index primer in the library PCR (*see* Subheading 3.13) [18]
Support in Illumina sequencers	No	Native
Possibility of bias	Higher	Lower
Sequence read	Read 1	Index
Position relative to the cDNA	5′-end, before the UMI and the GGG linker	3′-end, between the sequencing adaptor and the flow cell adaptor

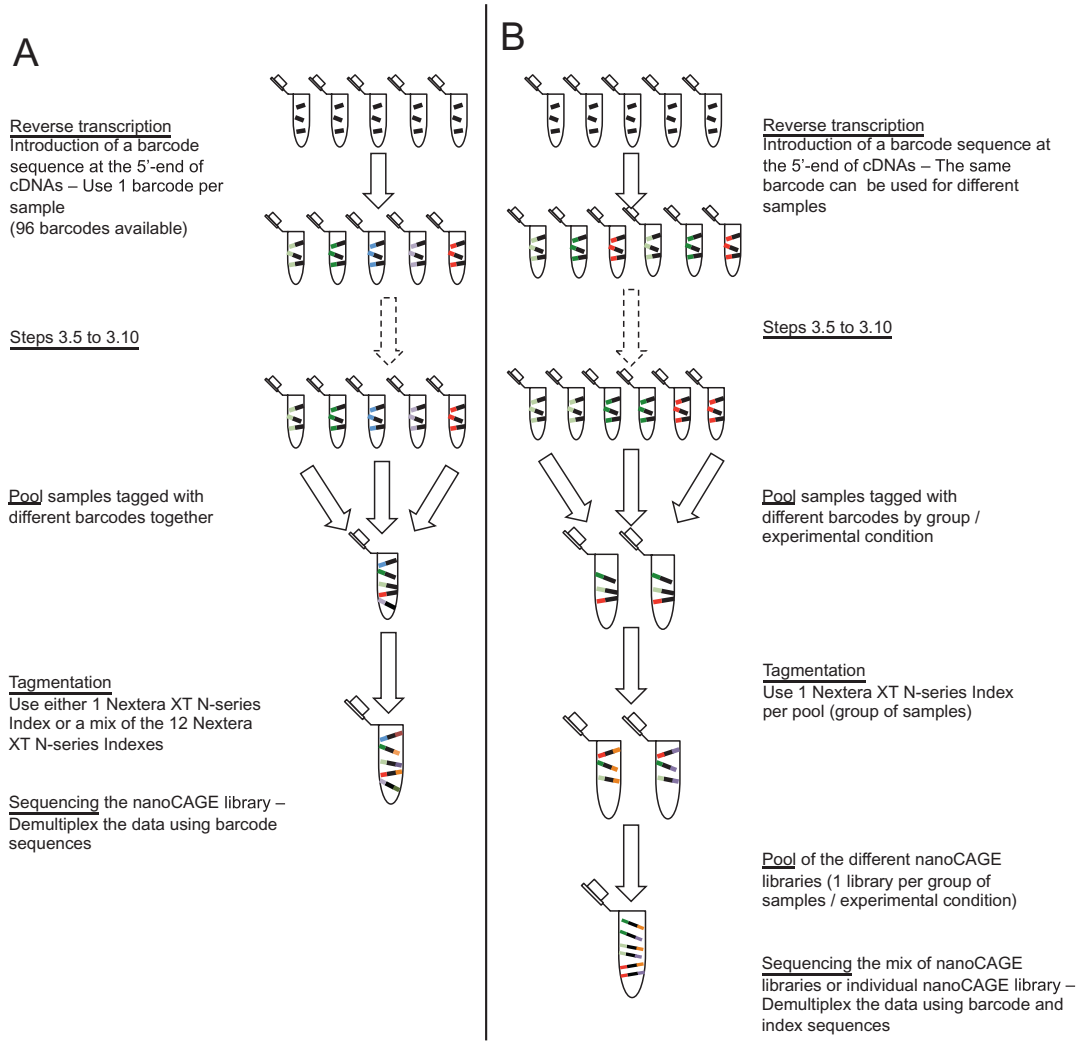


Fig. 3 Presentation of the different strategies offered for the multiplexing of nanoCAGE libraries. (**a**) Multiplexing strategy of nanoCAGE libraries based on barcodes. (**b**) Multiplexing strategy of nanoCAGE libraries based on indexes, or using a combination of barcodes and indexes

2. Pool cDNA PCR products tagged with different barcodes before the tagmentation step (Fig. 3a). This allows to quality-control the cDNAs for each sample separately.
3. Use the same barcode for all the nth technical replicate of a sample, and then multiplex with different indexes (Fig. 3b). This ensures that the replicates are exposed to the same sequence biases during template switching [28].
4. Use the same barcode for each sample, and multiplex on indexes. This way, the demultiplexing will be done entirely during the sequencing process.

3.3 (Optional Step) Pretreatment of Total RNA Samples with Terminator 5′-Phosphate-Dependent Exonuclease

Timing: 1 h.

This step is performed prior to nanoCAGE library preparation to digest rRNAs and most of the truncated mRNAs contained in a total RNA sample. This allows reducing the number of sequencing reads obtained from rRNA and increasing the number of reads obtained from capped 5′-ends of mRNAs. This treatment is suited for preparing nanoCAGE libraries from RNA samples containing a high proportion of rRNAs. Because we only tested this step when starting the protocol with at least 500 ng of total RNA, and the Terminator enzyme may digest more than 90 % of the RNA in the sample, we recommend performing this step only when a comfortable amount of total RNA is available and allows for the user to prepare replicates of the nanoCAGE library.

5. Prepare on ice, in a 0.2 mL low binding PCR tube, a master mix containing one volume of the reagent described in Table 6 per total RNA sample to be digested. Add an extra volume of each reagent to account for pipetting imprecisions.
6. Transfer 2 μL of master mix in a PCR tube low-bind, and then add 1 μL of total RNA sample (i.e., ≥500 ng) in the tube. Mix by pipetting slowly up and down and spin down briefly.
7. Incubate the samples at 30 °C for 1 h in a thermocycler.
8. Transfer the tubes on ice immediately until the next step.

3.4 First-Strand cDNA Synthesis

Timing: 1.5 h.

This step produces first-strand cDNAs from total RNA by random priming. The 5′-ends of RNAs are captured through template switching in the presence of a trehalose/sorbitol solution. A barcode sequence specific for each RNA sample is introduced by the template switching oligonucleotide at the 3′-end of the first-strand cDNA molecules synthesized by the reverse transcriptase.

3.4.1 Preparation of Primers Premix Stock Solutions

1. Prepare primers premix stock solutions in 0.2 mL low binding PCR tubes by mixing on ice the reagents indicated in Table 7. Prepare one tube of primers premix stock solution per barcoded

Table 6
Mix of reagents to prepare for the pretreatment of total RNA samples with the Terminator 5′-Phosphate-Dependent Exonuclease

Reagent	Volume (μL)	Final concentration
Water	1.3	–
SuperScript III First-Strand Buffer (5×)	0.2	3.33×
Terminator Exonuclease (1 U/μL)	0.5	0.5 U
Total Volume	2	–

Table 7
Preparation of primers premix stock solutions

Reagent	Volume (μL)	Concentration in primers premix	Final concentration in RT mix
Trehalose/sorbitol stock solution (0.66 M/3.3 M) (*see* **Note 3**)	8	0.528 M/2.64 M	52.8 mM/264 mM
Reverse transcription random primers (100 μM)	1	10 μM	1 μM
Template switching oligonucleotide (1 mM)	1	100 μM	10 μM
Total volume	10	–	–

Table 8
Mix of reagents to prepare for the denaturation of RNA samples

Reagent	Volume (μL)
Primers premix stock solution (*see* **Note 3**)	1
Total RNA (i.e., 50 ng) or Terminator-treated RNA	1
Total Volume	2

template switching oligonucleotide to be used in the nanoCAGE experiment.

2. Store the primers premix stock solutions at −20°C.

3.4.2 Denaturation of RNA

1. Mix on ice in 0.2 mL low binding PCR tubes 1 μL of each total RNA sample (i.e., 50 ng) or 1 μL of Terminator-treated RNA with 1 μL of primers premix prepared in step 1, as indicated in Table 8 (*see* **Note 10**). Prepare a negative control by mixing 1 μL of a primers premix stock solution with 1 μL of water. We also recommend preparing a positive control by mixing 1 μL of a commercially available RNA sample or a well-characterized total RNA with 1 μL of a primers premix stock solution.
2. Briefly spin down the tubes, and incubate the mixtures at 65°C for 10 min in a thermocycler.
3. Transfer the tubes on ice immediately for at least 2 min.

3.4.3 Reverse Transcription

1. Prepare on ice 8 μL of reverse transcription reaction mix per sample as indicated in Table 9.

   Prepare a master mix when processing several samples and replicates.
2. Add 8 μL of the reverse transcription reaction mix prepared in **step 6** to each RNA and primers premix solution prepared in **step 3**.

Table 9
Mix of reagents to prepare for the reverse transcription

Reagent	Volume per sample (μL)	Final concentration in RT reaction
First-strand buffer, 5×	2	1×
DTT (0.1 M)	1	10 mM
dNTPs (2.5 mM each)	2.5	0.625 mM each
Betaine (5 M)	1.5	0.75 M
SuperScript III (200 U/μL) (*see* **Note 1**)	1	20 U/μL
Total volume	8	–

Table 10
Temperature program used for the reverse transcription

Step number	Temperature (°C)	Time
1	22	10 min
2	50	30 min
3	75	15 min
4	4	Hold

3. Mix carefully on ice by pipetting (total volume = 10 μL, *see* **Note 3**).
4. Incubate the samples in a thermocycler and perform the reverse transcription reaction following the temperature program described in Table 10.
5. The reaction products are the first-strand cDNAs.

3.5 Purification of First-Strand cDNAs with Agencourt AMPure XP Kit

Timing: 1 h.

This step is necessary to remove smaller artifacts and primer dimers from the cDNAs that can make it difficult to determine the number of PCR cycles required for the cDNA PCR.

1. Add 18 μL of Agencourt AMPure XP beads to each 10 μL of first-strand cDNAs, and mix thoroughly by pipetting ten times up and down. Incubate the mix at room temperature for 5 min.
2. Transfer the microtubes containing the beads/cDNA solutions onto a magnetic stand, and wait for 5 min. Aspirate the cleared solution, and discard (*see* **Note 11**). Do not disturb the beads while removing the supernatant.

3. Keep the samples on the magnetic stand, and wash the beads with 200 μL of a freshly prepared 70 % (vol/vol) ethanol solution. After 30 s, aspirate the cleared solution and discard. Repeat this washing step for two additional times.
4. Remove the samples from the magnetic stand and wait 2 min for the ethanol remaining on the tube walls to dry up.
5. Add 25 μL of water to each tube, and mix by pipetting 15 times up and down to completely elute the cDNA from the beads.
6. Incubate the samples at room temperature for 5 min, and then transfer the tubes on the magnetic stand for 5 min to collect the beads. Transfer the supernatants to new tubes.
7. The resulting products are the purified first-strand cDNAs. Keep them on ice until the next step.

   Note: This is a safe stopping point. Store the samples at −20°C.

3.6 Real-Time Quantitative PCR

Timing: 2.5 h.

This step determines the number of amplification cycles required for each sample during the following cDNA PCR. This allows preparing equivalent amounts of DNA for each sample that will be subsequently used in the tagmentation reaction.

1. Prepare on ice a master mix containing 3 volumes (3 × 8.5 μL) of Real-Time qPCR mixture per purified first-strand cDNA sample (including the controls), as indicated in Table 11. Analyze each sample and controls in triplicates. Prepare one more volume of master mix for a negative control of the Real-Time qPCR reaction.
2. Transfer 8.5 μL of master mixture per well of a 96-well Real-Time PCR plate.

Table 11
Mix of reagents to prepare for the Real-Time quantitative PCR

Reagent	Volume (μL)	Final concentration
Takara SYBR Premix Ex Taq (2×) (Tli RNase H Plus)	5	1×
cDNA PCR primer, Forward (10 μM)	0.1	100 nM
cDNA PCR primer, Reverse (10 μM)	0.1	100 nM
Rox Reference Dye II (50×)	0.2	1×
Water	3.1	–
Total volume	8.5	–

Table 12
Temperature program used for the Real-Time quantitative PCR

Step number	Temperature (°C)	Time	Fluorescence Detection step
1	95	30 s	
2	95	15 s	
3	65	10 s	
4	68	2 min	X (amplification)
5	Repeat steps 2–4	40 cycles	
6	95	15 s	
7	65	1 min	
8	65–95 by increments of 0.3	15 s	X (melting curve)

3. Add 1.5 µL of purified first-strand cDNA or negative control to each well and mix by pipetting slowly. Add 1.5 µL of water to the negative control of the Real-Time qPCR reaction.
4. Seal the plate with an adhesive PCR seal. Ensure a tight seal with a sealing applicator.
5. Centrifuge the plate for 30 s at 1200 × *g*.
6. Transfer the 96-well plate in a Real-Time PCR system. Perform the reaction setup according to the manufacturer's instructions, following the temperature program described in Table 12.

3.7 Analysis of Real-Time Quantitative PCR Results and Determination of Optimal Cycle Number for the cDNA PCR

1. Determine the "Cycle threshold" values (Ct) for each cDNA sample and for the controls after Real-Time quantitative PCR by following the instructions presented below (*see* **Note 12**).
   - Adjust the baseline of the Real-Time PCR reaction carefully to allow an accurate determination of the Ct values.
   - Manually set the fluorescence threshold of the nanoCAGE Real-Time qPCR assay to half of the exponential zone of the amplification curves (Fig. 4a).
   - Determine the Ct value assigned to each sample replicate.
2. Calculate the average Ct values for each sample and for the controls.

The average Ct value for purified cDNA is usually ~20 to 34, depending on the source of RNA (50 ng). The reverse transcription reaction negative control should either give no signal, or a Ct value significantly higher (≥35) than for the samples (the formation of primer dimers may cause the negative control to produce a Ct value).

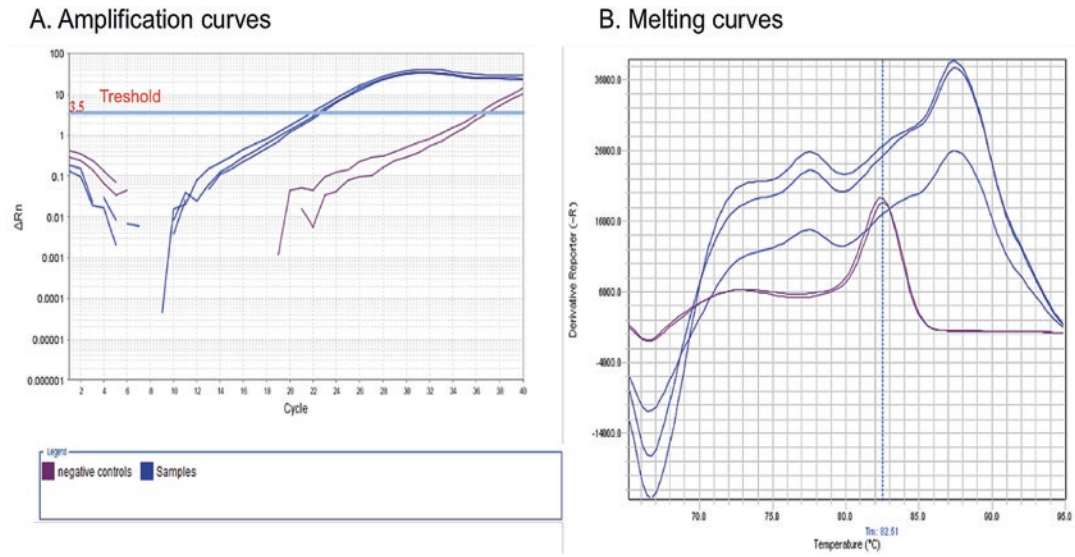


Fig. 4 Analysis of Real-Time quantitative PCR data. (**a**) A cycle threshold (Ct) value is determined by Real-Time quantitative PCR as the number of cycles required for the fluorescent signal to cross the threshold, i.e., to exceed fluorescence background levels. The optimum number of cycles (N) to be performed for the cDNA PCR is determined based on the average Ct value of the sample. The reverse transcription reaction negative control should either give no signal, or a Ct value significantly higher (≥35) than for the samples. (**b**) The melting curve analysis of the amplified cDNA must show a fluorescence signal covering a large range of temperatures, attesting for the presence of abundant number of cDNA molecules of different sizes and sequences. The negative control sample shows a single peak at a lower melting temperature, indicating the presence of primer dimers

Table 13
Determination of the number of cycles (N) to perform for the cDNA PCR based on the Ct value obtained in Real-Time quantitative PCR

Average Ct value determined for a cDNA sample by Real-Time qPCR	Number (N) of cDNA PCR cycles to perform
≤15	N = 15
15< average Ct value ≤30	N = Ct value
30< average Ct value <35	N = 30
≥35	N = 30, or troubleshoot first (*see* **steps 4–7** of Subheading 3.7)

3. Determine the optimal number of cDNA PCR cycles (N) based on the average Ct values obtained from Real-Time quantitative PCR data, as shown in Table 13.

   In case of the Ct value is unexpectedly higher than 35, troubleshoot by following the instructions provided in **steps 4–7**.

4. Compare the Ct value obtained for this sample to the Ct values obtained for the other samples and to the Ct value obtained for

the reverse transcription positive control, if any (Fig. 4a). If the Ct value obtained for this sample is significantly different, the total RNA might contain inhibitors of the reverse transcription reaction.

5. Compare the profile of the melting curve obtained for this sample to the profiles obtained with the other samples and to the profile obtained for the reverse transcription negative control (Fig. 4b). If the profile of the melting curve obtained for this sample overlaps with the negative control, it might only contain primer dimers and no amplification products.
6. If an RNA sample backup is available, recheck the quality of the sample with a Bioanalyzer and eventually repurify it. Alternatively, redo the reverse transcription using a greater amount of total RNA as starting material.
7. If no RNA sample backup is available, split the volume of purified cDNA sample by 2 and perform 2× cDNA PCR reactions as detailed in Subheading 3.8. (N = 30, perform 30 PCR cycles).
8. Purify the 2× PCR reactions as indicated in Subheading 3.9, and elute the purified cDNA PCR products from each reaction using the same 25 µL of water to concentrate the cDNA PCR products. The quantity and dynamic ranges of libraries needing more than 30 PCR cycles may be suboptimal. Only use material if you have no other options.

3.8 Second-Strand cDNA Synthesis and Amplification by cDNA PCR

Timing: 1.5 h.

1. Prepare on ice a PCR master mix for N reactions as indicated in Table 14. The number of reactions is 1 per purified first-strand cDNA sample and control sample, plus one for a PCR negative control, plus one to account for pipetting imprecisions (*see* **Note 13**).
2. Transfer 30 µL of PCR master mix per well of a 96-well low binding PCR plate.

Table 14
Mix of reagents to prepare for the cDNA PCR

Reagent	Volume (µL)	Final concentration
Kapa HiFi HotStart Ready Mix (2×)	25	1×
cDNA PCR primer, Forward (10 µM)	0.5	100 nM
cDNA PCR primer, Reverse (10 µM)	0.5	100 nM
Water	4	–
Total volume	30	–

Table 15
Temperature program used for the cDNA PCR

Step Number	Temperature (°C)	Time
1	95	3 min
2	98	20 s
3	65	15 s
4	72	2 min
5	Repeat steps 2–4	Na cycles
6	72	2 min
7	4	Hold

a N = optimum number of PCR cycles determined for each sample by Real-Time quantitative PCR ($15 \leq N \leq 30$)

3. Add 20 μL of purified first-strand cDNA sample, including the controls. Add 20 μL of water in the well prepared for the PCR negative control.
4. Close the wells of the PCR plate with an adhesive PCR seal or with PCR tubes strip caps. Vortex and centrifuge briefly at 1200 × *g*.
5. Perform the cDNA PCR reaction in a thermocycler, following the temperature program described in Table 15 (*see* **Note 14**).
6. The reaction products are the cDNA PCR products.

3.9 Purification of cDNA PCR Products by Agencourt AMPure XP Kit

Timing: 1 h.

1. Add 90 μL of Agencourt AMPure XP beads in each 50 μL cDNA PCR reaction and mix thoroughly by pipetting ten times. Incubate the mix at room temperature for 5 min.
2. Repeat **steps 2–7** of Subheading 3.5, using 25 μL of water for the elution of each cDNA PCR products sample.
3. The recovered samples are the purified cDNA PCR products. Keep them on ice until the next step.
   Note: This is a safe stopping point. Store the samples at −20°C.

3.10 Quantification of Purified cDNA PCR Products by Quant-iT PicoGreendsDNA Assay Kit

Timing: 1 h.

1. Prepare a 1:5 dilution of each purified cDNA PCR product by adding 3 μL of DNA sample in 12 μL of TE buffer 1× in 0.2 mL low binding PCR tubes.
2. For the quantification of the cDNA PCR products, prepare a Lambda DNA standard curve by performing serial dilutions of a Lambda DNA standard stock solution (1 ng/μL), as indicated in Table 16.

Table 16
Serial dilutions of Lambda DNA standard for PicoGreen dsDNA Assays

Standard ID	Volume of TE buffer 1× (µL)	Volume of Lambda DNA solution
S1	–	60 µL of a 1 ng/µL Lambda DNA standard stock solution
S2	30	30 µL from tube S1
S3	27	3 µL from tube S1
S4	27	3 µL from tube S2
S5	27	3 µL from tube S3
S6	27	3 µL from tube S4
S7	30	–

Table 17
Mixing of Lambda DNA standards with PicoGreen reagent for PicoGreen dsDNA Assays

Standard ID	Volume of Lambda DNA standard solution (µL)	Volume of 1:200 diluted Quant-iT Picogreen reagent (µL)	Total volume (µL)	Final concentration of Lambda DNA in Quant-iT Picogreen assay (ng/µL)
S1	20	20	40	500
S2	20	20	40	250
S3	20	20	40	50
S4	20	20	40	25
S5	20	20	40	5
S6	20	20	40	2.5
S7	20	20	40	0

3. Transfer 20 µL of each standard into new low binding PCR tubes.
4. Prepare a 1:200 dilution of Quant-iT PicoGreen Reagent in a 1.5 mL low binding microcentrifuge tube by adding 2 µL of PicoGreen reagent into 398 µL of TE buffer 1×. Adjust the total volume of diluted reagent prepared to the number of samples to quantify.
5. As indicated in Table 17, add 20 µL of diluted PicoGreen Reagent in the 20 µL prepared for each Lambda DNA standard and add 15 µL of diluted PicoGreen Reagent in the 15 µL of 1:5 dilution prepared for each sample.

The final concentration of each Lambda DNA standard is indicated in Table 17 and the cDNA PCR products are diluted ten times.

6. Transfer 3 × 10 μL of each standard and each sample prepared in **step 3** in triplicate wells of a 384-well microplate designed for fluorescence assays.
7. Measure the fluorescence of each sample with a microplate reader according to the instructions provided by the manufacturer of the Quant-iT PicoGreen dsDNA Assay Kit.
8. Determine the average concentration of each cDNA PCR product based on the fluorescence values obtained from the replicates of each sample.

3.11 Pooling of the Purified cDNA PCR Products and Quantification of the Mix by Quant-iT PicoGreendsDNA Assay Kit

Timing: 1 h (2 h, if performing the optional step).

This step allows preparing and quantifying the mix(es) of cDNA PCR products that will subsequently be used in the tagmentation reaction.

1. Based on the quantification of the purified cDNA PCR products, prepare a mix containing equivalent quantities of each sample to be sequenced (i.e., 2–20 ng of each cDNA PCR products sample). We recommend mixing at least 2 ng of each sample at this step.

   As indicated in Subheading 3.2.5 and shown in Fig. 3, it is also possible to prepare several pools of cDNA-derived PCR products at this stage. Samples tagged with different barcodes can be mixed as groups representing different experimental conditions or sample replicates.
2. Measure the concentration(s) of the pooled cDNA PCR products in triplicate by Quant-iT PicoGreen dsDNA Assay Kit, following the steps described in Subheading 3.10. The concentration of the mix should be at least 0.1 ng/μL and the volume of at least 2.5 μL to perform the tagmentation reaction.
3. (optional step) Visualize the size profile(s) of the pool(s) of cDNA PCR products and confirm the final concentration(s) of the mix(es) on an Agilent Bioanalyzer High Sensitivity DNA chip (Fig. 5). Based on the concentration determined in **step 2**, prepare 5 μL of a 1 ng/μL dilution of the pool of cDNA PCR products and measure the concentration of the dilution in triplicate by applying 3 × 1 μL on the Bioanalyzer chip. For the quantification of the pool with the Bioanalyzer software, follow the steps indicated in Subheading 3.15, using the 100 bp–9000 bp size range for the quantification of cDNA PCR products.
4. These are the pooled cDNA PCR products.

3.12 Preparation of a nanoCAGE Library with Nextera XT DNA Library Preparation Kit

Timing: 1.5 h.

According to Illumina (Nextera XT Sample Preparation Guide, 15031942 Rev. C), "The Nextera XT DNA Library Preparation Kit uses an engineered transposome to simultaneously fragment

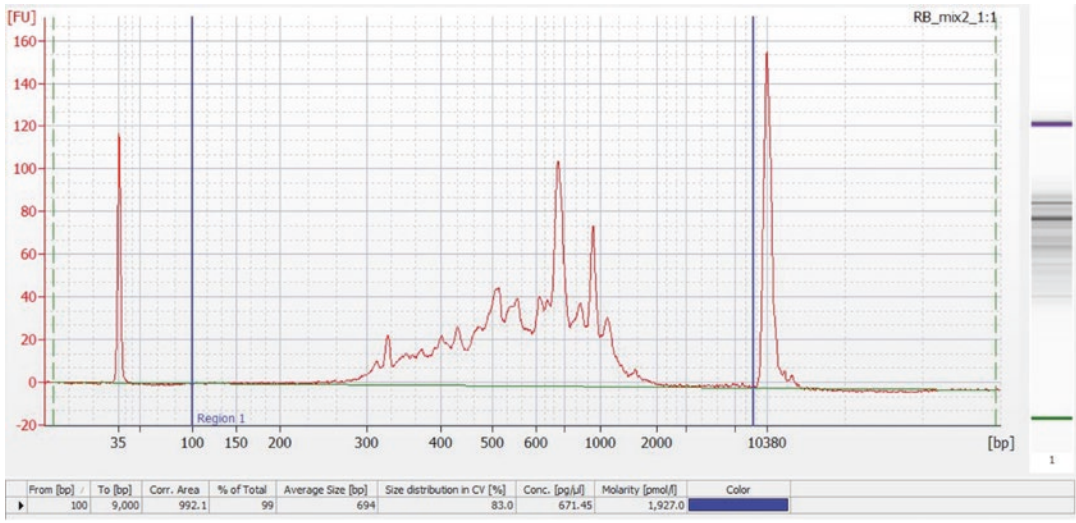


Fig. 5 Visualization of the size-profile and quantification of a pool of cDNA PCR products on an Agilent Bioanalyzer High Sensitivity DNA chip.22 cDNA PCR products were prepared from rat Purkinje cell's RNA samples and tagged with different barcodes. A pool of all cDNA PCR products was prepared by adding an equivalent amount (2 ng) from each individual sample. The concentration of the mix was adjusted to 1 ng/μL based on PicoGreendsDNA Assay quantification, and 1 μL of the dilution was applied in triplicate on an Agilent Bioanalyzer High Sensitivity DNA chip. The pool of cDNA PCR products was quantified using the Bioanalyzer software to select DNA fragments with a size included in the 100–9000 bp range

and tag ("tagment") input DNA, adding unique adapter sequences in the process. A limited-cycle PCR reaction uses these adapter sequences to amplify the insert DNA." The PCR reaction also adds index sequences at the 3′-end of the nanoCAGE DNA fragments (Fig. 1), thus enabling sequencing of pooled nanoCAGE libraries on any Illumina Sequencing System.

The tagmentation reaction advantageously replaced the library PCR step described in the previous version of the nanoCAGE protocol (*see* Subheading 3.13) as it requires very low amounts of DNA as input material, allows the multiplexing of nanoCAGE libraries in a single reaction, and provides a more optimal size distribution for sequencing.

3.12.1 Tagmentation Reaction

1. Prepare a dilution of the pooled cDNA PCR products to adjust the final concentration to 0.1–0.2 ng/μL. Keep the sample on ice.
2. Using Illumina's Nextera XT DNA Library Preparation Kit, prepare on ice, in 0.2 mL low binding PCR tube(s), the mixture(s) indicated in Table 18.
3. Mix the reagents by pipetting ten times up and down slowly, and incubate in a thermocycler as shown in Table 19.
4. As soon as the temperature reaches 10 °C, add 2.5 μL of NT buffer to neutralize the reaction.

Table 18
Mix of reagents to prepare for the tagmentation reaction

Reagent	Volume (μL)	Total quantity
Tagment DNA Buffer	5	–
Pool of cDNA PCR products (0.1–0.2 ng/μL)	2.5	0.25–0.5 ng
Amplicon Tagment Mix	2.5	–
Total volume	10	–

Table 19
Temperature program used for the tagmentation reaction

Step Number	Temperature (°C)	Time
1	55	10 min
2	10	Hold

Table 20
Mix of reagents to prepare for the Reduced-Cycle PCR Amplification

Reagent	Volume (μL)	Final concentration
Sample	12.5	–
nanoCAGE custom S-series primer (10 μM)	2.5	1 μM
Nextera XT N-series index primer (*see* **Note 14**)	2.5	1 μM
Nextera PCR Mastermix	7.5	–
Total volume	25	–

5. Mix by pipetting ten times up and down carefully to avoid the formation of bubbles and incubate the sample(s) for 5 min at room temperature.

3.12.2 Reduced-Cycle PCR Amplification

1. Choose one Nextera XT N-series index primer per pooled cDNA PCR product (*see* **Note 15**).
2. Add the reagents detailed in Table 20 into the sample(s) prepared at the **step 5** of Subheading 3.12.1.
3. Mix carefully by pipetting up and down to prevent the formation of bubbles. Centrifuge the tube briefly.
4. Incubate the sample(s) in a thermocycler following the PCR program described in Table 21.
5. The product of the PCR reaction is the nanoCAGE library.

Table 21
Temperature program used for the Reduced-Cycle PCR Amplification

Step number	Temperature (°C)	Time
1	72	3 min
2	95	30 s
3	95	10 s
4	55	30 s
5	72	1 min
6	Repeat steps 3–5	12 cycles
7	72	5 min
8	10	Hold

3.13 (Optional Alternative Step) Preparation of nanoCAGE Libraries by Library PCR

Timing: 2.5 h.

Despite the clear advantages provided by the tagmentation reaction in terms of yield, multiplexing, nanoCAGE fragment size distribution, and bioinformatics, it is also possible to prepare nanoCAGE libraries by performing a library PCR reaction as previously described by Salimullah et al. [18].

The standard library PCR reaction adds Illumina sequencing adaptors sequences at the 5′- and 3′-ends of the purified cDNA PCR products. However, the nanoCAGE fragments produced by library PCR have a broader size distribution profile on a Bioanalyzer chip, ranging from around 200–300 bp to over 8000 bp. The fragments larger than 1500 bp cannot be used to generate sequencing clusters on Illumina's MiSeq and HiSeq platforms. Therefore, the quantification of nanoCAGE libraries produced with this method, prior to sequencing, should be performed carefully with the Bioanalyzer software by precisely selecting fragments with a size included in the 150–1500 bp range only.

The library PCR can also be used to add an Index sequence (*see* Table 22) at the 3′-end of the purified cDNA PCR products to produce sequenceable nanoCAGE fragments having the same structure as those obtained with Illumina's Nextera XT DNA Library Preparation Kit (Fig. 1). Hence, for the multiplexing of nanoCAGE libraries using different Indexes, the library PCR should be performed, for each purified cDNA PCR products sample (or group of pooled purified cDNA PCR products samples tagged with different barcodes), using the forward nanoCAGE custom S-series primer in combination with one of the library PCR reverse Index primers described in Table 22. Although we have not tested it, it also appears possible to perform the library PCR reaction using an equimolar mix of the 12 library PCR reverse Index primers.

Table 22
Detailed sequences of the 12 library PCR reverse Index primers

Index ID	Index sequence	Library PCR reverse Index primer sequence (5′→3′)
1	ATCACG	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
2	CGATGT	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
3	TTAGGC	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
4	TGACCA	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
5	ACAGTG	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
6	GCCAAT	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
7	CAGATC	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
8	ACTTGA	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
9	GATCAG	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
10	TAGCTT	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
11	GGCTAC	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
12	CTTGTA	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Table 23
Mix of reagents to prepare for the library PCR

Reagent	Volume (μL)	Final concentration
Kapa HiFi HotStart Ready Mix (2×)	25	1×
nanoCAGE custom S-series primer (10 μM)	1.5	300 nM
Library PCR reverse Index primer (10 μM)	1.5	300 nM
Water	2	–
Total volume	30	–

We found a significant improvement in the efficiency of the library PCR reaction when using the Kapa HiFi HotSart Ready Mix. By following the protocol described below with this mix, we could prepare nanoCAGE libraries starting from 16 ng of cDNA PCR products instead of the 50 ng of starting material initially required [18].

1. Prepare on ice a library PCR master mix for N reactions as indicated in Table 23. The number of reactions is 1 per purified cDNA PCR products sample and control sample, plus one for a library PCR negative control, plus one to account for pipetting imprecisions.
2. Transfer 30 μL of library PCR master mix per well of a low binding 96-well PCR plate.
3. Add 20 μL of purified cDNA PCR products sample, including the controls (total volume of each PCR reaction = 50 μL). Add 20 μL of water in the well prepared for the library PCR negative control.
4. Close the wells of the PCR plate with a cap or with an adhesive PCR seal. Vortex and centrifuge briefly at 1200 × *g*.
5. Perform the library PCR reaction in a thermocycler, following the temperature program described in Table 24.
6. The products of the reaction are the library PCR products.
7. Purify the library PCR products with AMPure XP beads as indicated in Subheading 3.9.

   Note: This is a safe stopping point. Store the samples at –20°C.
8. Quantify the purified library PCR products by Quant-iT PicoGreendsDNA Assay Kit as indicated in Subheading 3.10.
9. Pool the purified library PCR products to be sequenced as described in Subheading 3.11 to obtain the nanoCAGE library.

Table 24
Temperature program used for the library PCR

Step number	Temperature (°C)	Time
1	95	3 min
2	98	20 s
3	55	15 s
4	72	2 min
5	98	20 s
6	65	15 s
7	72	2 min
8	Repeat steps 5–7	6 cycles
9	4	Hold

10. Quantify the nanoCAGE library prior to sequencing as indicated in Subheading 3.15, selecting the 150–1500 bp size range for the quantification of the nanoCAGE fragments.
11. Sequence the nanoCAGE library on an Illumina sequencer according to the instructions provided in Subheadings 3.16 and 3.17.
12. Analyze the sequencing data according to the instructions provided in Subheading 3.18 and visualize nanoCAGE tags as described in Subheading 3.19.

3.14 Purification of the nanoCAGE Library with Agencourt AMPure XP Kit

Timing: 30 min.

1. Add 22.5 μL of Agencourt AMPure XP beads to the 25 μL PCR reaction(s) from **step 10** of Subheading 3.12.2, and mix by pipetting ten times up and down slowly to avoid the formation of bubbles.
2. Incubate the mix at room temperature for 5 min.
3. Repeat **steps 2–7** of Subheading 3.5, using 20 μL of water for the elution of the nanoCAGE library.
4. The resulting product is the purified nanoCAGE library.

Note: This is a safe stopping point. Store the sample(s) at −20°C.

3.15 Quantification of the Purified nanoCAGE Library with a Bioanalyzer DNA High Sensitivity Chip

Timing: 1 h—(1.5 h, if performing the optional step).

It is important to check the removal of PCR primer dimers/artifacts, as well as to know the profile of the nanoCAGE library (e.g., size distribution, molar concentration) to calculate the amount of molecules to apply for the sequencing (*see* **Note 16**).

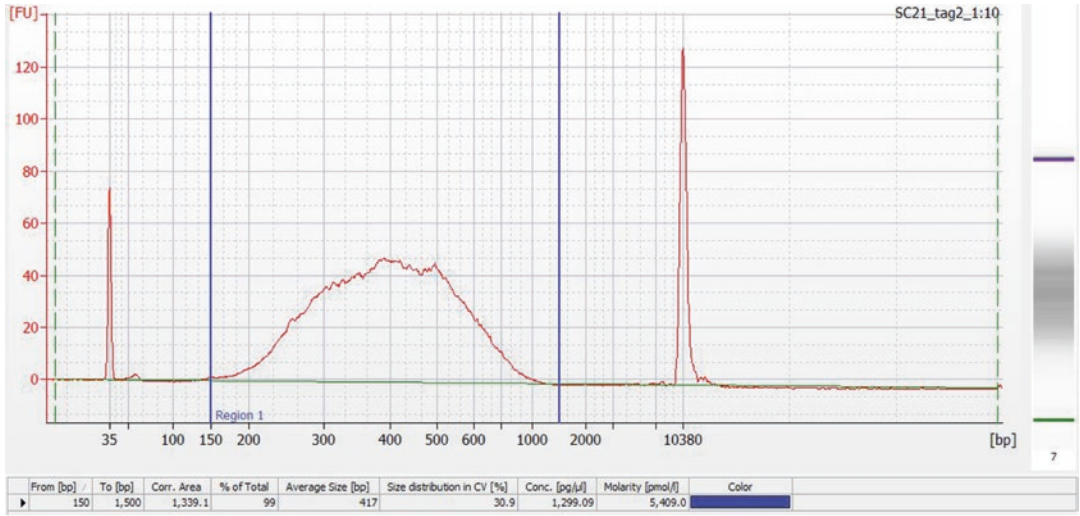


Fig. 6 Visualization of the size-profile and quantification of a nanoCAGE library on an Agilent Bioanalyzer High Sensitivity DNA chip. The nanoCAGE library sample was diluted ten times in ultrapure water and analyzed in triplicate on an Agilent Bioanalyzer High Sensitivity DNA chip. The molarity and the average size of the library were determined with the Bioanalyzer software by selecting nanoCAGE fragments with a size ranging from 150 to 1500 bp

1. Prepare a 1:10 (2 µL of the purified nanoCAGE library + 18 µL of water) and a 1:20 (1 µL of the purified nanoCAGE library + 19 µL of water) dilution of the library. If processing several nanoCAGE libraries tagged by different indexes, prepare and quantify dilutions of each library at this step.
2. Visualize the DNA size-profile and measure the concentration of the purified nanoCAGE library on an Agilent Bioanalyzer High Sensitivity DNA chip (Fig. 6). Measure the concentrations of the non-diluted library and the 1:10 and 1:20 dilutions prepared in **step 1** in triplicates by applying 3 × 1 µL of each sample on the Bioanalyzer chip. There should be no highly concentrated peak shorter than 150 bp visible on the Bioanalyzer profile.
3. Follow the different steps described below to quantify and determine the average size of the nanoCAGE library with the Agilent 2100 Expert Bioanalyzer software. This procedure assumes that only the nanoCAGE fragments with a size ranging from 150 to 1500 bp are efficiently used for the generation of sequencing clusters on the Illumina MiSeq and HiSeq sequencers.

1. Open the nanoCAGE library DNA analysis file (.xad).
2. Double-click on the electropherogram data of the sample to be analyzed.
3. Click on the "Region table" icon under the page. Scroll into the table.

4. Right-click the mouse button. Select "Modify region."
5. Input 150 and 1500 in the boxes "From [ ] bp" and "To [ ] bp," respectively.
6. Click the "OK" button to determine the molarity.

4. Calculate the average molarity of the purified nanoCAGE library based on the values obtained for the triplicates of each dilution. If processing several nanoCAGE libraries tagged by different indexes, prepare a pool by mixing equivalent quantities of each non-diluted library. We recommend mixing the nanoCAGE libraries based on their average molarities. Then repeat the **steps 1–4 (or 5)** to quantify the pool of nanoCAGE libraries prior to sequencing.
5. (optional step) Confirm the concentration of the nanoCAGE library by Quant-iT PicoGreendsDNA Assay Kit. Perform the assay following the procedure detailed in Subheading 3.10, using 3 μL of the 1:10 and 1:20 dilutions prepared in **step 1**. Determine the average concentration of the purified nanoCAGE library in ng/μL, based on the values obtained for the triplicates of each dilution. Calculate the molarity of the purified nanoCAGE library as follows:

Library concentration (inng/μL): as determined by PicoGreendsDNA Assay

Average fragment size (in bp): as determined by the Bioanalyzer software

Library molecular weight (g/mol): Average fragment size (bp) × 650 g/mol/bp

$$\text{Molar concentration}\left(nM\right) = \frac{\text{Library concentration}\left(ng\ /\ \mu L\right) \times 10^6}{\text{Library molecular weight}\left(\text{g}\ /\ mol\right)}$$

3.16 Sequencing the nanoCAGE Library

Timing: 2 h for the preparation, overnight or more for the sequencing

See Subheading 3.17 for a detailed description of how to prepare nanoCAGE libraries for sequencing on Illumina MiSeq Sequencing System.

1. Prepare the nanoCAGE library sample to be sequenced on a MiSeq or HiSeq Sequencing System by following either Illumina's recommendations or your service provider's protocol for sample preparation according to the device of your choice. We advise the user to prepare a fresh dilution of the purified nanoCAGE library for the preparation of the sequencing sample.
2. Sequence the nanoCAGE library at the concentration recommended in Table 25, based on the average size of the purified nanoCAGE library determined with the Bioanalyzer DNA High Sensitivity chip analysis in **step 4** of Subheading 3.15.

Table 25
Determination of the nanoCAGE library concentration to apply for the sequencing

Average size of the nanoCAGE library from Bioanalyzer (in bp)	Concentration of nanoCAGE library to apply for cluster generation
250	6–12 pM
500	6–12 pM
1000–1500	12–20 pM

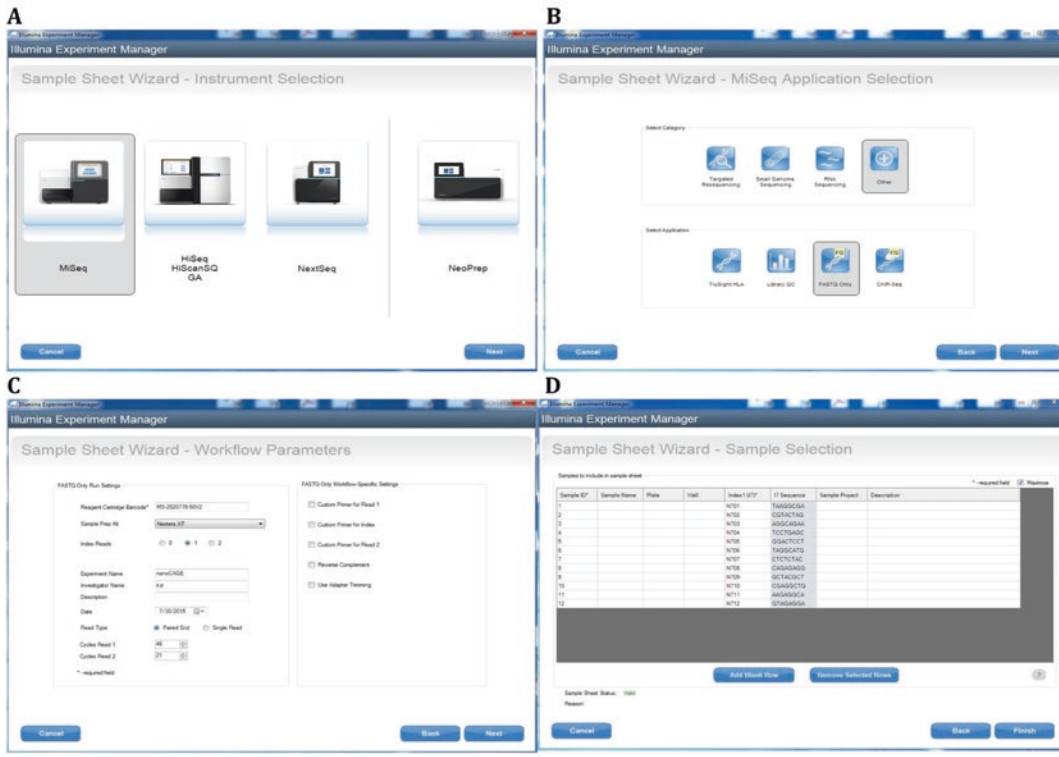


Fig. 7 Screenshots of Illumina Experiment Manager software. Panels (**a**–**d**) show the different steps and parameters required for preparing the Illumina sample sheet when sequencing a nanoCAGE library on a MiSeq sequencing system with a MiSeq Reagent Kit v2

3. Use the Illumina Experiment Manager software to adjust the sequencing parameters (type of sequencing data files generated, number of sequencing cycles performed for each read, etc.), and indicate the sequences of the Index used in the preparation of the nanoCAGE library for further demultiplexing of the sequencing data (Fig. 7). The Illumina sample sheet prepared with the software (Fig. 8) can further be double-checked and eventually modified using a text editor.

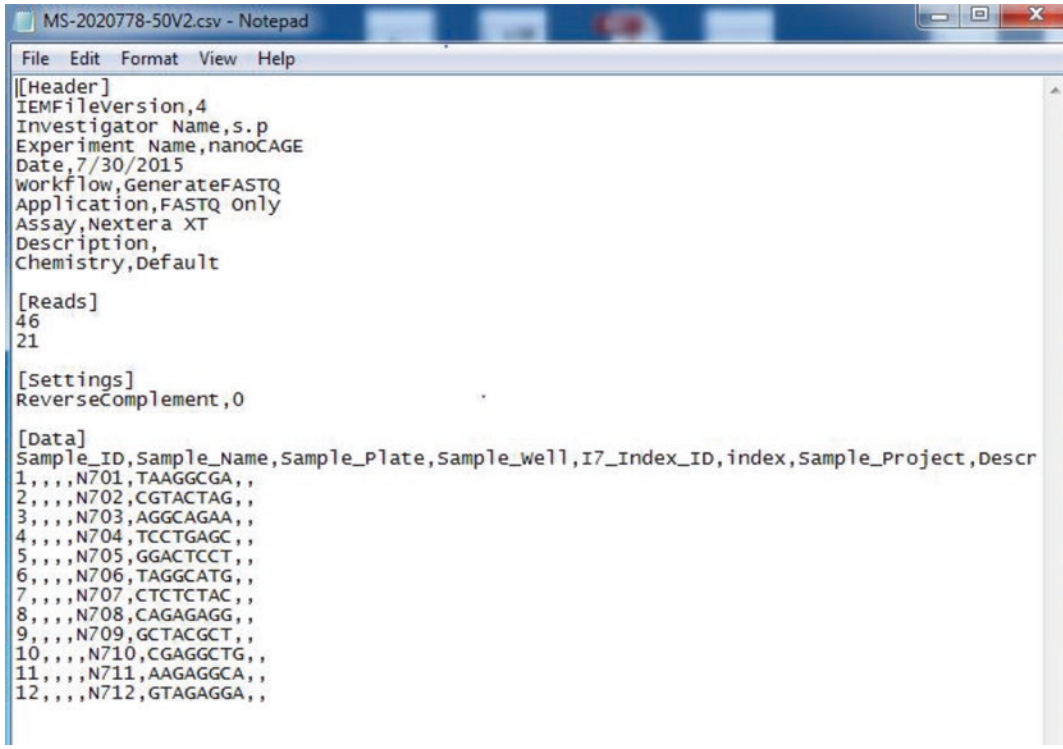


Fig. 8 Example of Illumina sample sheet. The sample sheet was designed for sequencing a nanoCAGE library prepared with a mix of 12 Nextera XT N-series index primers on a MiSeq sequencer with a MiSeq Reagent Kit v2

4. Once the sequencing started, important quality metrics generated by the Real Time Analysis software of the sequencing system (i.e., total number of clusters and quality of the sequencing reads generated) can be monitored in real time using Illumina Sequencing Analysis Viewer software.

3.17 Specific Instructions for Sequencing nanoCAGE Libraries on Illumina MiSeq Sequencing System

Timing: 2 h for the preparation, overnight for the sequencing.

This section assumes that the nanoCAGE library will be sequenced using a MiSeq Reagent Kit v2. Follow the manufacturer's instructions for preparing and loading the MiSeq cartridge, with the following modifications (some original instructions are repeated unchanged for the sake of clarity).

1. Prepare a fresh dilution of the purified nanoCAGE library to be sequenced. Adjust the concentration of the dilution to 0.6–1.2 nM (600–1200 pmol/L) in a total volume of at least 10 μL, as indicated in **step 2** of Subheading 3.16. Keep the sample on ice until use.
2. Thaw the tube of Hybridization buffer (Hyb buffer) included in the sequencing kit at room temperature, mix well by vortexing, and then transfer the tube on ice until use.

3. Thaw the reagents contained in the cartridge included in the sequencing kit by transferring the cartridge into a polystyrene box filled with an adapted volume of ultrapure water. Incubate for 30 min. Once the reagents are completely thawed, mix the reagents by inverting the cartridge gently several times, and then transfer the cartridge on ice.
4. Prepare 500 μL of a fresh 0.1 M NaOH solution by adding 10 μL of 5 M NaOH, pH ≥ 13–490 μL of water. Vortex the tube and centrifuge briefly.
5. Prepare a 20 pM PhiX Sequencing Control stock solution (*see* **Note 17**) in a 1.5 mL low binding microcentrifuge tube by mixing on ice 4 μL of Qiagen EB buffer with 1 μL of 10 nM PhiX Control v3. Add 5 μL of 0.1 M NaOH prepared in **step 4** and mix carefully by pipetting ten times up and down, then incubate the tube at room temperature for 5 min. Add 490 μL of Hyb buffer thawed in **step 2** to the DNA mix, vortex the tube and centrifuge briefly, then transfer the tube on ice until use (*see* **Note 18**).
6. Prepare 100 μL of an 8 pM PhiX Sequencing Control solution by mixing 40 μL of 20 pM PhiX stock solution prepared in **step 5** with 60 μL of Hyb buffer in a 1.5 mL low binding microcentrifuge tube. Mix well by vortexing, centrifuge briefly, and transfer the tube on ice.
7. Transfer 6 μL of 0.1 M NaOH in a 1.5 mL low binding microcentrifuge tube; add 6 μL of the nanoCAGE library dilution prepared in **step 1**, and mix carefully by pipetting ten times up and down. Incubate the sample for 5 min at room temperature, and then add 588 μL of Hyb buffer to the sample. Vortex the tube and centrifuge briefly, transfer the sample on ice.
8. In another 1.5 mL microcentrifuge tube, transfer 558 μL of the nanoCAGE library prepared in **step 7** and 62 μL of the 8 pM PhiX Sequencing Control solution prepared in **step 6**. Mix well by vortexing, centrifuge briefly, and transfer the tube on ice. The final nanoCAGE sequencing sample has a concentration of 6–12 pM and contains 10 % of the PhiX Sequencing Control library.
9. Pierce the aluminum foil covering the slots no. 17 and no. 12 of the sequencing cartridge with a plastic tip.
10. Load carefully 600 μL of the nanoCAGE sequencing sample prepared in **step 8** into the slot no. 17 ("Load sample") of the cartridge.
11. Load 3.4 μL of nanoCAGE Sequencing primer stock solution (100 μM) into the slot no. 12 of the sequencing cartridge. Mix the primers thoroughly by pipetting up and down using a long 200 μL filtered tip.

12. Sequence the nanoCAGE library according to Illumina's instructions on how to operate the MiSeq device. For instance, if sequencing paired-end a nanoCAGE library prepared with a mix of the 12 Nextera N-Series Indexes, use the Illumina Experiment Manager software as indicated in Fig. 7 to set an Illumina sample sheet as shown in Fig. 8. Generate sequencing data as FASTQ files for further bioinformatics analysis. Perform 46 sequencing cycles for the read 1 and 21 cycles for the read 2. 8 cycles are used by default for the sequencing of the Index. Indicate the sequence of each Index for the demultiplexing of the sequencing reads.

3.18 Analysis of nanoCAGE Sequencing Data

While a full description of the bioinformatics analysis is beyond the scope of this protocol, here we are providing some directions for quick quality controls on the sequence output. The nanoCAGE protocol is for the Illumina sequencers, and so is this section.

3.18.1 Quantity of Reads

The quantity of reads should be in the order of magnitude of 10 millions or above on MiSeq and 100 millions or above on HiSeq 2000. More than 200 million reads on HiSeq 2000 is not necessarily a problem, but it indicates that the lanes are close to overloading with DNA templates, which means that if there is variability in the loading, there will be cases of overloading from time to time. Lower quantity of reads can also be a sign of overloading, since in that case the basecaller may have difficulties to analyze the raw images that contain clusters that are too close to each other. A high cluster density together with a low frequency of reads passing the filter also suggests overloading. Inspection of the raw sequencing images can also help to distinguish between over- and under-loading.

3.18.2 Quality and Processing

Currently, the sequencing output is given in a FASTQ format. Tools like FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) or FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can be used to compute global statistics on the Phred quality scores.

Despite the cDNA PCR, some reads contain either oligonucleotide artifacts or empty constructs (no cDNA between the linkers). The tool "TagDust2" [35] can be used to detect and remove these artifacts, as well as the reads aligning to rRNA, demultiplex the barcodes, and transfer the UMI in the sequence name.

3.18.3 Mismatches at the 5′-End

The hallmark of CAGE is the addition of extra guanosines at the 5′-end of the tags. But for nanoCAGE, these guanosines become part of the linker, and most of them are removed from the sequence. However, the mismatches at the 5′-end of the tags are still expected to be a majority of G. The contrary may indicate a problem in the library, for instance degradation of the starting material (and therefore a much lower fraction of capped 5′ends is covered by the library).

3.19 Visualization of nanoCAGE Sequencing Data

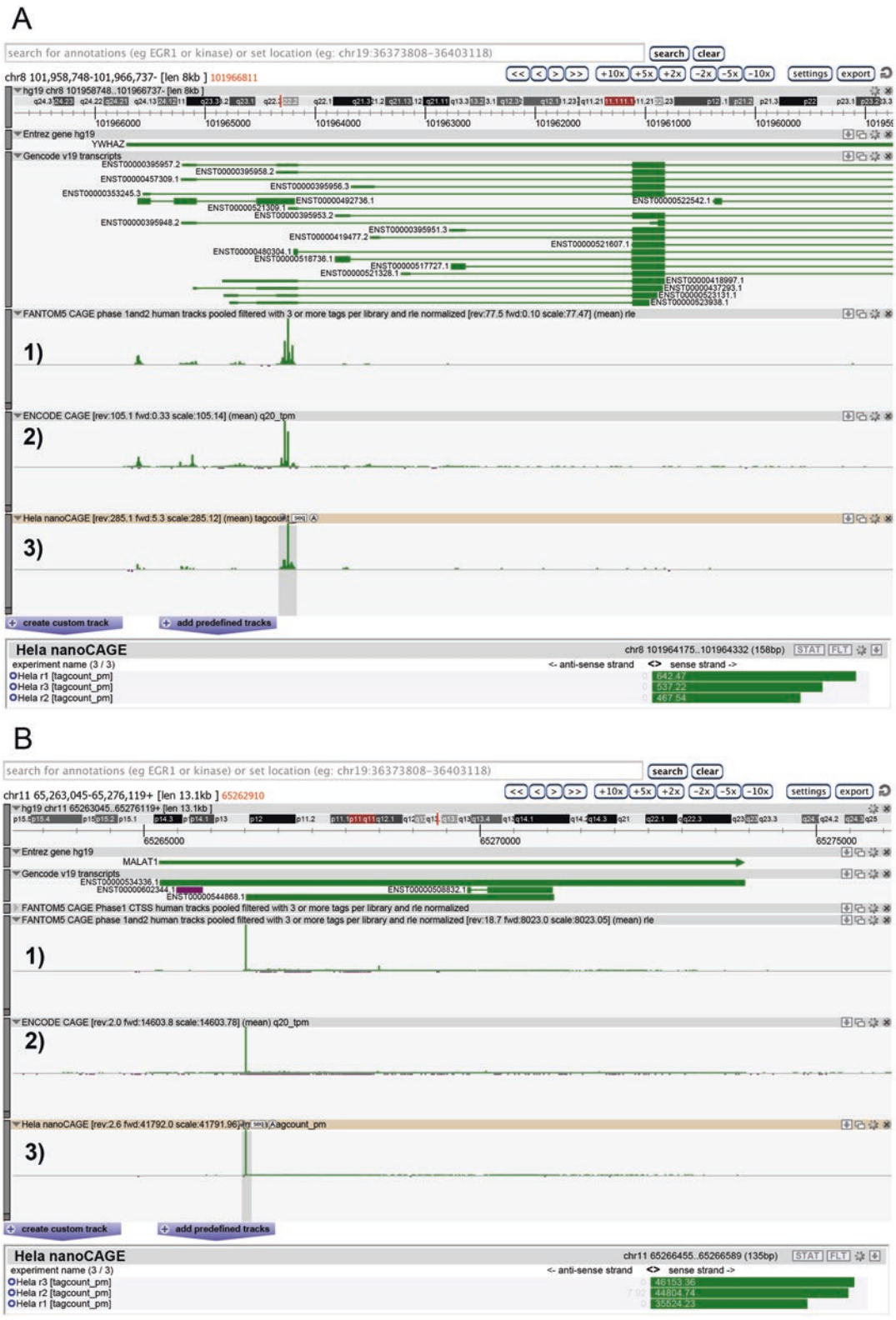
After demultiplexing, filtering, and alignment on a reference genome, it is possible to create files in BAM and BED12 formats that contain information about the coordinates of each CAGE tag specific of each RNA sample (one file per RNA sample), respectively, before and after removal of PCR duplicates and non-properly paired data. These files can be uploaded on a genome browser such as Zenbu (http://fantom.gsc.riken.jp/zenbu/) or UCSC (https://genome.ucsc.edu/). We recommend Zenbu [36] for its support of the BED12 format, allowing for a switch between single-nucleotide-resolution TSS expression histogram and block-style visualization of the paired-end information, using the same data source.

The distribution of nanoCAGE tags in a sample can be compared with reference CAGE assemblies like the FANTOM5 data, publicly available in Zenbu and UCSC (http://fantom.gsc.riken.jp/views/). The expression of a given gene can be quantified by counting the number of nanoCAGE tags at specific TSS coordinates. In order to compare the expressions of a given gene between different samples, the nanoCAGE tags must be displayed as transcript per million (tpm) in the genome browser as shown in Fig. 9.

4 Notes

1. We use and recommend SuperScript III for first-strand cDNA synthesis. We successfully prepared nanoCAGE libraries using different commercially available reverse transcriptases (i.e., SuperScript II, Life Technologies, and PrimeScript, TaKaRa Bio). However, we found that, in our hands, SuperScript III was the best enzyme for promoter detection and library preparation. We have not tested the SuperScript IV (Life Technologies) reverse transcriptase yet.

Fig. 9 Zenbu screenshots showing the quantitative distribution of 5′-ends of nanoCAGE transcript fragments on the human genome reference assembly hg19. NanoCAGE data were obtained from triplicate samples prepared from HeLa cells total RNA. The cDNAs prepared from each sample replicate were tagged with different barcodes at their 5′-end during the reverse transcription. Then, the cDNA PCR products obtained for each replicate were pooled and tagged with a single index at their 3′-end after the tagmentation reaction. The nanoCAGE library was sequenced on a MiSeq system using the MiSeq Reagent Kit v2. Upon bioinformatics analysis, filtering, and demultiplexing of the data based on barcode sequences, the nanoCAGE sequencing reads were aligned to the reference genome assembly hg19 to create BED12 files that contain the coordinates of the nanoCAGE transcript fragments found in each sample replicate. Panels **a** and **b** show the quantitative distribution of the 5′-ends of the nanoCAGE fragments mapping to the promoter region of the YWHAZ coding gene (**a**) and to the MALAT1 noncoding RNA (**b**). On tracks (1) and (2) are shown the CAGE reference HeLa data publicly available from the FANTOM5 and ENCODE consortia, respectively. On track (3) are shown the nanoCAGE data. The quantification of each nanoCAGE replicate is displayed under track (3), normalized as tag per millions of transcript counts. On panel **a**, CAGE data mapping at different coordinates and representing possible alternative promoters of the gene can be visualized on the three tracks


A
search for annotations (eg EGR1 or kinase) or set location (eg: chr19:36373808-36403118)
search
clear
chr8 101,958,748-101,966,737- [len 8kb] 101966811
settings
export
hg19 chr8 101958748..101966737- [len 8kb]
101966000
101965000
101964000
101963000
101962000
101961000
101960000
Entrez gene hg19
YWHAZ
Gencode v19 transcripts
ENST00000395957.2
ENST00000395958.2
ENST00000457309.1
ENST00000395956.3
ENST00000353245.3
ENST00000492736.1
ENST00000522542.1
ENST00000521309.1
ENST00000395953.2
ENST00000395948.2
ENST00000395951.3
ENST00000419477.2
ENST00000521607.1
ENST00000480304.1
ENST00000518736.1
ENST00000517727.1
ENST00000521328.1
ENST00000418997.1
ENST00000437293.1
ENST00000523131.1
ENST00000523938.1
FANTOM5 CAGE phase 1and2 human tracks pooled filtered with 3 or more tags per library and rle normalized [rev:77.5 fwd:0.10 scale:77.47] (mean) rle
1)
ENCODE CAGE [rev:105.1 fwd:0.33 scale:105.14] (mean) q20_tpm
2)
Hela nanoCAGE [rev:285.1 fwd:5.3 scale:285.12] (mean) tagcount
3)
create custom track
add predefined tracks
Hela nanoCAGE
chr8 101964175..101964332 (158bp)
STAT
FLT
experiment name (3 / 3)
<- anti-sense strand
sense strand ->
Hela r1 [tagcount_pm]
Hela r3 [tagcount_pm]
Hela r2 [tagcount_pm]
642.47
537.22
467.54
B
search for annotations (eg EGR1 or kinase) or set location (eg: chr19:36373808-36403118)
search
clear
chr11 65,263,045-65,276,119+ [len 13.1kb] 65262910
settings
export
hg19 chr11 65263045..65276119+ [len 13.1kb]
65265000
65270000
65275000
Entrez gene hg19
MALAT1
Gencode v19 transcripts
ENST00000534336.1
ENST00000602344.1
ENST00000508832.1
ENST00000544868.1
FANTOM5 CAGE Phase1 CTSS human tracks pooled filtered with 3 or more tags per library and rle normalized
FANTOM5 CAGE phase 1and2 human tracks pooled filtered with 3 or more tags per library and rle normalized [rev:18.7 fwd:8023.0 scale:8023.05] (mean) rle
1)
ENCODE CAGE [rev:2.0 fwd:14603.8 scale:14603.78] (mean) q20_tpm
2)
Hela nanoCAGE [rev:2.6 fwd:41792.0 scale:41791.96] (mean) tagcount_pm
3)
create custom track
add predefined tracks
Hela nanoCAGE
chr11 65266455..65266589 (135bp)
STAT
FLT
experiment name (3 / 3)
<- anti-sense strand
sense strand ->
Hela r3 [tagcount_pm]
Hela r2 [tagcount_pm]
Hela r1 [tagcount_pm]
46153.36
44804.74
35524.23

2. Trehalose and sorbitol should be of high quality and essentially free of heavy metals that could cause nucleic acid degradation.
3. Trehalose/sorbitol solutions are viscous; mix very carefully by pipetting up and down slowly for at least ten times.
4. The Nextera XT N-series index primers are included in the Nextera XT DNA Library Preparation Kit (Illumina) or can also be purchased separately as Nextera XT DNA Library Preparation Index Kit (Illumina).
5. The Read 1 nanoCAGE sequencing primer must be added manually into the cartridge provided with Illumina's sequencing kit. We usually add and mix carefully 3.4 μL of a 100 μM primer stock solution into the slot number 12 of a MiSeq Reagent Kit v2 (Illumina) (*see* Subheading 3.17).
6. The Index and Read 2 Nextera sequencing primers are included in the MiSeq Reagent Kit v2 (Illumina), but are provided separately in the TruSeq Dual Index Sequencing Primer Box (Paired End) (Illumina) when sequencing the nanoCAGE library on HiSeq 2000 (Sequencing Dual-Indexed Libraries on the HiSeq System User Guide, 15032071, Rev. B).
7. The layout of the nanoCAGE bench workflow was inspired in part by the "Quick Start Lambda Burn-in checklist protocol" provided by Oxford Nanopore Technologies.
8. Standard RIN values were established for working with mammalian RNA, and are not reliable for other RNA samples, e.g., from insects or plants. Note further that the Bioanalyzer is very sensitive to contaminations by salt or ethanol, and can therefore strongly underestimate the quality of RNA preparations.
9. Because of the presence of a cap structure, short RNAs captured by template switching should in principle not be affected by the Terminator 5′-Phosphate-Dependent Exonuclease treatment. These short RNAs molecules should therefore be detected in the nanoCAGE libraries prepared with this method.
10. When processing Terminator-treated RNA, we advise to use the 2 μL of the remaining digested RNA for preparing replicates of the reverse transcription reaction.
11. Aspirate carefully. In order to avoid losing the sample, be sure not to aspirate beads in the pipette tip together with the solution.
12. For a detailed explanation on baselines and Ct values, refer to a qPCR manual, for instance, the Real-time PCR handbook from Life Technologies.
13. We observed an improvement of Real-Time qPCR reaction performances (lower Ct values) when using the Kapa Sybr Fast qPCR Master Mix.
14. When different numbers of cDNA PCR cycles have to be performed for the different samples, try to group the samples by

number of cycles to be performed and do the PCR amplification on different thermocyclers, if available.

15. Choose the Nextera XT N-series indexes according to the instructions provided by the manufacturer. In case only one *pool of cDNA PCR products* was made (same index used for all samples), program the sequencer or instruct your facility to skip index sequencing. Alternatively, it is also possible to perform the Reduced-Cycle PCR Amplification reaction with a mix of the 12 Nextera XT N-series index primers. This introduces a diversity in the sequence of the index and prevents the appearance of ultra-bright clusters during the sequencing of the index read, which can further impair the quality of the sequencing and alter the number of reads passing the filter which are finally used for the downstream analysis of the sequencing data.
16. Alternatively, it is also possible to quantify the nanoCAGE library by Real-Time quantitative PCR (i.e., with KAPA Library Quantification Kits For Illumina sequencing platforms, Kapa Biosystems). However, we found it more accurate and convenient to perform the quantification by the methods described in this protocol.
17. According to Illumina: "The PhiX Control v3 (Illumina) is a reliable, adapter-ligated library used as a control for Illumina sequencing runs. The library is derived from the small, well-characterized PhiX virus genome, offering several benefits for sequencing and alignment. The PhiX library provides a quality control for cluster generation, sequencing, and alignment, and a calibration control for cross-talk matrix generation, phasing, and prephasing. It can be rapidly aligned to estimate relevant sequencing by synthesis (SBS) metrics such as phasing and error rate." The PhiX Control library is generally mixed with the nanoCAGE library to represent 10 % of the final sequencing sample and thus can provide information about the quantity of nanoCAGE sample actually used for the generation of the sequencing clusters. It is therefore a useful control for troubleshooting cluster generation problems, allowing the user to quickly determine whether an error is related to the preparation or to the quantification of the nanoCAGE library.
18. The 20 pM PhiX Sequencing Control stock solution can be stored for up to 3 weeks at −20°C.

Acknowledgments

We thank Alexandre Fort for critically reading the manuscript and his helpful comments, and Laia Masvidal Sanz for helpful discussions and suggestions on the nanoCAGE protocol. This work was founded by a Research Grant from the Japanese Ministry of

Education, Culture, Sports, Science and Technology (MEXT) to the RIKEN Center for Life Science Technologies, and a JSPS Grant-in-Aid for Young Scientists A (number 25710018).

References

1. Shiraki T, Kondo S, Katayama S et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100:15776–15781. doi:10.1073/pnas.2136655100
2. Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2:495–502. doi:10.1038/nmeth768
3. Kodzius R, Kojima M, Nishiyori H et al (2006) CAGE: cap analysis of gene expression. Nat Methods 3:211–222
4. Carninci P, Nishiyama Y, Westover A et al (1998) Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. Proc Natl Acad Sci USA 95:520–524. doi:10.1073/pnas.95.2.520
5. Carninci P, Shiraki T, Mizuno Y et al (2002) Extra-long first-strand cDNA synthesis. Biotechniques 32:984–985
6. Spiess A-N, Ivell R (2002) A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. Anal Biochem 301:168–174. doi:10.1006/abio.2001.5474
7. Suzuki H, Forrest ARR, Nimwegen E et al (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet 41:553–562. doi:10.1038/ng.375
8. Balwierz PJ, Pachkov M, Arnold P et al (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res 24:869–884. doi:10.1101/gr.169508.113
9. Andersson R, Gebhard C, Miguel-Escalada I et al (2014) An atlas of active enhancers across human cell types and tissues. Nature 507:455–461. doi:10.1038/nature12787
10. Kellis M, Wold B, Snyder MP et al (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA 111:6131–6138. doi:10.1073/pnas.1318948111
11. Plessy C, Bertin N, Takahashi H et al (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat Methods 7:528–534. doi:10.1038/nmeth.1470
12. Kratz A, Beguin P, Kaneko M et al (2014) Digital expression profiling of the compartmentalized translatome of Purkinje neurons. Genome Res 24:1396–1410. doi:10.1101/gr.164095.113
13. Klerk E, Dunnen JT, 't Hoen PAC (2014) RNA sequencing: from tag-based profiling to resolving complete transcript structure. Cell Mol Life Sci 71:3537–3551. doi:10.1007/s00018-014-1637-9
14. Klerk E, 't Hoen PAC (2015) Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet 31:128–139. doi:10.1016/j.tig.2015.01.001
15. Harbers M, Kahl G (eds) (2012) Tag-based next generation sequencing. Wiley-Blackwell, Hoboken, NJ
16. Carninci P, Kvam C, Kitamura A et al (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics 37:327–336. doi:10.1006/geno.1996.0567
17. https://en.wikipedia.org/wiki/Cap_analysis_gene_expression
18. Salimullah M, Mizuho S, Plessy C, Carninci P (2011) NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. Cold Spring Harb Protoc. doi:10.1101/pdb.prot5559
19. Hirzmann J, Luo D, Hahnen J, Hobom G (1993) Determination of messenger RNA 5′-ends by reverse transcription of the cap structure. Nucleic Acids Res 21:3597–3598. doi:10.1093/nar/21.15.3597
20. Ohtake H, Ohtoko K, Ishimaru Y, Kato S (2004) Determination of the capped site sequence of mRNA based on the detection of cap-dependent nucleotide addition using an anchor ligation method. DNA Res 11:305–309. doi:10.1093/dnares/11.4.305
21. Lavie L, Maldener E, Brouha B et al (2004) The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res 14:2253–2260. doi:10.1101/gr.2745804
22. Kulpa D, Topping R, Telesnitsky A (1997) Determination of the site of first strand transfer during Moloney murine leukemia virus reverse transcription and identification of strand transfer-associated reverse transcriptase errors.

EMBO J 16:856–865. doi:10.1093/emboj/16.4.856

23. Islam S, Zeisel A, Joost S et al (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 11:163–166. doi:10.1038/nmeth.2772
24. König J, Zarnack K, Rot G et al (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 17:909–915. doi:10.1038/nsmb.1838
25. Kivioja T, Vähärautio A, Karlsson K et al (2011) Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 9:72–74. doi:10.1038/nmeth.1778
26. Fejes-Toth K, Sotirova V, Sachidanandam R et al (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature 457:1028–1032. doi:10.1038/nature07759
27. Yan B, Ma J (2012) Promoter-associated RNAs and promoter-targeted RNAs. Cell Mol Life Sci 69:2833–2842. doi:10.1007/s00018-012-0953-1
28. Tang DTP, Plessy C, Salimullah M et al (2013) Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. Nucleic Acids Res. doi:10.1093/nar/gks1128
29. Imbeaud S, Graudens E, Boulanger V et al (2005) Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. Nucleic Acids Res 33:1–12. doi:10.1093/nar/gni054
30. Mueller O, Schroeder A (2004) RNA Integrity Number (RIN)—Standardization of RNA quality control application. Nano:1–8
31. Rio DC (2015) Denaturation and electrophoresis of RNA with formaldehyde. Cold Spring Harb Protoc 2015:219–222. doi:10.1101/pdb.prot080994
32. Mansour FH, Pestov DG (2013) Separation of long RNA by agarose-formaldehyde gel electrophoresis. Anal Biochem 441:18–20. doi:10.1016/j.ab.2013.06.008
33. Khamis AM, Hamilton AR, Medvedeva YA et al (2015) Insights into the transcriptional architecture of behavioral plasticity in the honey bee Apis mellifera. Sci Rep 5:11136. doi:10.1038/srep11136
34. Cumbie JS, Ivanchenko MG, Megraw M (2015) NanoCAGE-XL and CapFilter: an approach to genome wide identification of high confidence transcription start sites. BMC Genomics 16:597. doi:10.1186/s12864-015-1670-6
35. Lassmann T (2015) TagDust2: a generic method to extract reads from sequencing data. BMC Bioinformatics 16:24. doi:10.1186/s12859-015-0454-y
36. Severin J, Lizio M, Harshbarger J et al (2014) Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat Biotechnol 32:217–219. doi:10.1038/nbt.2840

Chapter 5

Deep Cap Analysis of Gene Expression (CAGE): Genome-Wide Identification of Promoters, Quantification of Their Activity, and Transcriptional Network Inference

Alexandre Fort and Richard J. Fish

Abstract

Among the most significant findings of the post-genomic era, the discovery of pervasive transcription of mammalian genomes has tremendously modified our understanding of the genome output seen as RNA molecules. The increased focus on non-protein-coding genomic regions together with concomitant technological innovations has led to rapid discovery of numerous noncoding transcripts (ncRNAs). Biological relevance and functional roles of the vast majority of these ncRNAs remain largely unknown.

The cap analysis of gene expression (CAGE) technology allows accurate transcript detection and quantification without relying on preexisting transcript models. In combination with complementary data sets, generated using other technologies, it has been shown as an efficient approach for exploring transcriptome complexity.

Here, we describe the use of CAGE for the identification of novel noncoding transcripts in mammalian cells providing detailed information for basic data processing and advanced bioinformatics analyses.

Key words Transcriptomics, Noncoding RNAs, Cap Analysis of Gene Expression (CAGE), Transcription start sites (TSSs)

1 Introduction

The comprehensive annotation of protein-coding genes encoded by mammalian genomes, with 19,814 and 22,032 genes for the human (GENCODE V22) [1] and mouse (GENCODE M4) respectively, results from intensive efforts over more than a decade. Meanwhile, a growing part of genomics studies shifted focus onto the identification of functional non-protein-coding regions, such as distant regulatory enhancer elements or long-noncoding RNA (lncRNAs) genes. Quickly, taking advantage of significant technological innovations, this novel field of investigation led to a major amendment of the human genome annotation, currently rich of 15,900 human lncRNAs (GENCODE V22).

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_5,

1.1 Technology-Driven Research

Precursor technologies such as cDNA 5' and 3'-end sequencing [2], or tiling-arrays [3] have been supplanted with RNA sequencing (RNA-Seq) [4] and tagging technologies combined with high-throughput sequencing [5, 6]. A few milestone studies are briefly reviewed hereafter, illustrating methodological approaches used for cataloging mammalian noncoding genes.

High-resolution tiling-arrays (covering 1.5Gb of the human genome) were used in a pioneer study identifying 10,595 transcriptionally active regions (TARs) mainly intergenic, producing polyadenylated transcripts expressed in the liver [3]. Most of these TARs are conserved among mammalian species and defined as likely to encode proteins and producing ncRNAs with alternate regulatory or structural functions. Cross-hybridizations and incomplete genome coverage are among the main limitations of such hybridization-based methods, which are by design inappropriate for comprehensive analyses of transcriptomes. In 2005, isolating full-length cDNAs and sequencing 5' and 3'-ends from four tissues known for rich transcript diversity (embryo, brain, testis, and thymus), Carninci et al. [7] suggested the presence of as many non-protein-coding as protein-coding genes within the mouse genome. However, this full-length cDNA cloning approach is very labor intensive and necessitates large infrastructure. Others have mapped genome-wide epigenetic histone marks, specific for promoter and gene body regions (*i.e.* H3K4me3 and H3K36me3), using chromatin immuno-precipitation followed by high-throughput sequencing (ChIP-Seq), to identify over 1,600 noncoding transcripts [8]. Taken alone, such epigenetic information carries limited functional prediction values, but when combined with transcriptomics data, it allows accurate identification and classification of ncRNAs. The ENCODE (Encyclopedia Of DNA Elements) consortium designed such a strategy to construct a genome-wide catalogue of human transcripts [9]. Their extensive collection of transcripts, annotated and continuously curated by the GENCODE annotation group [1], currently includes 60,483 genes with 33 % coding for proteins, 26 % for lncRNAs, 16 % for short ncRNAs, and the remaining 25 % being annotated as pseudogenes (GENCODE V22, October 2014). Production of such extensive epigenetic, chromatin accessibility and transcriptomics datasets from multiple cell lines imply large collaborative consortium, gathering the necessary resources and technical skills.

1.2 Detection of Novel Functional Transcripts

Exploration of the mammalian transcriptional landscape complexity remains incomplete and lack of annotated functional features for lncRNAs functional assignments constitutes a great challenge, which necessitates approaches combining several complementary technologies (Fig. 1a). Among others, subcellular transcript localization, dynamic expression patterns, and implication in regulatory network analyses currently provide only circumstantial

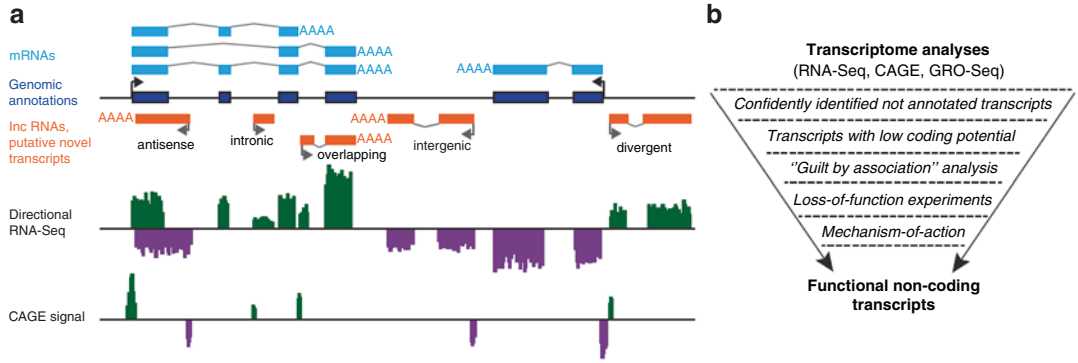


Fig. 1 Identification of functional noncoding RNAs in a complex transcriptional landscape. (**a**) Schematic representation of a genomic context for noncoding transcripts (*orange*) that are generally categorized based on their position and orientation to protein-coding genes (*blue*). Typical directional RNA-seq and CAGE signals are shown for the plus (*green*) and minus (*purple*) genomic strands. (**b**) Multiple bioinformatics and experimental steps required for assigning biological function to newly identified transcripts

information for hypothesis generation. In-depth in silico analyses actually constitute "only" the first few steps in the full description of functional ncRNAs that requires complex experimental approaches for demonstrating function (Fig. 1b).

Recently described classes of ncRNA include enhancer RNAs (eRNAs) [10], enhancer-associated lncRNAs (elncRNAs) [11], multiexonic mRNA isoforms with alternative first exons (meRNA) [12], the very long intergenic ncRNAs (vlincRNA) [13], promoter upstream transcripts (PROMPTs) [14], as well as promoter-associated ncRNAs [15] and lncRNA associated with retrotransposon elements [16–18]. Carefully chosen complementary technologies are common features of strategies used in these novel ncRNA classes' discoveries. Detecting transcriptional start sites (CAGE [5], GRO-Seq [19]) and retrieving transcript structures (RNA-Seq [4]), with high-throughput analyses of chromatin immuno-precipitation (ChIP-Seq) for key factors of the transcriptional machinery or histone modifications appear highly efficient for deciphering complex mammalian transcriptomes.

In the following subheadings, the use of the CAGE technology for the investigation of transcriptome complexity will be detailed, from basic CAGE-data processing to advanced bioinformatics analyses for the identification and characterization of novel transcripts.

1.3 CAGE Technologies

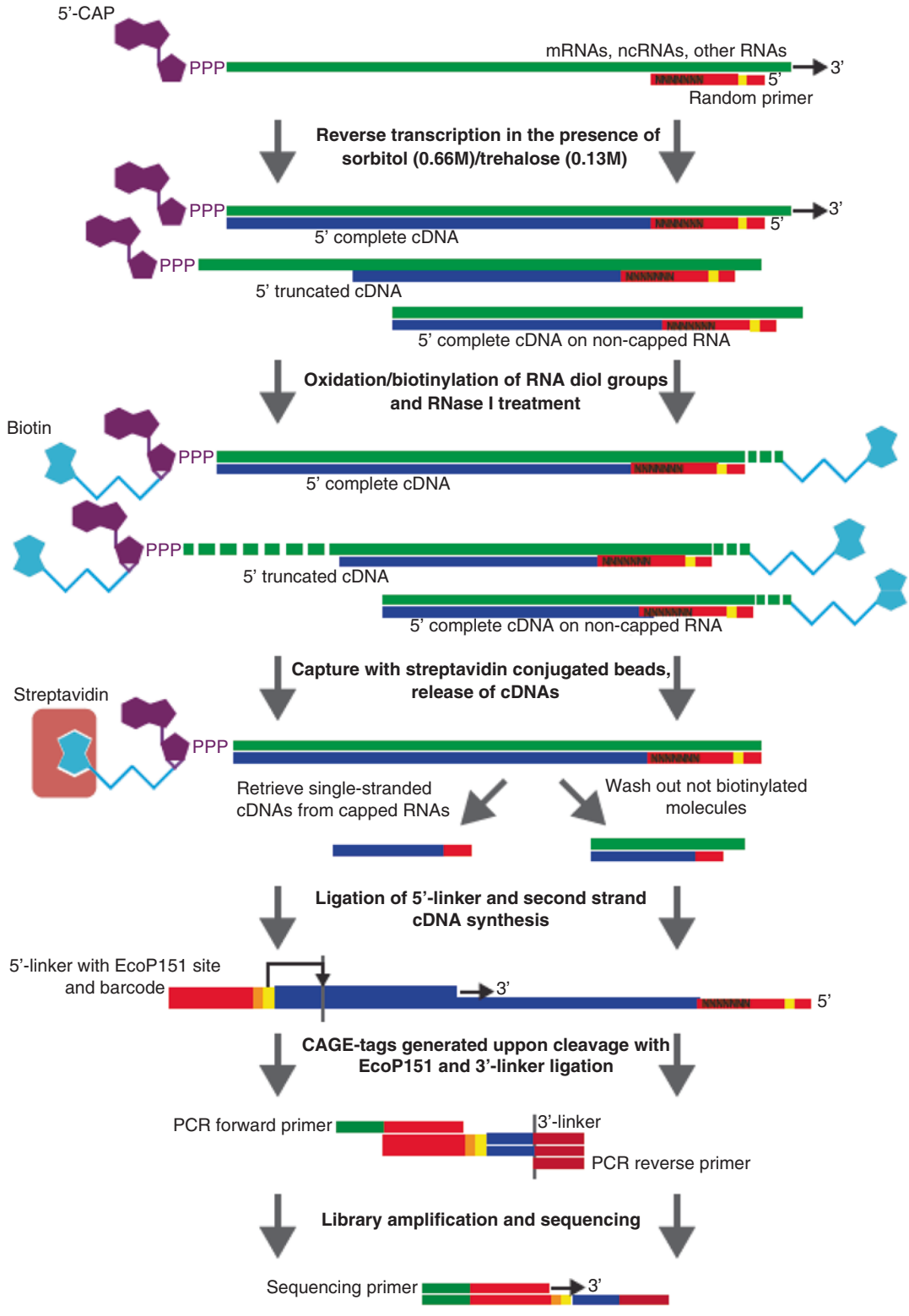
Focusing on the capture of capped 5'-ends of RNAs, the versatile CAGE technology allows unsupervised genome-wide mapping of TSSs and therefore associated core promoter as well as it provides tag frequencies based measure of transcripts expression [5]. It thus enables identification of promoter regions associated with documented and putative novel genes [20], as well as to study

alternative usage of TSSs [21, 22]. The combination of CAGE-defined promoters together with transcription factors binding sites analyses also allows deciphering complex transcriptional regulatory networks and identifying major contributions of specific transcription factors, as published for several cell differentiation processes [23, 24]. Furthermore, while combined with histone marks profiling and detection of balanced bidirectional transcriptional patterns characteristic of enhancer *cis*-regulatory regions, CAGE analysis is informative for the identification of enhancers and analysis of their activity in specific cell types [25].

Since the description of the original CAGE methods [26], six variant protocols have been developed, including (1) optimization for low amount of starting material (nano-CAGE) [27], (2) sequencing paired-end to connect novel promoters to known annotations by skipping the enzymatic restriction (CAGEscan) [27], (3) avoiding library amplification and sequencing using single molecule sequencer (HeliScope-CAGE) [28], (4) combining "*CAP trapper*" [29] and template switching for "extremely" specific 5′-end selection (RAMPAGE) [30], as well as (5) no amplification or tagging steps (nAnTi-CAGE) [31]. Here, the (6) tagging-CAGE library preparation protocol, including enzymatic restriction, library amplification, and analysis using next-generation sequencers, will be briefly explained (Fig. 2, detailed protocol in [32]).

CAGE library construction begins with reverse transcription using random-primer, allowing the 5′-ends of capped RNAs to be effectively and equally reached regardless of their size or polyadenylation status. The reaction takes place in the presence of sorbitol and trehalose increasing the reverse transcriptase activity at high temperatures and allowing efficient cDNA extension for highly structured RNA templates [33]. Then, the specific enrichment of capped RNAs is enabled with the "*CAP trapper*" approach [29], including oxidation of the 5′-cap and 3′-end RNA diol groups, resulting in dialdehyde ends that are then coupled with long-arm biotin hydrazide (Fig. 2). Before capturing biotinylated RNA/cDNA hybrids on streptavidin-coated magnetic beads, samples are submitted to RNase-I treatment, cleaving single-stranded RNA and allowing to discard cDNA not reaching the 5′-cap or bound by their 3′-ends. Finally, RNA/cDNA hybrids are denatured with alkali allowing the recovery of cap-selected single-stranded cDNAs.

Fig. 2 Key steps in the tagging CAGE library preparation. cDNA (*blue*) is synthesized from polyadenylated and non-polyadenylated RNA (*green*) templates with reverse transcription using random primers. Cap and 3'-ends are biotinylated and following the digestion of single stranded RNA (*green* and *dashed lines*) by RNAse-I, capped-RNA/cDNA hybrids are captured on streptavidin-conjugated magnetic beads. cDNAs are then released and ligated to 5′-linkers, including barcode (*orange*) and EcoP15I recognition site (*yellow*). Second strand cDNA synthesis followed by EcoP15I restriction reactions is then conducted, generating CAGE-tags. Ligation of the 3′-linker allows amplification of CAGE-tags before next-generation sequencing


5'-CAP
PPP
mRNAs, ncRNAs, other RNAs
3'
5'
Random primer
Reverse transcription in the presence of sorbitol (0.66M)/trehalose (0.13M)
PPP
3'
5'
5' complete cDNA
PPP
5' truncated cDNA
5' complete cDNA on non-capped RNA
Oxidation/biotinylation of RNA diol groups and RNase I treatment
Biotin
PPP
5' complete cDNA
PPP
5' truncated cDNA
5' complete cDNA on non-capped RNA
Capture with streptavidin conjugated beads, release of cDNAs
Streptavidin
PPP
Retrieve single-stranded cDNAs from capped RNAs
Wash out not biotinylated molecules
Ligation of 5'-linker and second strand cDNA synthesis
5'-linker with EcoP151 site and barcode
3'
5'
CAGE-tags generated uppon cleavage with EcoP151 and 3'-linker ligation
PCR forward primer
3'-linker
PCR reverse primer
Library amplification and sequencing
Sequencing primer
3'

Sample multiplexing is achieved by the introduction of barcode sequences in the 5′-linkers ligated to the 3′-extremity of first-strand cDNA. The 5′-linkers are also used for priming the second-strand cDNA synthesis and for CAGE-tag generation using the EcoP15I restriction enzyme (Fig. 2). Following the 3′-linker ligation, final CAGE-library amplification is performed before to be loaded on high-throughput sequencing platform.

2 Materials

1. *TagDust2* [34] (*http://sourceforge.net/projects/tagdust*).
2. *Delve* [9] (*http://sourceforge.net/projects/tometools*).
3. *BWA* [35] (*http://bio-bwa.sourceforge.net*).
4. *Bowtie* [36] (*http://bowtie-bio.sourceforge.net*).
5. *Paraclu* [37] (*http://www.cbrc.jp/paraclu*).
6. *RECLU* [38] (*http://en.sourceforge.jp/projects/reclu*).
7. *ChromHMM* [39] (*http://compbio.mit.edu/ChromHMM*).
8. *EPO* [40] (*http://www.ensembl.org/info/genome/compara/epo_anchors_info.html*).
9. *liftOver* [41] (*http://genome.ucsc.edu/cgi-bin/hgLiftOver*).
10. *PhyloCSF* [42] (*http://github.com/mlin/PhyloCSF/wiki*).
11. CPAT [43](*http://rna-cpat.sourceforge.net*).
12. *WGCNA* [44] (*http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA*).
13. *bnlearn* [45] (*http://cran.r-project.org/web/packages/bnlearn/index.html*).
14. *DAVID* [46] (*http://david.abcc.ncifcrf.gov*).

3 Methods

3.1 CAGE-Data Processing

CAGE-data processing consists of extracting CAGE-tags from raw sequencing files, mapping these back to the genome, identifying active transcriptional start sites (TSSs), and counting the number of reads for each site for quantification of downstream transcripts (Fig. 3). In the following subheading, each step of CAGE-data processing will be detailed.

3.1.1 CAGE-Tag Extraction and Sample De-Multiplexing

Bioinformatics tools have been developed to extract and sort raw sequences from raw sequencing files produced by next-generation sequencing platforms (typically in FASTQ format). *TagDust2* [34] takes as input FASTQ files and the expected composition (starting with barcode, EcoP15I recognition site, CAGE tag, and 3′-linker

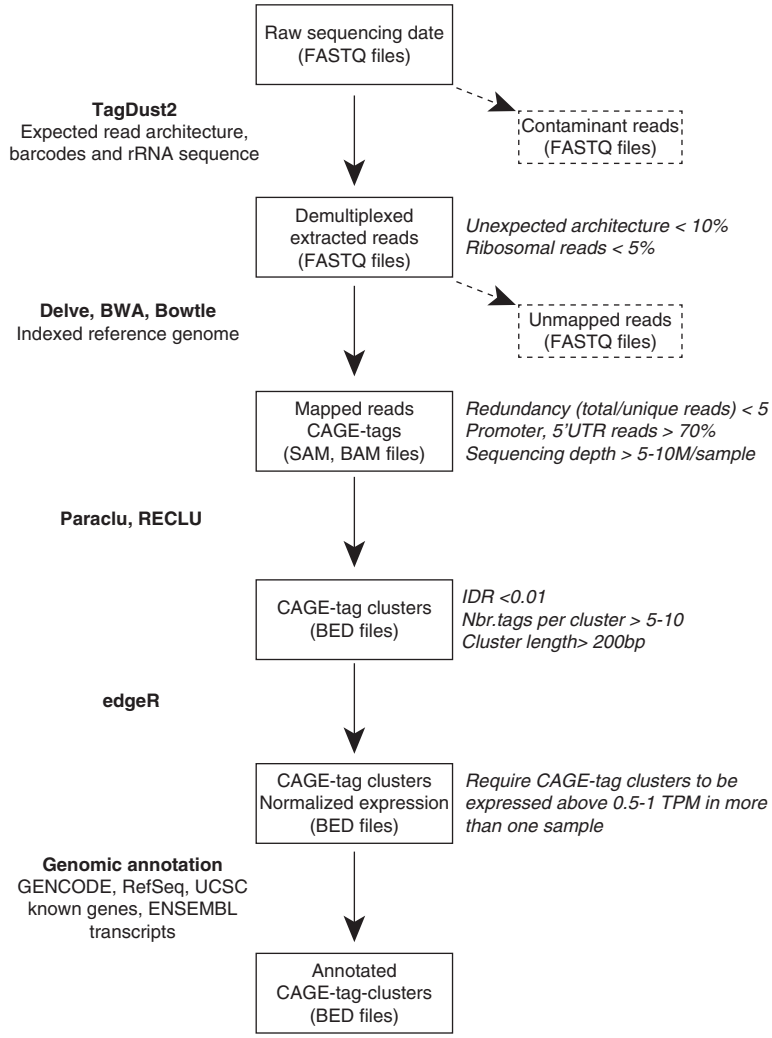


Fig. 3 Schematic view of the basic processing of CAGE data. Bioinformatics tools and approaches are shown in *bold*. Quality control expectation and usual threshold are depicted in italic

sequence; Fig. 2) of sequencing reads to be extracted. Building hidden Markov models, the algorithm assigns each read to a particular barcode that corresponds to an individual sample. It also flags unwanted "contaminants" such as sequences not matching expected architectures (*i.e.* linkers, primer dimers, or other contaminants) or mapping to reference ribosomal RNA sequences.

De-multiplexed reads, trimmed for 5′ and 3′-linker sequences, are then mapped to reference genomes using alignment algorithms optimized for short reads (<50 bp). We recommend using *Delve* [9], which in combination with *BWA* [35], provides accurate mapping to short read, modeling alignments (using hidden Markov model) from the data itself. This modeling approach allows

discrimination between genuine and spurious alignments, taking into consideration detected biases toward certain substitutions. Alternatively, *Bowtie* [36] can also be used for mapping CAGE-tags. Importantly, multi-mapping reads, assigned to multiple locations of the genome, are filtered out at this step, based on BWA MapQ values (*see* **Note 1**).

3.1.2 Sequencing Depth and Quality Controls

There is no definitive consensus for the required sequencing depth of CAGE libraries; however, 5–10 million mapped reads are generally used when comparing multiple samples [32]. For deeper exploration of transcriptomes, including rare transcripts at less or close to one copy per cell, one would opt for 30–50 million mapped reads. To improve detection of rare transcripts, one could start with RNA extracted from subcellular compartments as nucleus, nucleoplasm, chromatin or cytoplasm.

Evaluating CAGE-library quality before proceeding with further analyses is imperative for reasonable interpretation of results. Contaminating reads of unexpected architectures or those that map to ribosomal regions should ideally count for less than 10 % and 5 % of total reads respectively. The read redundancy, how many times a read is observed on average in a library, controls for molecular bottlenecks introduced by PCR amplification. Mammalian promoters generally depict multiple TSSs within core promoters [21], relatively low and spread CAGE-tag distributions are thus expected at promoter regions. When studying complex mammalian transcriptomes, redundancy values (total/unique reads) above 5 are suggestive of a molecular bottleneck introduced in the amplification step [32]. Finally, counting hits of CAGE-tags to annotated promoters controls for efficient enrichment of capped transcripts. Starting with total RNA, around 70 % of reads mapping to 5′-UTRs and core promoters are expected.

3.1.3 CAGE-Tag Clustering and Normalization of Expression Values

The next processing step consists of identifying TSSs based on clustering CAGE-tags (*see* **Note 2**). Notably, CAGE technology maintains strand information with clustering on each chromosomal strand performed independently. *Paraclu* [37] is a parametric clustering method developed for TSS-associated tags. It reports genomic intervals containing more tags than surrounding areas. The method tolerates CAGE-tag clusters that are contained within each other, therefore resulting in a hierarchy of CAGE-tag clusters. However, clusters contained within larger ones are commonly excluded, restricting downstream analyses to a single hierarchical level. In order to gain insights into transcription initiation events from CAGE-data, the *RECLU* [38] clustering analysis pipeline was developed. Starting with *Paraclu* defined CAGE-tag clusters, *RECLU* aims to detect reproducible TSSs from the top and bottom of a clusters hierarchy. It detects alternative TSSs and their differential usage among samples, selecting reproducible signals based

on "irreproducible discovery rate" analyses (IDR) [47]. Then, expression values are calculated, counting CAGE-tags per defined cluster and normalizing counts as TPM (tags per million) based on the library size (*i.e.* sequencing depth). Alternatively, as computed for the FANTOM5 consortium atlas of mammalian promoters [20], expression values can be calculated using the relative logarithmic expression method (RLE) [48], considering the normalization factor estimated by *edgeR* [49].

3.1.4 CAGE-Tag Cluster Filtering

Applying several filters to the detected CAGE-tag clusters carries out construction of permissive or robust TSS sets. To this end, expression level is commonly used as a first filtering, discarding lowly expressed clusters based on a minimum expression cutoff. A minimum number of CAGE-tags are therefore required (typically 5-10 tags) when clustering with *Paraclu* [37]. On the other hand, the *RECLU* method [38] aims at retaining lowly expressed CAGE-tag clusters if reproducibly found in multiple samples. Therefore, while using *RECLU*, instead of considering a total tag count cut-off, a normalized expression threshold will be subsequently assigned, typically omitting clusters with expression levels below 0.1–0.5 TPM. Relevance for low expression cutoff is supported by the observation that only a minority of annotated lncRNAs are stable and abundant molecules. It has been shown that only a minority of noncoding transcripts are resistant to ribo-nucleolytic RNA exosome complex-mediated decay and thus potentially carrying their role as high-copy molecules in *trans*, similarly to messenger RNA (mRNAs) [50].

Filtering CAGE-tag clusters based on statistical support of their reproducibility is an effective way to filter out spurious TSSs. For example, Derrien et al. [51] discarded clusters with IDR below 0.1. When using other clustering approaches, one could consider a minimal requirement for all CAGE-tag clusters to be expressed above a certain threshold (typically 0.5–1 TPM) in more than one sample.

Finally, independent evidences from other transcriptomics technologies (*i.e.* RNA-Seq, GRO-Seq or EST data) or characteristic chromatin features (DNase hypersensitive sites, presence of H3K4me3, H3K4me1 or H3K27ac histone marks) are very much informative for selecting genuine TSSs.

3.2 Genomic and Putative Functional Annotation

Assigning each CAGE-tag cluster selected after filtering to either annotated or yet not defined genes is the initial step in the course of identifying novel transcripts. Multiple annotation sources are available and should be considered together. These include GENCODE [1], ENSEMBL transcripts [40], RefSeq [52], and UCSC known gene [53]. We recommend treating multiple genomic annotations with the following hierarchy starting with (1) sense TSSs and exons, (2) antisense TSSs and exons, (3) sense and

3.2.1 Genomic Annotation for Known Transcripts

antisense introns, (4) sequence ±1 kb relative to TSSs and (5) intergenic sequences. Generally, CAGE-tag clusters are considered to be associated with an annotated transcript when residing within 500 bp of its 5′-end [17, 20].

3.2.2 Classification Based on Epigenetic Features and Transcript Stability

Genome-wide chromatin profiling, based on epigenetic histone marks, emerged as an efficient approach to annotate functional regulatory elements, such as promoters, enhancers or insulators [54]. The chromatin state discovery and characterization method (*ChromHMM*) [39] has been developed to define functional chromatin segments based on presence of distinctive histone marks. Overlapping such *ChromHMM* functional segments with nonannotated CAGE-tag clusters allows classification of potential novel transcripts as originated from *bona fide* promoters, being associated with enhancer regions or as not supported by epigenetic marks characteristic of active transcription.

3.2.3 Evolutionary Conservation Features

An ancient precept, founded on natural selection theory, proposes sequence evolutionary conservation of genomic elements as suggestive of biological function (reviewed in [55]). Although highly conserved lncRNAs have been reported [56, 57], modest sequence conservation is generally observed among them [8, 57, 58, 59]. Lack of conservation should thus not be used as an excluding criterion for functional ncRNA candidates. However, when evolutionary conservation is detected, it pleads in favor of biological function. Therefore, if CAGE data for multiple species in homologous cell types are available, one should search for the presence of conserved CAGE-tag clusters. To this end, TSSs identified in a species can be projected onto another species genomes using the Ensemble *EPO12* eutherian mammal multiple sequence alignments [60]. *LiftOver* [41], available on the UCSC Genome browser, would also be appropriate for this type of analysis. CAGE-tag cluster boundaries are usually used to define intervals to be analyzed in multiple species alignments.

3.2.4 Assessing Coding Potential

An obvious limitation of the CAGE technology concerns the lack of information related to transcript structures. However, evaluating the coding potential of putative novel regulatory ncRNAs is an initial requirement for planning further functional loss/gain of function experiments. To this end, full-length sequences with accurate 5′ and 3′-ends as well as exon boundaries are required. Comprehensive description of exon structures and splicing variants can be obtained from RNA-Seq analysis [4]; however, coverage of 5′-ends is often too low for precise TSS definition. Accurate detection of transcript extremities can be achieved with complementary technologies using either rapid amplification of 5′ cDNA ends (5′-RACE) designed for a set of candidate RNAs, or genome-wide CAGE for TSSs mapping and 3′ Seq [61] or SAGE [62] for 3′-ends.

Bioinformatics tools have been developed to distinguish between protein-coding and noncoding sequences identified from deep transcriptome sequencings. Among the different available methods *PhyloCSF* (Phylogenetic Codon Substitution Frequency) [42], use statistical models from phylogenetic codon models to detect conserved protein-coding regions. To this end, it evaluates (1) the presence of evolutionary signatures of conserved coding regions, (2) the frequency of synonymous codon substitutions, and (3) the frequency of missense and nonsense substitutions. Alternatively, CPAT (Coding Potential Assessment Tool) [43] method uses (1) the size of open reading frame, (2) the open reading frame coverage, (3) high statistic for compositional bias with a periodicity of three [63], and (4) bias in hexamer usage. Not relying on multispecies alignments, it efficiently and quickly evaluates coding potential of not conserved lncRNAs as well as for lncRNAs overlapping or antisense to protein coding genes.

3.2.5 *"Guilt-by-Association" Derived Hypothetic Functions*

Functional annotation of lncRNAs remains a challenging task since we lack comparable features to make assignments in an analogous way to the homology of functional protein domains. Epigenetic and conservation features support the identification of genuine novel transcripts but additional information is needed for hypothesis generation regarding function. Building regulatory networks including putative novel transcripts and protein-coding genes, followed by gene ontology analysis, the "*guilt-by-association*" (Fig. 4) approach has been proven to be effective for constructing functional hypotheses for novel ncRNAs [59, 64].

Starting with transcript profiles, constructed with CAGE or RNA-Seq data, interrelation between novel transcripts and protein-coding genes is built using correlation matrices, weighted gene co-expression network analysis (*WGCNA*) [44], or a Bayesian network method (*bnlearn*) [45]. Then gene ontology term

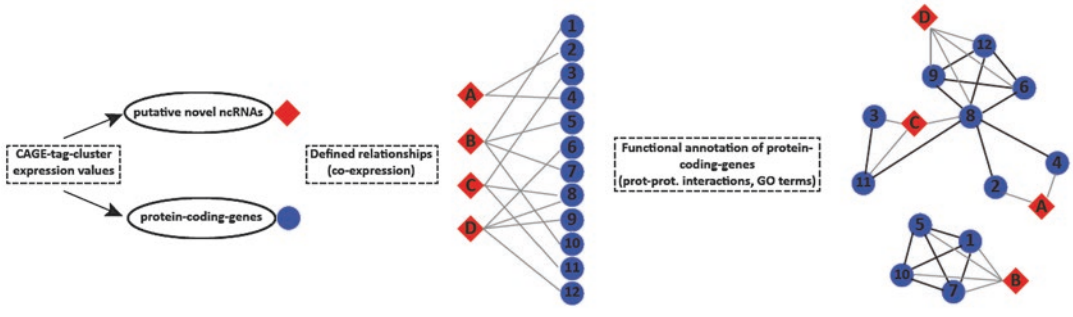


Fig. 4 Schematic workflow of "guilty-by-association" analysis. Normalized expression values for putative novel nc-RNAs and protein-coding genes are retrieved from a single dataset. Relationships between noncoding and protein-coding RNAs are computed based on co-expression analyses. Each ncRNA is mapped to a protein-protein interactions network and predicted functions are derived from functional enrichment of connected genes

enrichments for gene lists extracted from the network analysis is conducted. The *DAVID* Knowledgebase (Database for Annotation, Visualization, and Integrated Discovery) [46] provides tools to perform automated functional annotation including statistical testing.

Finally, tissue-specific or dynamic expression patterns through multiple developmental stages or cell conditions are used to assign putative function to novel transcripts. Currently, two main databases are available: the Genotype-Tissue Expression project (GTEx) [65] and the FANTOM5 promoter atlas [20]. Importantly, inherent methodological limitations, such as restriction on the polyadenylated fraction or lack of genomic strand information, for each database have to be considered when estimating expression levels of rare nuclear lncRNAs.

3.3 Functional Validations

We would like to conclude with some considerations for testing experimentally the functional roles of confidently detected novel transcripts (Fig. 1b) and refer the reader to more detailed reviews published elsewhere [66, 67].

Functional ncRNAs have diverse mechanism-of-action, generally classified as *cis* when regulating the expression of genomically neighboring genes, or *trans* when targeting distant genes. Among other roles, regulatory ncRNAs are implicated in the modulation of protein activities, recruiting effective factors at specific genomic loci or acting as primary host transcripts for small ncRNAs. The great variability of mechanisms implies the need for a good understanding of interacting factors. Specific technologies to detect interacting proteins (crosslinking-immunoprecipitation) [68], RNAs (crosslinking analysis of synthetic hybrids) [69], or DNA loci (capture hybridization analysis of RNA targets, chromatin isolation of RNA purification) [70, 71] are critical to select specific interaction candidates. In-depth functional investigations will be typically carried on cellular or whole organism models with genetic modification, over-expression, or knockdown strategies. Noteworthy, the act of transcription itself could be critical for modifying chromatin structure and to guarantee access to chromatin; expressed RNA molecules are in this case purely incidental by-products. Therefore, multiple and complementary strategies are needed to discriminate the role of the genomic locus from the role of the RNA.

Loss-of-function followed by complementation experiments provides prime evidence for RNA-dependent ncRNA functions. The design of such perturbation experiments needs to consider subcellular localization and *cis*- or *trans*-predicted mechanisms, preferring RNAi knockdown for cytoplasmic transcripts with *trans*-activity, while nuclear *cis*-active ncRNAs can be efficiently targeted with antisense DNA or LNA (lock nucleic acids) oligonucleotides inducing RNase H mediated degradation [72]. Targeted genetic modification, using the CRISPR-Cas9 or TALEN technologies

(reviewed in [73]), is certainly the current method of choice for chromatin-bound ncRNAs acting in *cis*. However, one will need to carefully design such manipulation, minimizing interference with other transcripts or regulatory regions when editing noncoding loci overlapping or in close proximity of protein-coding genes (Fig. 1a).

In addition to loss-of-function strategies, cis overexpression of specific noncoding endogenous loci has also been shown to be efficient for deciphering *cis*-acting ncRNA transcript function [74, 75].

4 Notes

1. Commonly CAGE analyses are carried out with sets of sequencing reads mapping at unique position in the reference genome. However, transcription occurring at repeated regions, such as retrotransposons, can be estimated mapping CAGE-tag sequences to all repeat elements as defined by RepeatMasker [76] and normalizing for sequencing depth. This will allow comparisons of repeat elements or family between different samples such as conducted in [17].
2. A few integrated bioinformatics tools have been developed specifically for the processing and analysis of CAGE data and would be very useful for users of CAGE with limited bioinformatics skills.

ZENBU [77] (*http://fantom.gsc.riken.jp/zenbu*) is a web-based system to view and process CAGE-data with predefined scripts. Starting by loading the binary version of sequence alignment outputs (BAM files), *ZENBU* supports quality filtering, signal normalization, annotation, as well as expression difference visualization across multiple samples.

CAGEr [78] (*http://www.bioconductor.org/packages/CAGEr*) is a comprehensive toolbox for CAGE-datasets developed for R. Its workflow detects CAGE-tag clusters from BAM files, analyses promoter width and differential TSS usage, and calculates normalized expression values. It also provides easy access to publicly available CAGE data from human [20], mouse [9], zebrafish [79], and the fruit fly [80].

CAGEexploreR [81], (*https://github.com/edimont/CAGExploreR*) is a R package specifically developed to investigate promoter dynamics and TSS usage across multiple samples [21].

Acknowledgments

We thank Charles Plessy for critical reading and suggestions. This work was supported by a Swiss National Science Foundation Ambizione grant (PZ00P3–154728) grant to A.F.

References

1. Harrow J, Frankish A, Gonzalez JM (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22(9):1760–1774
2. Ng P, Wei CL, Sung WK, Chiu KP (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2(2):105–111
3. Bertone P, Stolc V, Royce TE (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306(5705): 2242–2246
4. Cloonan N, Forrest AR, Kolle G (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5(7): 613–619
5. Kodzius R, Kojima M, Nishiyori H (2006) CAGE: cap analysis of gene expression. Nat Methods 3(3):211–222
6. Hoen PA, Ariyurek Y, HH T (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res 36(21):e141
7. Carninci P, Kasukawa T, Katayama S (2005) The transcriptional landscape of the mammalian genome. Science 309(5740):1559–1563
8. Guttman M, Amit I, Garber M (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458(7235):223–227
9. Djebali S, Davis CA, Merkel A (2012) Landscape of transcription in human cells. Nature 489(7414):101–108
10. Kim TK, Hemberg M, Gray JM (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465(7295): 182–187
11. Marques AC, Hughes J, Graham B (2013) Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol 14(11):R131
12. Kowalczyk MS, Hughes JR, Garrick D (2012) Intragenic enhancers act as alternative promoters. MollCell 45(4):447–458
13. St Laurent G, Shtokalo D, Dong B (2013) VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol 14(7):R73
14. Preker P, Nielsen J, Kammler S (2008) RNA exosome depletion reveals transcription upstream of active human promoters. Science 322(5909):1851–1854
15. Hung T, Wang Y, Lin MF (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet 43(7):621–629
16. Kelley D, Rinn J (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 13(11):R107
17. Fort A, Hashimoto K, Yamada D (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet 46(6):558–566
18. Lu X, Sachs F, Ramsay L (2014) The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol 21(4):423–425
19. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322(5909):1845–1848
20. The FANTOM Consortium, RIKEN PMI, Forrest AR (2014) A promoter-level mammalian expression atlas. Nature 507(7493):462–470
21. Carninci P, Sandelin A, Lenhard B (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38(6):626–635
22. Haberle V, Li N, Hadzhiev Y (2014) Two independent transcription initiation codes overlap on vertebrate core promoters. Nature 507(7492):381–385
23. The FANTOM Consortium, Suzuki H, Forrest AR (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet 41(5):553–562
24. Mina M, Magi S, Jurman G (2015) Promoter-level expression clustering identifies time development of transcriptional regulatory cascades initiated by ErbB receptors in breast cancer cells. Sci Rep 5:11999
25. Andersson R, Gebhard C, Miguel-Escalada I (2014) An atlas of active enhancers across human cell types and tissues. Nature 507(7493):455–461
26. Shiraki T, Kondo S, Katayama S (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100(26):15776–15781
27. Plessy C, Bertin N, Takahashi H (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat Methods 7(7):528–534
28. Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T (2011) Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res 21(7):1150–1159

29. Carninci P, Kvam C, Kitamura A (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics 37(3):327–336
30. Batut P, Dobin A, Plessy C (2013) High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res 23(1):169–180
31. Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M (2014) Detecting expressed genes using CAGE. Methods Mol Biol 1164:67–85
32. Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7(3):542–561. doi:10.1038/nprot.2012.005nprot.2012.005[pii]
33. Carninci P, Nishiyama Y, Westover A (1998) Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. Proc Natl Acad Sci USA 95(2):520–524
34. Lassmann T (2015) TagDust2: a generic method to extract reads from sequencing data. BMC Bioinform 16(1):24
35. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760
36. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. doi:10.1186/gb-2009-10-3-r25
37. Frith MC, Valen E, Krogh A (2008) A code for transcription initiation in mammalian genomes. Genome Res 18(1):1–12
38. Ohmiya H, Vitezic M, Frith MC (2014) RECLU: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). BMC Genom 15:269
39. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9(3):215–216. doi:10.1038/nmeth.1906
40. Flicek P, Amode MR, Barrell D (2014) Ensembl 2014. Nucleic Acids Res 42(Database issue):D749–D755
41. Hinrichs AS, Karolchik D, Baertsch R (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34(Database issue):D590–D598
42. Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27(13):i275–i282. doi:10.1093/bioinformatics/btr209
43. Wang L, Park HJ, Dasari S (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41(6):e74
44. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 9:559. doi:10.1186/1471-2105-9-559
45. Scutari M (2010) Learning Bayesian networks with the bnlearn R package. J Stat Softw 35(3):1–22
46. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57
47. Li QH, Brown JB, Huang HY, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. Ann Appl Stat 5(3):1752–1779
48. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106
49. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140
50. Andersson R, Refsing Andersen P, Valen E (2014) Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun 5:5336
51. Derrien T, Johnson R, Bussotti G (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22(9):1775–1789
52. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33(Database issue):D501–D504
53. Hsu F, Kent WJ, Clawson H (2006) The UCSC known genes. Bioinformatics 22(9):1036–1046
54. Ernst J, Kheradpour P, Mikkelsen TS (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473(7345):43–49
55. Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences—an unexpected feature of mammalian genomes. Nature Rev Genet 6(2):151–157
56. Chodroff RA, Goodstadt L, Sirey TM (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol 11(7):R72

57. Necsulea A, Soumillon M, Warnefors M (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505(7485):635–640
58. Marques AC, Ponting CP (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol 10(11):R124
59. Cabili MN, Trapnell C, Goff L (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25(18):1915–1927
60. Hubbard TJ, Aken BL, Ayling S (2009) Ensembl 2009. Nucleic Acids Res 37(Database issue):D690–D697
61. Beck AH, Weng Z, Witten DM (2010) 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. PLoS One 5(1):e8768
62. Nielsen KL, Hogh AL, Emmersen J (2006) DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res 34(19):e133
63. Fickett JW (1982) Recognition of protein coding regions in DNA sequences. Nucleic Acids Res 10(17):5303–5318
64. Pauli A, Valen E, Lin MF (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22(3):577–591
65. The GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nature Genet 45(6):580–585
66. Kapusta A, Feschotte C (2014) Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. TrendsGenet 30(10):439–452
67. Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC, Gingeras TR, Haerty W, Higgs DR, Miska EA, Ponting CP (2014) Considerations when investigating lncRNA function in vivo. eLife 3:e03058
68. Huppertz I, Attig J, D'Ambrogio A (2014) iCLIP: protein-RNA interactions at nucleotide resolution. Methods 65(3):274–287
69. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 153(3):654–665
70. Simon MD, Wang CI, Kharchenko PV (2011) The genomic binding sites of a noncoding RNA. Proc Natl Acad Sci U S A 108(51): 20497–20502
71. Chu C, Qu K, Zhong FL (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell 44(4):667–678
72. Sahu NK, Shilakari G, Nayak A, Kohli DV (2007) Antisense technology: a selective tool for gene expression regulation and gene targeting. Curr Pharm Biotechnol 8(5):291–304
73. Gaj T, Gersbach CA, Barbas CF 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31(7):397–405
74. Cheng AW, Wang H, Yang H (2013) Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res 23(10):1163–1171
75. Xiang JF, Yin QF, Chen T (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res 24(5):513–531
76. Jurka J, Kapitonov VV, Pavlicek A (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1–4):462–467
77. Severin J, Lizio M, Harshbarger J (2014) Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat Biotechnol 32(3):217–219
78. Haberle V, Forrest AR, Hayashizaki Y, Carninci P, Lenhard B (2015) CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res 43(8):e51
79. Nepal C, Hadzhiev Y, Previti C (2013) Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res 23(11):1938–1950
80. Hoskins RA, Landolin JM, Brown JB (2011) Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21(2):182–192
81. Dimont E, Hofmann O, Ho Sui SJ (2014) CAGExploreR: an R package for the analysis and visualization of promoter dynamics across multiple experiments. Bioinformatics 30(8): 1183–1184

Part II

Characterization of Promoter-Associated RNA Features

Chapter 6

Deep-RACE: Comprehensive Search for Novel ncRNAs Associated to a Specific Locus

Chiara Pastori, Dmitry Velmeshev, and Veronica Julia Peschansky

Abstract

Deep-RACE (or RACEseq) is a recently described method (Olivarius et al. BioTechniques 46(2):130–132, 2009) that applies next-generation sequencing to the Rapid Amplification of cDNA End (RACE) protocol to define the 5′ and 3′ ends of RNA transcripts. Conventional mapping of 5′ and 3′ ends is achieved by manually cloning the PCR product of RACE followed by Sanger sequencing; this process can become costly and time-consuming when investigating multiple transcripts. High-throughput sequencing of the RACE products streamlines this process by eliminating the need to manually cut bands from an agarose gel and to clone each product individually. Importantly, in addition to these advantages, next-generation sequencing can detect low abundance fragments that would be difficult to extract from gel and clone for Sanger sequencing. For these reasons, Deep-RACE is an ideal protocol for the comprehensive study of noncoding transcripts from both intergenic regions of the genome and from within the loci of protein coding genes.

Key words Long noncoding RNAs, Deep-RACE, 5′ and 3′ Rapid amplification of cDNA ends, Transcriptome, Next-generation sequencing

1 Introduction

The majority of the human genome is transcribed but not translated, giving rise to a plethora of noncoding RNAs (ncRNAs), including long ncRNAs, that are implicated in a wide variety of gene-regulatory functions. Although next-generation RNA sequencing (RNAseq) can identify thousands of lncRNAs throughout the genome, the precise start site (5′-end) and termination (3′-end) of these transcripts must always be verified by an independent method. Deep-RACE is an excellent tool for this purpose.

RACE kits (5′ and 3′) are available from several companies, including Life Technologies, Clontech, Roche, Life Science, and Ambion. These protocols are optimized for the detection of protein coding genes; therefore, it can be more difficult to map

*Author contributed equally with all other contributors.

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_6,

lncRNAs as they often have lower expression compared with protein coding genes, may not be polyadenylated and due to splicing processes, might be missing the cap at the 5′-end. Protocols of commercially available kits are generally very similar, with the exception of the RML-RACE kit (Ambion) that relies on a capped 5′ end containing a methylated guanosine. Capping of the 5′-end is part of the normal process of transcription mediated by Polymerase II. Although most eukaryotic RNAs are capped, post-transcriptional processing events such as splicing can give rise to mature noncoding RNAs with an uncapped 5′-end. Therefore, the application of RACE kits to the study of lncRNAs might require some modifications of the protocol. We have appended next-generation sequencing to the traditional RACE protocol to define the 5′ and 3′-ends and splicing of sense- and antisense-oriented noncoding transcripts in a specific locus [1, 2]. Other studies have used similar approaches [3–5].

2 Materials

1. cDNA synthesis kit.
2. cDNA purification kit.
3. TdT enzyme.
4. dCTP.
5. DNA Polymerase.
6. Gene specific and anchor primers: (*see* Table 1) (*see* **Note 1**).
7. Polyadenylation enzyme (as needed).
8. Thermocyclers.
9. Electrophoresis apparati.
10. Bioanalyzer.
11. Next-generation DNA sequencing library preparation kit (i.e., NEB, Illumina, Takara, Clontech).

Table 1
Primer sequences

	Primer name	Primer sequence
3′ RACE	AP	GGC CAC GCG TCG ACT AGT ACT TTT TTTTTTTTTTT T
5′ RACE	AAP	GGC CAC GCG TCG ACT AGT ACG GGi* iGGGii GGG iiG
3′/5′ RACE	AUAP	GGC CAC GCG TCG ACT AGT AC

3 Methods

Before starting a RACE experiment, it is necessary first to collect information about the expression of noncoding RNAs in the locus of interest (*see* **Note 2**, Fig. 1).

3.1 Rapid Amplification of 5′-End

Here, we describe the protocol for identifying the 5′-end of transcripts, regardless the presence of a cap.

Below is a schematic of this 5′-RACE protocol (Fig. 2) (*see* **Note 3**).

3.1.1 Primer Design

The RACE protocol consists of reverse transcription and several PCR reactions that require correct primer design for efficient and specific amplification. To this end, there are several web sites and software packages for primer design that take GC content, *T*m, and other factors into consideration.

The primer for first-strand cDNA synthesis (gene-specific primer 1, GSP1) should be designed to have a melting temperature (*T*m) that will facilitate annealing to RNA molecules during the reverse transcription step (*see* **Note 4**). GSP2 and GSP3 need to have *T*m compatible with the adapter primers (*see* **Note 5**).

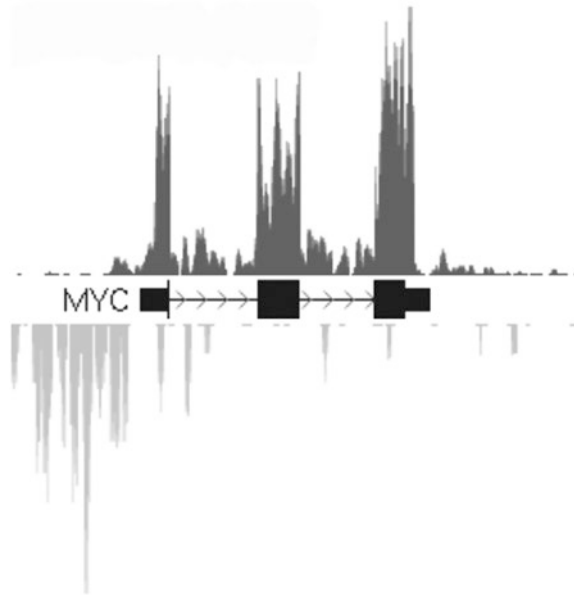


Fig. 1 Strand-specific RNA sequencing shows the presence of RNA transcribed from the plus strand of DNA corresponding to c-MycmRNA. The minus strand corresponds to antisense noncoding transcription along the c-Myc gene locus. Mapping and visualization of noncoding transcription can facilitate the design of RACE primers

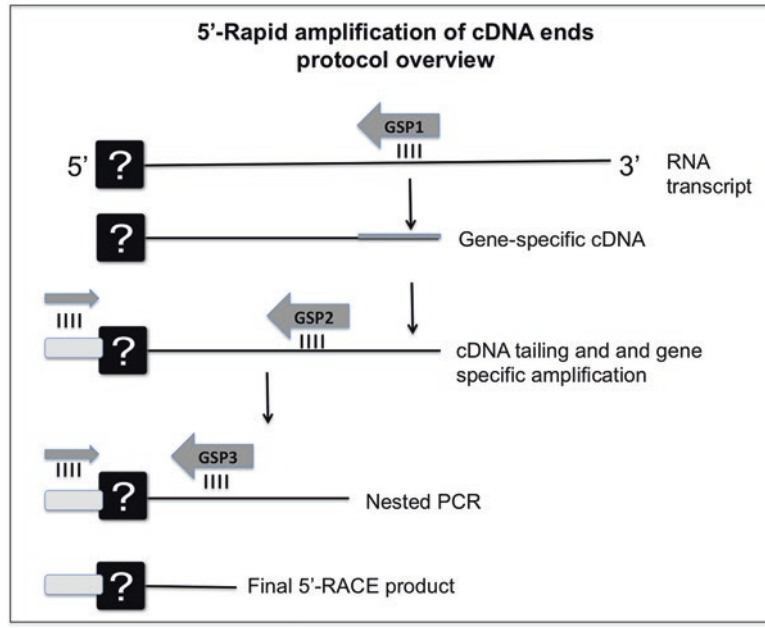


Fig. 2 Schematic of 5′-RACE. Using a gene-specific primer (GSP1) the first-strand cDNA is synthesized from total RNA. A tail is then added to the 3′-end of the cDNA using TdT and dCTP. A first PCR amplification is accomplished using a second gene-specific primer (GSP2) that anneals to a site located within the cDNA molecule, and an anchor primer provided with the kit. Following amplification, 5′ RACE products can be subjected to a second round of PCR (nested PCR) using an anchor primer and a gene-specific primer GSP3. The final products can be purified and used for next-generation sequencing library preparation

For the 5′ RACE, design all GSPs as reverse primers (aka left primer)—meaning that they are the reverse complement of the sequence of interest.

3.1.2 RNA Isolation (See Note 6)

1. Collect ~5 × 10 E6 cells and resuspend the pellet in 1 ml of TriZol and transfer to a 1.5 ml tube.
2. Add 200 μl of chloroform, vortex and centrifuge at maximum speed for 15 min in a refrigerated microcentrifuge.
3. Collect 350 μl of aqueous phase (upper clear layer) and mix it with an equal volume of isopropanol (*see* **Note 7**).
4. Add ammonium acetate to achieve a final concentration of 0.5 M and precipitate the RNA at −20 °C overnight or −80 °C for 2 h.
5. After precipitation centrifuge at 4 °C in a refrigerated microcentrifuge, remove the supernatant and gently wash the pellet (containing RNA) with ethanol (70 % EtOH).

6. Repeat the washing **step 5** and let the pellet dry at room temperature.
7. Resuspend the RNA in RNAse/DNAse free water (*see* **Notes 8** and **9**).

3.1.3 cDNA Synthesis (See Note 10)

This will be performed using gene-specific primer 1 (GSP1), which is a primer designed to be complementary to the transcript of interest (*see* Fig. 2).

1. Prepare the first-strand cDNA synthesis reaction in a microcentrifuge tube by adding the following components:

   Up to 5 μg of total RNA (*see* **Note 11**).

   RNAse-free water to 15.7 μl.

   1 μl of GSP1 (10 μM stock).
2. Incubate at 65 °C for 5 min.
3. Place on ice for 10 min and add the following reagents:

   2 μl of 10×AffinityScript RT Buffer.

   0.8 μl of dNTP mix (100 mM, 25 mM each).

   1 μl of AffinityScript Multiple Temperature RT.
4. Place the tubes in a thermocycler and run the reverse transcription for 60 min at a temperature 42–55 °C (*see* **Note 12**).
5. Terminate the reaction at 70 °C for 15 min and then store at −20 °C or proceed to the cDNA purification (*see* **Note 5**).
6. Purify cDNA, according to manufacturer's instructions (*see* **Note 13**).

3.1.4 Homopolymeric Tailing of cDNA with TdT Enzyme

The terminal deoxynucleotidyl transferase (TdT) is a DNA polymerase that catalyzes the addition of nucleotides to the 3′-end of a single stranded DNA molecule (*see* **Note 14**)

1. Start the reaction by mixing:

   5.0 μl of 10×TdT Buffer.

   5.0 μl of 2.5 mM $CoCl_2$ solution provided.

   20 μl of purified cDNA.

   0.5 μl 10 mM dCTP.

   0.5 μl Terminal Transferase (20 units/μl).

   Bring to 50 μl with water.
2. Incubate at 37 °C for 30 min. Stop the reaction by heating to 70 °C for 10 min.

3.1.5 PCR of Tailed cDNA and Nested PCR (See Note 15)

1. In a PCR tube mix the following:

   0.5 μl (2.5 units) of Taq polymerase.

25 μl of 2× GC rich buffer (*see* **Note 16**).

8 μl of dNTPs.

1–2 μl of tailed cDNA.

2.5 μl of GSP2 (from 10 μM stock).

2.5 μl of anchor primer (AAP).

Water to 50 μl.

2. In a thermocycler perform the PCR reaction using the following conditions:

   1 min at 94 °C.

   (Repeat 25–30 cycles) 30 s at 94 °C, 30 s at 55–65 °C, 2 min at 72 °C.

   72 °C 5 min.

3. The product of this reaction can be analyzed by agarose gel electrophoresis and ethidium bromide staining according to standard protocols.

4. If no bands appear in this gel another round of PCR (nested PCR) is required to amplify low abundance products not visible in the first PCR reaction (*see* **Note 17**). Perform a nested PCR as described:

   0.5 μl of TaKaRa LA Taq.

   25 μl of 2× GC rich buffer.

   8 μl of dNTPs.

   1 μl of DNA obtained in the first PCR reaction.

   2.5 μl of GSP3 (from 10 μM stock).

   2.5 μl of anchor primer (AUAP) (for sequence *see* Table 1).

   Water to 50 μl.

5. In a thermocycler perform the PCR reaction using the following conditions:

   1 min at 94 °C

   (Repeat 25–30 cycles) 30 s at 94 °C followed by 30 s at 55–65 °C.

   2 min at 72 °C.

   72 °C 5 min.

6. A portion of the product of the nested PCR can be analyzed on agarose gel to visualize the bands that will be submitted for next-generation sequencing (*see* **Note 18**, Fig. 3).

3.2 Rapid Amplification of 3′-End

To determine the termination of an RNA transcript the 3′ RACE protocol takes advantage of the natural poly(A) tail found in mRNA as a priming site for the reverse transcription step. However,

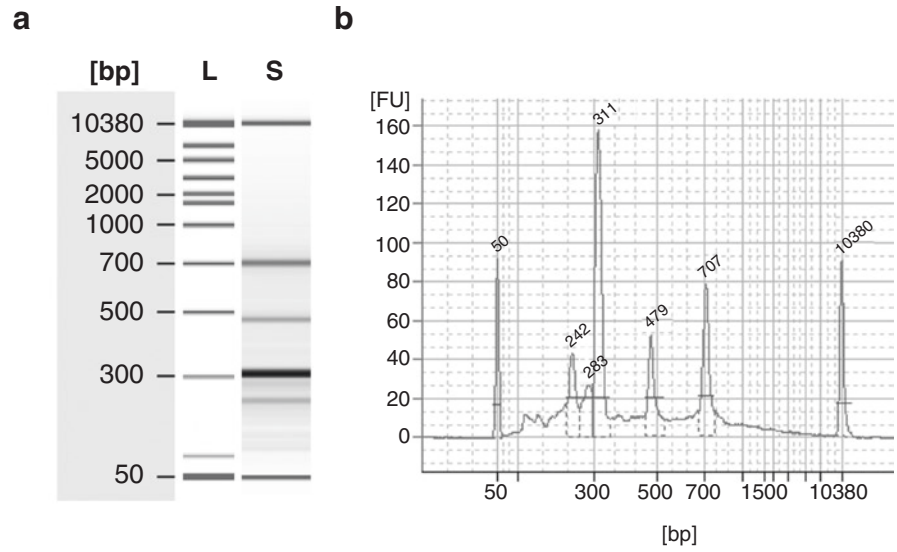


Fig. 3 (**a**) Densitometry plot of a representative 5′-RACE sample run on a bioanalyzer shows multiple products of different lengths. This step is important to confirm the presence of RACE products, and to determine whether size fragmentation is necessary. (**b**) Bioanalyzer results can also be visualized in aelectropherogram tab that displays a data plot of size/migration time versus fluorescence intensity. L = ladder; S = sample

long noncoding RNAs are frequently not polyadenylated, which precludes the use of this procedure. Below is the schematic for the 3′-end RACE protocol (Fig. 4). Therefore before starting it is important to determine whether or not the transcript of interest is polyadenylated (*see* **Note 19**).

If the noncoding RNA is not polyadenylated, a polyA tail can be added artificially using the *E. coli* poly(A) Polymerase (*see* **Note 20**).

3.2.1 *Primer Design*

The gene-specific GSP1 and GSP2 primers must be designed as forward primers for the transcript of interest (right primers). Design GSP1 in the same region as was used for the 5′ RACE GSP1 (Fig. 5).

3.2.2 *RNA Isolation*

For RNA isolation follow the same suggestions described for 5′ RACE, (Subheading 3.1.2)

3.2.3 *cDNA Synthesis*

1. First-strand cDNA synthesis is initiated at the poly(A) tail (natural or artificial) using the adapter primer (AP). Prepare the first-strand cDNA synthesis in a microcentrifuge tube adding the following components:

   Up to 5 μg of total RNA (*see* **Note 10**).

   RNAse-free water to 15.7 μl.

   1 μl of AP (10 μM stock).

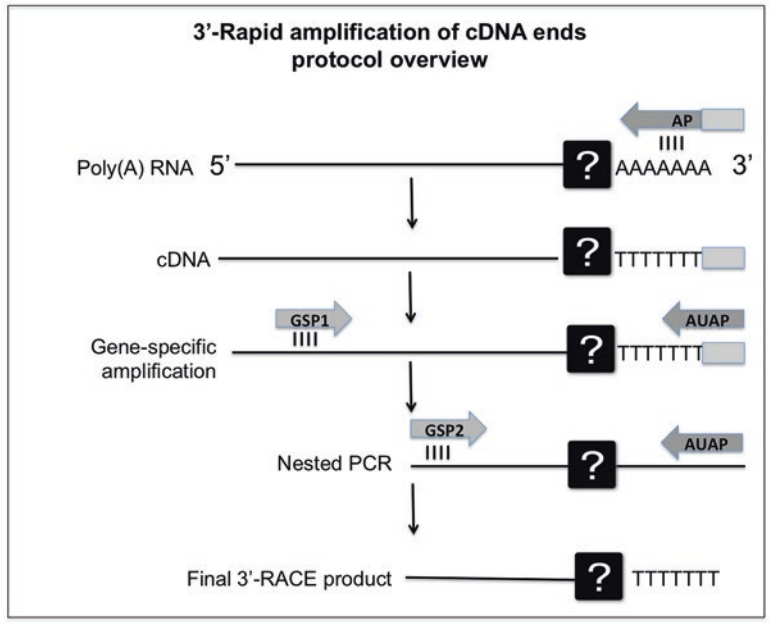


Fig. 4 Schematic of 3′-end RACE. First-strand cDNA synthesis is initiated at the poly(A) tail of RNA using the adapter primer (AP). After first-strand cDNA synthesis a first PCR amplification is performed using two primers: one is a user-designed GSP1 that anneals to a site located within the cDNA molecule; the other is a universal amplification primer (AUAP) that targets the sequence that was introduced by the adapter primer during cDNA synthesis. Since the 3′ RACE utilizes the poly(A) tail region as an initial priming site, multiple amplification products may be synthesized, depending on the degree of specificity conferred by the GSP1. To generate a specific amplification product, the user will design a second gene-specific primer (GSP2) and reamplify the RACE products

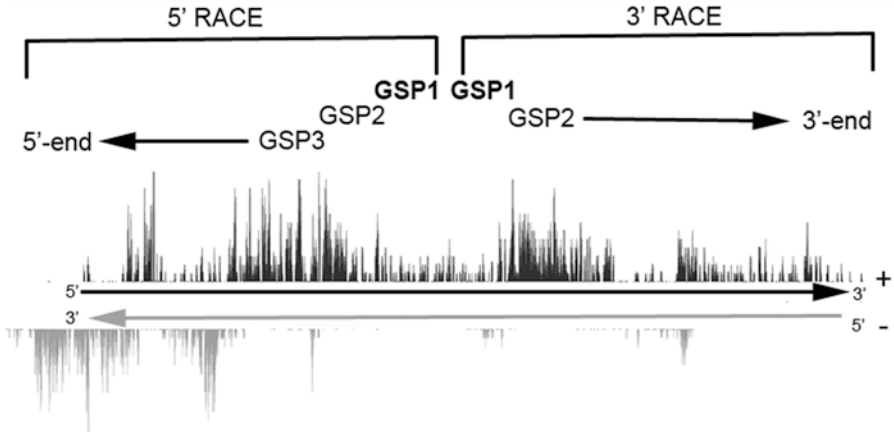


Fig. 5 As an example, we show an intergenic genomic region where transcription is generated from the plus and minus strands of the DNA. If 5′ and 3′-RACE will be performed to characterize the RNA transcribed from the plus strand, it is important to design gene-specific primers for the transcript of interest. GSP primers for the 5′ RACE will be the reverse sequence of the noncoding RNA and the 3′ RACE primers will be in the forward orientation. Moreover, we suggest designing the GSP1 for 5′ RACE and the GSP1 for the 3′ RACE as close together as possible to achieve better coverage of the locus

2. Incubate at 65 °C for 5 min.
3. Place on ice for 10 min and add the following reagents:

   2 μl of 10× Affinity Script RT Buffer.

   0.8 μl of dNTPs mix (100 mM, 25 mM each).

   1 μl of Affinity Script Multiple Temperature RT.
4. Place the tubes in a thermocycler and run the reverse transcription for 60 min at 55 °C. Terminate the reaction at 70 °C for 15 min and then store at −20 °C or proceed to the cDNA purification.
5. Purify cDNA according to manufacturer's instruction.

3.2.4 PCR

Amplification of cDNA is performed with the GSP1 and the Abridged Universal Amplification Primer (AUAP, the same used in the 5′ RACE) homologous to the adapter region of the AP primer used for cDNA synthesis. Thus far, the first-strand synthesis has produced a pool of all the poly(A) transcripts present in the template RNA. This amplification, using the GSP1, is therefore the first step generating a "gene-specific" product.

1. Set up the PCR reaction as described:

   0.5 μl (2.5 units) of Taq Polymerase.

   25 μl of 2× GC rich buffer.

   8 μl of dNTPs.

   1 μl of cDNA.

   2.5 μl of GSP1 (from 10 μM stock).

   2.5 μl of anchor primer (AUAP).

   Water to 50 μl.
2. In a thermocycler perform the PCR reaction using the following conditions:

   1 min at 94 °C.

   30 s at 94 °C → 30 s at 60 °C (25–30 cycles).

   2 min at 72 °C.

   72 °C 5 min.

   (see Notes **15, 16**, and **17**).
3. Another round of PCR (nested PCR) using a new gene-specific primer for the same target (GSP2) and AUAP is highly recommended to increase the specificity of the amplified product.

   0.5 μl (2.5 units) of Taq polymerase.

   25 μl of 2× GC rich buffer.

   8 μl of dNTPs.

   1 μl of DNA amplified in the first round.

2.5 μl of GSP2 (from 10 μM stock).

2.5 μl of anchor primer (AUAP).

Water to 50 μl.

4. In a thermocycler perform the PCR reaction using the following conditions:

   1 min at 94 °C.

   30 s at 94 °C → 30 s at 60 °C (25–30 cycles).

   2 min at 72 °C.

   72 °C 5 min.

   (see Notes **15, 16**, and **17**).

A portion of the product of the nested PCR can be analyzed on agarose gel to visualize the bands that will be submitted for next-generation sequencing (*see* **Note 18**).

3.3 Library Construction for Next-Generation Sequencing and Data Analysis

RACE products can be submitted for next-generation sequencing upon library preparation. A suitable kit would be the one used for chromatin immunoprecipitation sequencing (*see* **Note 21**).

3.3.1 Library Preparation

1. Quantify the material with systems like Qubit dsDNA HS assay (ThermoFischer Scientific) (*see* **Note 22**).
2. Run the material on the bioanalyzer: dilute 1:20 the sample and combine 2 μl of it with 2 μl of High Sensitivity D1000 sample buffer. Vortex, spin down, and load the sample on the chip (*see* **Note 22**).
3. A fragmentation step might be required if the bio analyzer results demonstrate bands bigger than 300 bp (*see* **Note 23**). DNA fragmentation kits using enzymatic digestion are commercially available.
4. Run fragments on an agarose gel to confirm that they are 200–300 bp in length.
5. 5–10 ng of double-strand DNA is enough to start the library preparation.
6. Before the library construction increase the cluster complexity (*see* **Note 24**).
7. Sequencing parameters: we recommend performing paired-end sequencing with reads length between 75 and 100 bp. The number of total million reads depends based on the sequencing machine model, in general one run can accommodate 8–16 samples. For more details discuss with the personnel of the sequencing facility.

3.3.2 Data Analysis

The raw reads resulting from sequencing of the 5′ - and 3′ RACE PCR products on the sequencer contain the RACE adapter primer sequences that prevent the reads from aligning to the genome. In order to remove the PCR primer sequences, read trimming

software such as TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) can be used to supply the PCR adapter sequences and trim the raw reads. After trimming, the reads can be aligned to the genome reference using a gapped aligner, such as TopHat (http://tophat.cbcb.umd.edu) or STAR(https://github.com/alexdobin/STAR), to perform discovery of spliced junctions. Aligned reads in the BAM or SAM format are further used to perform ab initio reconstruction of transcripts using Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/). ab initio approach allows reconstruction of transcript structures without a priori knowledge and thus enables discovery of novel splice variants. Cufflinks produces Gene Transfer File (GTF) that can be further used to visualize the transcript structures (i.e., splicing) revealed with RACE using visualization tools such as Integrated Genome Viewer (IGV) (https://www.broadinstitute.org/igv/) (Figs. 6 and 7). Optionally, GTF files from 5′ and 3′ RACE can be combined using Cuffmerge module of Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/cuffmerge/) to produce a file containing the full structure of the locus.

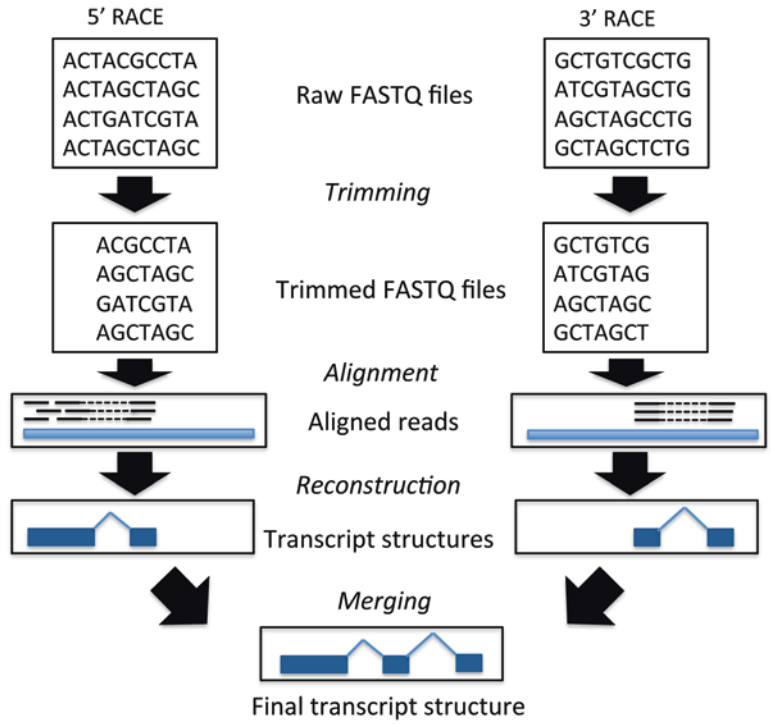


Fig. 6 RACEseq data analysis. First, the raw sequencing reads are trimmed to remove RACE adapters using a program such as TrimGalore. After that the reads are aligned to the genome using TopHat or STAR. Aligned reads are reconstructed with Cufflinks and merged with Cuffmerge to obtain complete transcript structures

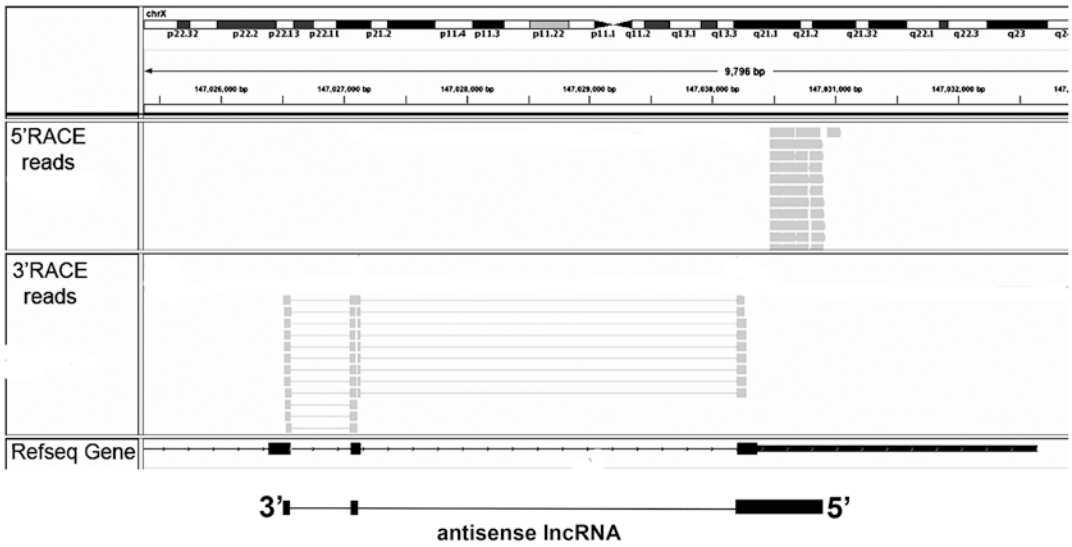


Fig. 7 The reads coming from next-generation sequencing can be visualized using the program IGV to display transcript mapping

4 Notes

1. The anchor primer AAP must incorporate the modified deoxyinosine base (i) in the 3′ region in the positions indicated (*).
2. RNAseq data for several cell types and cell lines are nowadays available at the Gene Expression Omnibus (GEO), the Sequence Reads Archive (SRA) websites (http://www.ncbi.nlm.nih.gov/geo/,http://www.ncbi.nlm.nih.gov/sra) and the UCSC genome browser (https://genome.ucsc.edu/) (Fig. 1). Expressed Sequence Tags (ESTs) (http://www.ncbi.nlm.nih.gov/unigene) can also provide information on the potential locations of lncRNAs. Finally, strand-specific RT-qPCR can verify that transcripts are present at a particular genomic locus and guide the design of gene-specific primers for use in the RACE protocol.
3. Many of the steps outlined here are analogous to those described in commercially available RACE kits. You can choose to follow manufacturer's instructions or if reagents were purchased separately, perform the steps as outlined here.
4. We suggest designing at least three GSP1 primers to achieve an efficient (and specific) reverse transcription of the transcript of interest.
5. The melting temperature (*T*m) of AAP is 66 °C, the *T*m of AUAP is 60 °C. Design the GSP2 and GSP3 to have *T*m that are compatible with their corresponding adapter primers.

6. To extract high quality RNA, we recommend using Trizol since it is free of protein, RNAse-free, and free of guanidinium salt (which inhibits the subsequent reverse transcription (RT) step).
7. Take care not to disturb the lower layer or the interface, as this can cause contamination of your RNA sample with DNA and/or protein.
8. We recommend resuspending the RNA in 30–50 μl of RNAse-free water.
9. We recommend including a DNAse treatment step after the RNA extraction (turbo DNAse or similar); however, a small amount of retained genomic DNA (gDNA) may not be detrimental to the 5′ RACE protocol. This is because the "cDNA tailing step" only uses single stranded DNA (meaning cDNA but not gDNA).
10. For the reverse transcription step we recommend using the Affinity Multiple Temperature cDNA Synthesis kit (Agilent Technologies) that allows for reverse transcription at the preferred reaction temperature.
11. Generally, RACE kits are not optimized to detect 5′-ends of low-abundance noncoding RNA transcripts; therefore, we recommend starting with the maximum amount of RNA allowed by the cDNA protocol.
12. A temperature of 55 °C is recommended in the reverse transcription step to facilitate cDNA synthesis throughout RNAs that have secondary structures. Any cDNA synthesis kit can be used in addition to the one mentioned here.
13. In order to prevent the excess GSP1 left over in your sample from interfering with cDNA tailing, it is necessary to clean up the cDNA. Several kits are available for the purification of nucleic acids.
14. In this step the TdT enzyme is used to attach a stretch of cytosines to the 3′ end of the purified cDNA. Since the 3′-end of the cDNA corresponds to the 5′-end of the transcript of interest, this results in attachment of nucleotides to the original 5′-end. The addition of cytosines to the 3′-end of the cDNA creates the binding site for the abridged anchor primer (AAP) which will be used in a PCR reaction in combination with a second gene-specific primer (GSP2). To amplify the tailed cDNA use a "long range" DNA polymerase.
15. A gradient PCR can be done to ascertain that the annealing temperatures used with AAP/AUAP and the gene-specific primers produce bands. Exclude any temperatures that do not produce bands. All other products can be combined for sequencing since nonspecific bands that do not align to the locus of interest will be excluded during the analysis.

16. If there are GC-rich sequences in the region of interest, use a buffer optimized for amplification of GC-rich sequences.
17. We suggest limiting cycles to not more than 25–30 since we have noticed that using fewer cycles reduces the appearance of nonspecific bands. If necessary, it is better to introduce a second nested PCR with GSP4.
18. To preserve as much as possible of the final product of the 5′ RACE protocol, we recommend visualizing the band with a bioanalyzer machine since as little as 1 μl can be sufficient to run the microfluidic chip (i.e., Agilent DNA analysis kit) (Fig. 3).
19. If the transcript of interest does not overlap a protein coding gene, it is possible to determine whether a transcript is polyadenylated by creating cDNA using oligodT primers, followed by a PCR amplifying the region where the ncRNA of interest is located. If the transcript is polyadenylated, a band will be visible on an agarose gel. If a protein coding gene overlaps with ncRNA it will not be possible to discriminate which strand gives rise to the polyadenylated transcript.
20. Protocol for poly(A)tailing of RNA using *E. coli* Poly(A) polymerase.

    5 μg of RNA (in 15 μl of water).

    2 μl of 10× reaction buffer.

    2 μl of ATP.

    1 μl of Poly(A) polymerase.

    Incubate reaction at 37 °C for 30 min. Stop the reaction directly proceeding to the cleanup step.
21. Several companies offer Kits for Chip sequencing, i.e., NEB Next ChIP-Seq Library Prep Reagent Set for Illumina (NEB), DNA SMART ChIP- Seq Kit (Clonetech), TruSeq ChIP Library Preparation Kit (Illumina).
22. Do not use nanodrop to quantify the material since it is not accurate enough.
23. RACE PCR product will be heterogeneous in length but next-generation sequencing instruments only sequence 75–100 bp. During the library preparation the size selection step will exclude the large fragments. Therefore, it is necessary to fragment the RACE PCR products to ensure that longer transcripts are completely sequenced from start to end.
24. Since the RACE products will contain almost exclusively DNA fragments belonging to a single locus; therefore, this can represent a "low complexity/clustering" issue during the sequencing. To solve this problem Illumina provides a kit for low-diversity sequencing. The protocol involves spiking-in 10 % of a viral genome (PhiX) to the RACE samples to increase

the cluster complexity. PhiX Control v3 is an adapter-ligated library that provides a quality control for cluster generation, sequencing, and alignment.

References

1. Napoli S, Pastori C, Magistri M et al (2009) Promoter-specific transcriptional interference and c-myc gene silencing by siRNAs in human cells. EMBO J 28(12): 1708–1719
2. Pastori C, Peschansky VJ, Barbouth D et al (2014) Comprehensive analysis of the transcriptional landscape of the human FMR1 gene reveals two new long noncoding RNAs differentially expressed in Fragile X syndrome and Fragile X-associated tremor/ataxia syndrome. Hum Genet 133(1): 59–67
3. Olivarius S, Plessy C, Carninci P (2009) High-throughput verification of transcriptional starting sites by Deep-RACE. BioTechniques 46(2):130–132
4. Denise H, Moschos SA, Sidders B et al (2014) Deep sequencing insights in therapeutic shRNA processing and siRNA target cleavage precision. Mol Ther Nucleic Acids 3:e145
5. Choi NM, Loguercio S, Verma-Gaur J et al (2013) Deep sequencing of the murine IgH repertoire reveals complex regulation of non-random V gene rearrangement frequencies. J Immunol 191(5):2393–2402

Chapter 7

In Silico Prediction of RNA Secondary Structure

Fariza Tahi, Van Du T. Tran, and Anouar Boucheham

Abstract

The secondary structure of an RNA molecule represents the base-pairing interactions within the molecule and fundamentally determines its overall structure. In this chapter, we overview the main approaches and existing tools for predicting RNA secondary structures, as well as methods for identifying noncoding RNAs from genomic sequences or RNA sequencing data. We then focus on the identification of a well-known class of small noncoding RNAs, namely microRNAs, which play very important roles in many biological processes through regulating post-transcriptionally the expression of genes and which dysregulation has been shown to be involved in several human diseases.

Key words RNA secondary structure, Pseudoknots, RNA secondary structure prediction, RNA identification, Noncoding RNA, MicroRNA, Integrative approach

1 Introduction

Knowledge on structures of macromolecules, their interactions, and associations is central to the understanding of the living. For many years, structural biology was mostly associated with the structure of proteins, and ignored that of non-protein-coding RNAs, which also adopt a three-dimensional structure (Fig. 1) and are involved in many biological processes via their structural interactions. In the 1980s, the discovery of catalytic RNAs had revolutionized the perception of the origins of life and led to the development of many studies on RNA. The discovery of small interfering RNAs, involved in the inhibition or the regulation of gene expression, opened a huge field of investigation, both in fundamental biology and medicine. And today, we are constantly accumulating the knowledge on RNA structures, their folding process, evolution, and catalysis. RNAs are now at the heart of many research studies as their roles are very important in a large number of biological processes.

The concept of secondary structure was introduced by Doty and Fresco in 1959 [1]. The RNA secondary structure, referring to the

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_7,

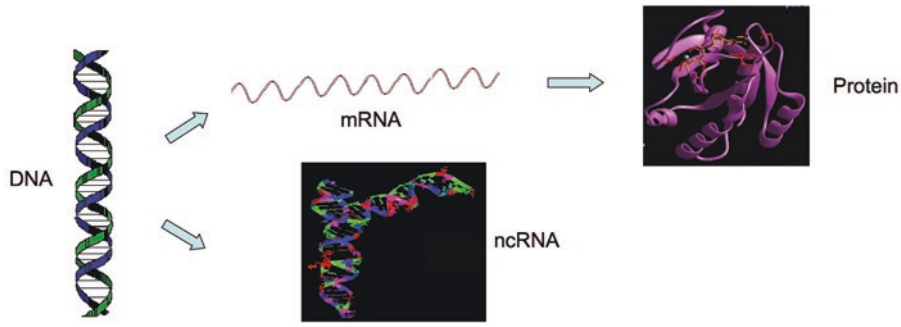


Fig. 1 Biological functions are expressed through two ways: proteins and RNAs. Proteins are encoded from mRNAs. Noncoding RNAs do not encode proteins

canonical interactions in the RNA molecule (composed of Watson-Crick and Wobble base pairings), is known as a simple well-defined representation of its complex tertiary structure, which also includes noncanonical interactions. The RNA secondary structure has interested biologists since the publication in 1965 by Holley et al. on the structure of alanine transfer RNA [2]. Knowledge on secondary structure is essential to understand the relation between structure and function. Identification of the former can contribute in some extent to the understanding of activities of the RNA in the involved biological process as well as phylogenetics. The last years have seen considerable developments in the principal experimental techniques to determine the structures of macromolecules (nuclear magnetic resonance, crystallography, X-ray diffraction, mass spectrometry, etc.). In particular, a new generation of experimental techniques, called SHAPE-seq, allows the identification of some substructures of an RNA secondary structure [3]. However, to deal with the molecular complexity and the huge volume of accumulating data, structural biology must coordinate its efforts with other domains like computer science and mathematics, whose expertise is crucial, especially in processing of big data and modeling of complex objects. The challenges of this cooperation are important and concern many areas of life sciences.

In this chapter, we focus on a particular field of this interdisciplinary cooperation, the RNA bioinformatics. With the new discoveries on RNAs, more particularly noncoding RNAs (ncRNAs), and on the importance of various roles they can play in biological processes at different levels, RNA bioinformatics has certainly become one of the most important fields in bioinformatics and computational biology. We will provide here an overview of RNA secondary structure and noncoding RNA prediction. The chapter will be organized as follows. We present in the first section definitions and descriptions on RNA secondary structures, and then in the second section the main approaches and available tools for the in silico prediction. We discuss in the next section the challenges in RNA secondary structure prediction, in particular the prediction of pseudoknots, and some

characteristics that are expected from developed tools. In the following section, we show how RNA secondary structures can be compared, in order to deduce some structure homology, for instance, and thus some function homology between different RNAs. Then, the fifth section shows some approaches and existing computational methods for ncRNA identification in genomic sequences and RNA-seq data, which are largely based on a first step of secondary structure prediction. Finally, we review the challenges in predicting a particular class of ncRNAs, microRNAs, and the existing identification tools that are mostly based on their proper structure and characteristics.

2 RNA Secondary Structure

RNA molecules have three levels of structure (*see* Fig. 2):

- Primary structure or the sequence of nucleotides.
- Secondary structure, representing the folding of the sequence through Watson-Crick base pairings (A-U and G-C) and Wobble base pairing (G-U). These pairings, known as canonical interactions, are formed by hydrogen bonds between the corresponding nucleotides.
- Tertiary structure or the three dimensional conformation of the nucleotide chain. It results from the folding of the secondary structure via diverse noncanonical pairings, of which over 150 types have been observed [4].

The secondary structure of RNA corresponds to the shape (or topology) induced by all base pairings A-U, G-C, and G-U from the single-stranded molecule. It is thus composed by matched regions, called stems (or helices), and unpaired regions, called loops (*see* Fig. 3). Loops can be classified into several types:

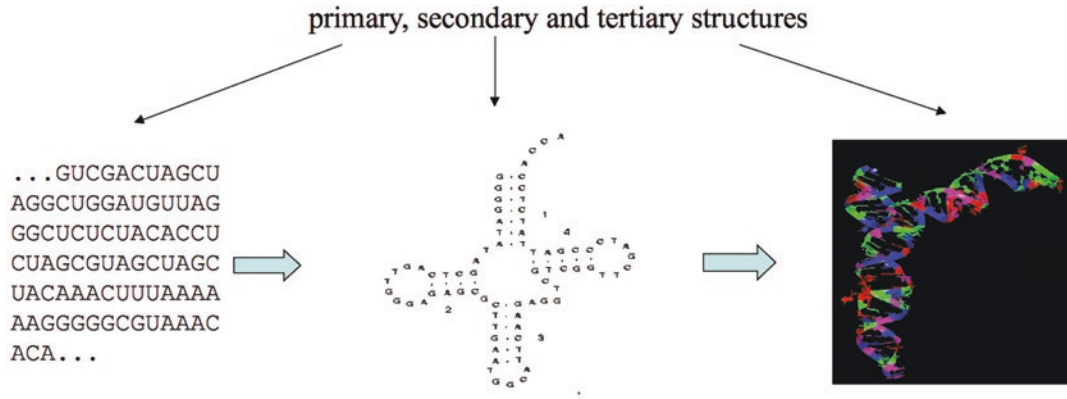


Fig. 2 The three structure levels of RNAs

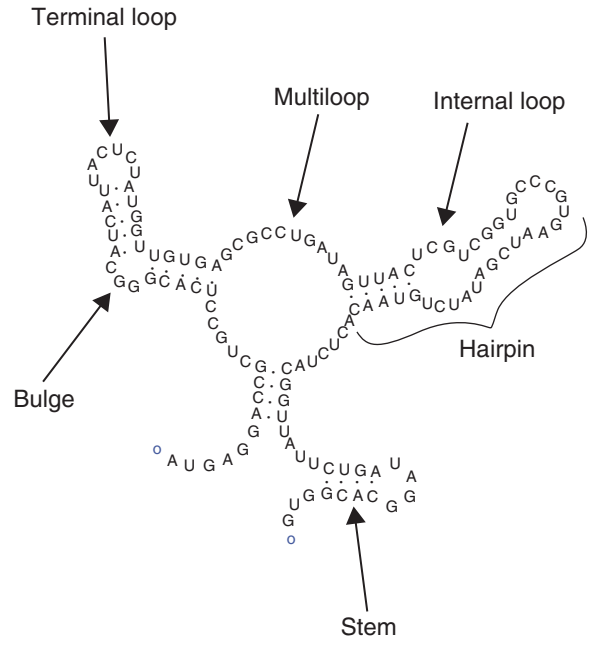


Fig. 3 Components of RNA secondary structures

- Internal loops, which connect two stems and are formed by two strands of equal sizes (symmetric loops) or unequal sizes (asymmetric loops).
- The bulges, which are unpaired strands occurring in only one side of the stems.
- Terminal loops, which are located at the extremity of a stem and have at least four nucleotides (nt).
- Multiloops, which connect three or more stems.

We can find several kinds of motifs in RNA secondary structures. The best known is the stem-loop or hairpin motif, which is composed of a stem, or possibly a series of stems separated by bulges or internal loops, and a terminal loop (*see* Fig. 3). Several small RNAs, such as microRNAs, possess this conformation (*see* Subheading 6, Fig. 6).

A secondary structure is essentially defined by its shape, and more precisely by the configuration of stems and the nature of unpaired bases. Paired and unpaired nucleotides can be represented in a simplified manner by brackets ("(" or ")") and dots (".") respectively. For example, the RNA structure of Fig. 3 can be represented as follows:

……((((…..(((((..(((((……)))))))))).…….(((((…(((((……..)))))…..)))))……))))….((((….))))…

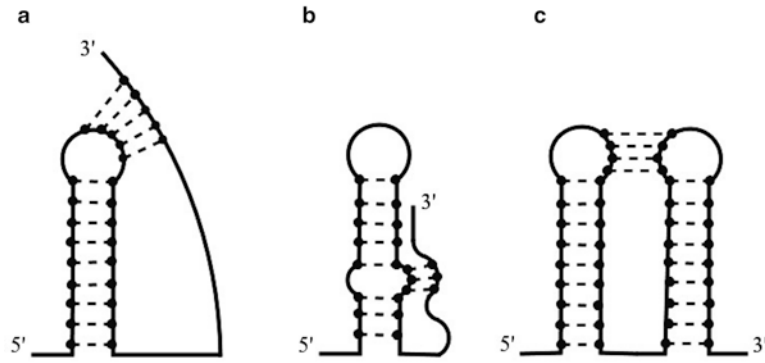


Fig. 4 Pseudoknots of different types: (**a**) H-type pseudoknot; (**b**) B-type pseudoknot; (**c**) Kissing-hairpin pseudoknot

Such a shape can be involved in: (1) formation of the tertiary structure (through noncanonical interactions); (2) biological functions of the RNA (e.g., the end loop of the clover-leaf structure of a transfer RNA, where the three bases forming the anticodon permit transcription of a codon into an amino acid); (3) interactions with other RNAs or with proteins.

Many RNA molecules contain pseudoknots in their secondary structure. A pseudoknot is a structure formed by Watson-Crick and Wobble base-pairings between a loop and a region of the same RNA that is outside the stem delimiting the loop (*see* Fig. 4). In other words, a pseudoknot is composed of at least two interleaved stems.

A secondary structure with pseudoknots loses its two-dimensional conformation. For this reason, the pseudoknots were excluded from the classical definition of RNA secondary structure, and were considered substructures of the tertiary structure. The pseudoknots were first demonstrated by Pleij et al. in 1985 [5]. They showed that these substructures were involved in various strategic regions of RNA molecules, such as ribosome functioning, RNA splicing, and recognition of tRNA-like structures. Following studies discovered significant roles of pseudoknots, especially in the regulation of biological processes. Experimental observations suggested, for instance, their role as "switches" or control elements in several biological functions [6].

3 RNA Secondary Structure Prediction

The prediction of RNA secondary structures is one of the most important studied issues in RNA bioinformatics.

Knowledge on those structures is implicated in several biological interests. The secondary structure prediction is a simple and

well-defined way toward the resolution of complex tertiary structures. Besides, comparing those secondary structures allows for sequence classification and phylogeny study, as they are often better preserved than the sequences themselves. In addition, some motifs in secondary structures, such as hairpins, can play important roles in some control processes.

3.1 Main Approaches

The computational prediction of RNA secondary structures has been developed with two main approaches. The first one is based on the thermodynamics of the molecules, whereas the second one relies on the conservation between organisms. Most of existing prediction tools used one or the other, while some tried to combine the two approaches.

3.1.1 Thermodynamic Approach

The principle is that the real structure is the most thermodynamically stable, i.e., the one of minimum free energy. The free energy is computed using thermodynamic parameters defined experimentally by Mathews et al. [7]. The limitation of this approach is partly due to the uncertainty of the energy model. Indeed, although it is admitted that, to be stable, the real structure should have a low energy, it is not necessarily the one of the minimal free energy. The reason is that the folding of RNA depends on its environment and is often affected by other macromolecules (other RNAs, protein, etc.).

Nussinov and Jacobson proposed the first efficient algorithm in this approach [8]. It implemented the dynamic programming technique, where all possible pairings were considered in the minimization of the global free energy. An improvement of this algorithm was proposed by Zuker, who developed later the Mfold program [9, 10], with a time complexity of $O(n^3)$, where n is the sequence length. This computational complexity gives the information about the ability of the algorithm to deal with the analyzed sequences. The higher the complexity, the lower the maximal size of the sequences the tool can analyze (*see* Subheading 3.3.5). Another algorithm, RNAfold, was proposed by Hofacker et al. [11]. It was based on McCastill model—a computation of base-pair probabilities, and exploited statistical methods on RNA structure thermodynamics (e.g., statistical sampling of secondary structures according to their equilibrium probabilities). Mfold and RNAfold are the two most widely used algorithms to predict the secondary structure of a given sequence, certainly because Mfold is one of the first tools proposed for automatic prediction of RNA secondary structures, and RNAfold is from the well-known ViennaRNA package. The two algorithms give fairly similar results, results that have been shown to be less good than the ones obtained by the comparative approach [12]. Note however that Mfold can return several possible secondary structures (*see* Subheading 3.3.2), which is not the case of RNAfold.

Other attempts in this approach were also implemented. We can mention, for example, the algorithm developed by Ninio [13] and the one developed by Martinez [14], which recursively search for stems, keeping at each step the stems forming a structure with an energy not exceeding a given threshold.

3.1.2 Comparative Approach

The comparative approach is applied when multiple homologous RNA sequences (i.e., sequences of the same RNA belonging to different species) are available [15]. The principle is to search for covariations between nucleotides of different sequences, i.e., compensatory mutations that maintain the pairings and thus sequence folding. With this approach, one could manually determine secondary structures of some ribosomal RNAs, such as 16S rRNA and 23S rRNA, which contain a few thousands of nucleotides [16–19], and therefore could not be treated by existing thermodynamics-based prediction tools (these tools, because of their high computational complexity in time, which is of $O(n^3)$, were indeed not able to deal with long RNAs).

The first algorithm, with reasonable time and space complexities, was proposed in [20], where the five most plausible secondary structures common to m homologous sequences of lengths equal to n are produced in time complexity of $O(m^*n^2+n^3)$. Likewise, another implementation was then proposed in [21]. The algorithm is based on dynamic programming technique, using SCFGs (Stochastic Context-Free Grammars). We can also cite Pfold [22], which is based on context-free grammars and has a complexity of $O(n^3)$. More recently, an algorithm with a complexity of $O(n^2)$, called Tfold [23], was introduced, using an iterative strategy where stems are identified recursively according to their length and covariation.

An important limitation of the comparative approach is that prediction results are highly dependent on the homologous sequences used for comparison and the quality of alignment. To avoid this dependency, some algorithms perform the sequence alignment simultaneously with the search for a common secondary structure. This produces then a high complexity, and thus they can only deal with very few sequences, usually two. We can cite, for instance, Foldalign [24], Dynalign [25], PARTS [26], and RAF [27]. Tfold [23] manages this dependency in a different manner. The algorithm takes as input a set of aligned sequences, obtained with ClustalW [28], and then selects a subset of homologous sequences that are the most informative for the secondary structure prediction with the algorithm SSCA [29] developed by the same group. For that selection purpose, SSCA chooses the sequences (1) that are neither too close to, nor too far from the target sequence for prediction, and (2) that are well aligned.

3.2 Pseudoknots Prediction

The prediction of pseudoknots in RNA secondary structure has been avoided for a long time. The prediction of pseudoknots is too time-consuming, because computational complexities in time are too high. Most algorithms proposed for this issue are limited to the simplest classes of pseudoknots, mostly the H-type class (*see* Fig. 4). For example, in [30], a method based on genetic algorithms approach was developed to predict secondary structures including H-type pseudoknots. In [31], a dynamic programming algorithm based on the energy minimization model, called pknots, was proposed for predicting some simple kinds of pseudoknots in time complexity of $O(n^6)$. Standard RNA folding parameters and those describing the thermodynamic stability of pseudoknots were used. Reeder and Giegerich presented an algorithm of $O(n^4)$, called pknotsRG [32], which can predict structures with pseudoknots of H-type class. Dawson et al. proposed vsfold [33] using a structure mapping and thermodynamics for RNA pseudoknot prediction. Besides, Akutsu's group also tried to solve the free energy minimization problem under constraints describing RNA secondary structures with recursive pseudoknots using integer programming [34]. They developed later another method, IPknot, using integer programming with a base-pairing probabilistic model [35], a model that was exploited before by Bellaousov and Mathews in an algorithm called ProbKnot [36]. IPknot, as well as ProbKnot, allow predicting any topology of pseudoknots. IPknot shows better prediction results and running time as compared to several other existing tools, including ProbKnot and pknotsRG [35].

Some assumptions stipulate that for kinetic reasons, the real RNA secondary structure often has a local free energy minimum rather than a global one [37]. A number of algorithms take into account these kinetic features to minimize the free energy in a local area. They attempt to simulate the RNA folding process iteratively by adding stems instead of base pairs [38–40]. The iterative strategy for stem selection allows reducing the search space and then exploring for the structures with pseudoknots. For example, ILM [38] implements dynamic programming to predict pseudoknots with an average complexity of $O(n^3)$, and a worst complexity of $O(n^4)$. It takes as input a single sequence or a set of aligned sequences, and uses either thermodynamic or comparative information or both. Another algorithm using this strategy is Tfold [23], which predicts in a first step the structures without pseudoknots, and then in a second step the pseudoknots. Tfold searches for all possible stems (verifying criteria of length and covariation) that are compatible, i.e. do not overlap and therefore do not form pseudoknots. Subsequently, it searches for all other possible stems, which are not compatible with the previously determined ones, forming thus pseudoknots. Compared to other existing algorithms, Tfold has the lowest time complexity, which is of $O(n^2)$, without restriction on topology of predicted pseudoknots. It also gives better predic-

Table 1
Results obtained by knotSeeker, pknotsRG, vsfold, and Tfold for predicting pseudoknots in several RNAs. The number of correctly predicted pseudoknots over the total number of known pseudoknots is provided as well as the number of false positives (FP)

	tRNA	5sRNA	u1RNA	srpRNA	tmRNA	RNAseP
knotSeeker	0/1 1 FP	0/0 0 FP	0/0 1 FP	0/0 1 FP	2/4 1 FP	0/2 1 FP
pknotsRG	0/1 0 FP	0/0 1 FP	0/0 0 FP	0/0 1 FP	0/4 0 FP	0/2 0 FP
vsfold5	0/1 1 FP	0/0 1 FP	0/0 2 FP	1/1 5 FP	1/4 0 FP	0/2 4 FP
Tfold	0/1 0 FP	0/0 1 FP	0/0 0 FP	1/1 0 FP	3/4 1 FP	2/2 0 FP

tion results as compared to several other existing algorithms, including ILM, vsfold, and pknotsRG. Some others tools were developed for detecting only pseudoknots but not the folding structures. These take as input a sequence in knotseeker [41], or an RNA secondary structure without pseudoknots in Hfold [42].

Table 1 shows the results of some existing tools on their ability to predict pseudoknots. To have a large comparison between several existing tools, refer to [23] and [35].

3.3 Challenges in RNA Secondary Structure Prediction

3.3.1 Combining Several Models

Because no model could be considered as the absolute model defining the RNA structure, it would be useful to combine several models to take benefits from each of them. It is the case of RNAalifold [43], which combines the thermodynamic and comparative models, to predict a secondary structure common to a set of homologous sequences in time complexity of $O(n^3)$. ILM [38] also allows using both of the two models for predicting pseudoknots (*see* above). Other more recent methods combine covariation and thermodynamic information using machine learning techniques, computational methods of artificial intelligence that allow learning from example inputs and then building models to perform predictions. In this category, we can cite KnotShape [44], which is based on support vector machine (SVM) method.

3.3.2 Prediction of Several Possible Structures

Most of the existing tools propose one solution, the optimal one regarding the considered model. But the folding of a given RNA depends not only on the sequence information but also on the RNA environment. Hence, the optimal solution returned with some specific model is naturally not always the good structure. The closest solution to the real structure can be one of the suboptimal solutions. It is therefore important to exploit tools that can return several possible structures.

The first method proposed for finding suboptimal folding structures of an RNA molecule was proposed by Zuker [45]. Very few tools can return several solutions: Mfold [10] based on the energy minimization model (*see* above), which has been subsequently updated to produce several suboptimal structures [46]; RNAsubopt [47] based on the same model, which enumerates all secondary structures within some energy range; and Tfold [23] based on the comparative approach, which provides all possible structures corresponding to different choices of incompatible stems (*see* above). Note that Tfold is able to return several possible structures that include pseudoknots (of any type). We can also cite pknotsRG [32] that predict pseudoknots of H-type class within a given interval of free energy.

3.3.3 Interactive Tools

As mentioned above, there is no RNA secondary structure predictor that perfectly works and always allows predicting an appropriate structure. Furthermore, biologists may have some information on the structure of the RNA of interest, so they are usually more interested by user-friendly tools with which they can interact in order to:

- Choose among a set of solutions, the one or the ones that are likely to be the most probable structure.
- Set up parameters with values that are suitable to their biological problem.
- Take into account some already known parts of the structure generally obtained from experimental data.

Very few tools are designed for this interaction. Among them we can cite Tfold [23], which allows users: (1) to take into account some already known stems, (2) to choose a subset of homologous sequences to consider for the prediction, (3) to specify if the prediction must include pseudoknots, and (4) to get several possible solutions or only one solution.

3.3.4 Integration of Experimental Data

In recent years, it has appeared a new generation of experimental techniques, which furnish more useful information on RNA secondary structure. Among them, the SHAPE-seq (*Selective 2′-Hydroxyl Acetylation analyzed by Primer Extension sequencing)* technique [3], developed in 2011, allows measuring the reactivity of the RNA nucleotides. This reactivity represents the capability of each nucleotide to be modified by a chemical agent. If the reactivity is high, the nucleotide is easily accessible, and therefore it is not paired. In contrary, if the reactivity is low, the nucleotide would be involved in a pairing. RNA structure predictors have made use of SHAPE data, such as RNAstructure [48]. In this work, it was shown that nucleotide-resolution information from a SHAPE experiment can be interpreted as a pseudo-free energy change term

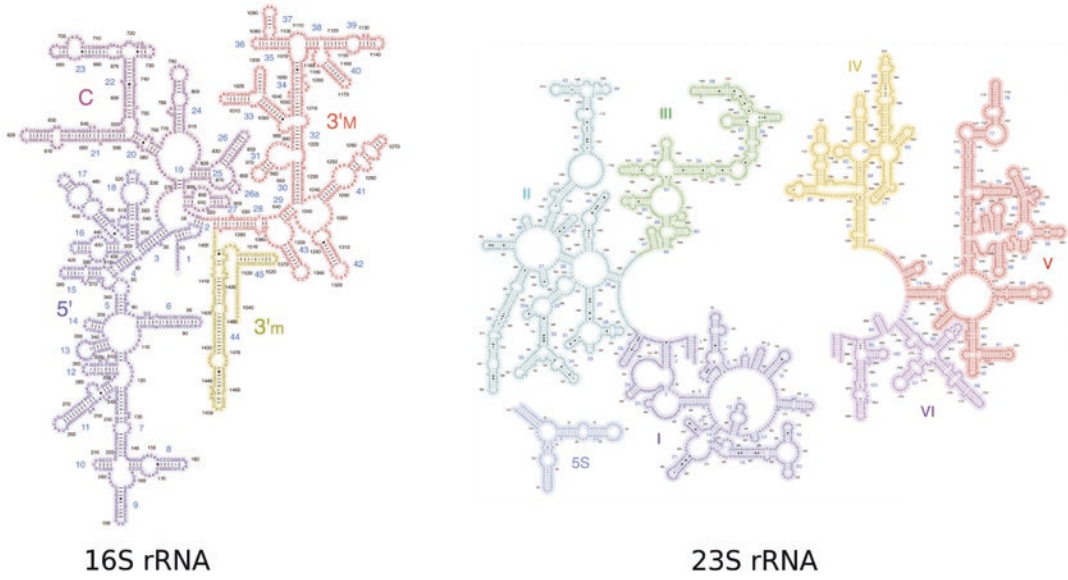


Fig. 5 Secondary structures of 16S and 23S ribosomal RNA

and used to determine RNA secondary structure with high accuracy. SHAPE pseudo-free energies, in conjunction with nearest neighbor parameters, were used for predicting the most probable structure with the lowest free energy.

3.3.5 Large-Scale Analysis

The computing time of RNA secondary structure prediction tools is often very long, especially when the pseudoknots are considered. More problematically, most of the existing tools are not able to analyze long sequences and predict large structures, such as 16S and 23S ribosomal RNAs that contain around 1500 and 2500 nt, respectively (*see* Fig. 5). This is due to the computational complexity in time that is always higher than or equal to $O(n^3)$, except for Tfold that has a complexity of $O(n^2)$. Some algorithms, for instance the ones using integer-programming techniques, have exponential complexities, which make them usable on only small sequences. For 23S rRNA, only algorithms with low complexities, like Tfold [23], are able to predict its structure, particularly when pseudoknots are considered.

4 RNA Secondary Structure Comparison

Computational methods for RNA structure comparison have received an increasing attention recently. This is due to its great importance in the classification of ncRNAs by comparing their structures with consensus structures. Another reason is that structure comparison can also serve to find the function of ncRNA

sequences, based on comparative analysis between their structures and several functionally annotated structural signatures [49].

Generally, structure comparison methods try to measure and calculate how much two structures are different, usually by determining the distance between them. Obviously, the first step is the prediction of the RNA secondary structure, which includes many elements (helices, hairpin loops, internal loops, bulges, and multiloops) and can be represented by trees, bracket strings, or graphs. Both the representation and the dimension of the information considered in the comparison affect the structure comparison techniques and divide them into many categories.

Tree representation of the RNA secondary structure is the most widely used one. It allows more illustration of the different complex elements of the structure (as the position of loops) [50]. Many methods for structure comparison based on the tree representation and using tree edit algorithms have been proposed [51–55]. Hofacker et al. proposed RNAdistance [11], a three-based tool for the comparison of RNA secondary structures. It transforms a tree into another one by many editing operations with specific costs and then considers the smallest sum of the costs along an editing path as the distance between two trees [11]. RNAalign [56] uses only the conventional editing operations on trees (node substitution and deletion). Allali et al. [57] have proposed another tree-based comparison algorithm by introducing two new operations (node fusion and edge fusion). They considered also the classical tree edit operations (deletion, insertion, and relabeling operations) used generally in the other tree-based comparison methods. They developed a dynamic programming algorithm to compare two RNA secondary structures by incorporating these two new operations, with the objective of alleviating some limitations of the other edit operations. Agius et al. [58] have introduced the relaxed base-pair (RBP) score to give a more biologically meaningful measure when differing between two secondary structures obtained from the same RNA. Moreover, an RNA secondary structure can be represented by an ordered labeled tree. Many algorithms have used this representation to perform RNA structure comparison. For example, Dulucq et al. [59] exploited the tree edit algorithm proposed by Zhang and Shasha [60] to efficiently compute the best edit score between two structures in ordered labeled tree representation.

Another way for comparing the RNA structures is the detection of the common RNA secondary structure motifs from a group of related RNA sequences [61–63]. In this direction, Ji et al. [64] have proposed a new graph-theoretical approach to automatic detection of common RNA secondary structure motifs in a set of functionally related RNA sequences. Without requiring a prior structural alignment, it allows finding groups of stable stems conserved across multiple sequences. Then, it assembles compatible conserved stems to form consensus secondary structure motifs [64].

5 In Silico Methods for ncRNA Identification

Many computational methods have been proposed to identify ncRNAs, either in genomic sequences or from RNA-seq data, or both. Many of them perform a classification between coding and noncoding RNAs. Others perform the prediction of a given class of ncRNAs, most of them devoted to microRNAs. Recently, very few try to predict and classify several classes of ncRNAs. In order to perform these predictions, computational methods based on machine learning techniques are largely adopted and used.

5.1 Secondary Structure in ncRNA Identification

Many methods have been proposed to predict ncRNAs based on structural information. They differ depending on the nature and the representation of the secondary structure information processed to perform a prediction.

One of the first works proposed for ncRNA prediction is based on comparative genomics by distinguishing conserved RNA secondary structures from a background of other conserved sequences, using probabilistic models of expected mutational patterns in pairwise sequence alignments [65]. In the same category, Fu et al. proposed Multifind [66], which first predicts the common secondary structure for multiple input sequences using the Multilign tool [67], and then uses several measures of structure conservation to estimate the probability that the input sequences correspond to a conserved ncRNA using an SVM classifier.

NcRNA identification tools are often specific for one class of ncRNA, principally the microRNAs (*see* Subheading 6). Xue et al. proposed a structure-based predictor called Triplet-SVM [68] to classify pre-miRNAs and pseudo pre-miRNAs. They first predict the RNA secondary structures using RNAfold [11], for which they define triplet structure-sequence elements, based on the brackets-dots representation (*see* Subheading 2). For any three adjacent nucleotides, there are 8 (2^3) possible structure compositions: "(((", "((.", "(..", "(.(", ".((", ".(.", "..(", and "…". They consider the middle nucleotide among the three representations, so there are 32 (4 × 8) possible structure-sequence combinations (such as "U(((", "A((.", etc.). This structural information is used to train an SVM classifier to finally do prediction. It should be noted that this technique could be easily used to predict other ncRNAs based on their structure. In the same direction, MiPred [69], proposed as an extension of Triplet-SVM, uses the Random forest model as well as three features: (a) local contiguous structure-sequence composition, (b) minimum free energy (MFE) of the secondary structure, and (c) *p*-value of randomization test. It performs better than Triplet-SVM when giving over 10 % more in accuracy [69].

Another family of methods tries to predict and classify different types of ncRNAs based on the secondary structure. They

mostly use a graph representation of the predicted secondary structure. Graph-based models have been primarily used in ncRNA prediction by Karklin et al. [70]. They used the basic geometric and topological properties of secondary structure in a labeled dual graph representation to train an SVM classifier for classification. Afterward, Childs et al. [71] proposed GraPPLE, a web-based tool that also uses graph properties to classify ncRNAs using the Rfam database [72]. More recently, Panwar et al. have developed RNAcon [73], where graph properties are used with random forest algorithm to classify different ncRNA classes.

5.2 Integrative Approaches for ncRNA Identification

The prediction of ncRNAs based only on structural information is not always sufficient [74], especially for ncRNAs that do not have a known secondary structure, such as piRNAs and many long noncoding RNAs. Integrative computational prediction techniques using other features that can characterize these molecules play now a crucial role in the ncRNA identification field.

The rapid advances in the experimental biological studies about biogenesis, processing, and function of ncRNAs have resulted by an important amount of characteristics and observations that could be useful for their identification and classification. The exploitation of these heterogonous features requires integrating data coming from many sources and in different forms [75]. At the same time, the fast progress in the development of sequencing technologies in recent years has enabled biologists to access to huge amount of different types of data (genomics, epigenomics, etc.).

In parallel to the experimental efforts, many computational ncRNAs predictors have been proposed in last years. They are mainly designed to predict a specific type or family of ncRNAs and based on sequence characteristics. Nevertheless, few works have been proposed to predict all types of ncRNAs. However, a majority of methods and tools aim to separate ncRNAs from mRNA transcripts by exploiting many features. incRNA is an example of these methods proposed for *Caenorhabditis elegans* ncRNA prediction [76]. It combines a large amount of sequence, structure, and large-scale expression data to distinct coding from ncRNA sequences and potentially differentiates between various ncRNA types. Furthermore, Lertampaiporn et al. [77] proposed a random forest based classifier to identify ncRNAs. They have used various types of features, which can be divided into five categories: Sequence-based features, Secondary structure features, Base-pair features, Triplet sequence-structure, and Structural robustness features.

Recently, Brayet et al. proposed a new methodology and tool (called piRPred) for piRNA prediction based on the multiple kernel learning (MKL) approach [75]. piRPred is composed of three kernels implementing, respectively, the following heterogeneous features: (1) presence of Uridine ("U") at the first position of the

sequence and k-mer motif frequency, (2) occurrence into clusters, and (3) distance to centromeric and telomeric regions.

Another way to predict ncRNAs is based on their function, by the exploitation of the interaction information of ncRNA sequences with other targets in the cell; as an example, the Piano tool that uses the piRNA-transposon interaction information to accurately predict piRNAs [78].

Long ncRNAs (lncRNAs), generally longer than 200 nt, have received an increasing attention recently due to their potential role as new and crucial layer of biological regulation [79]. Some tools have been proposed for lncRNA prediction based on several features [80], such as the pipeline proposed by Sun et al. [81] and the iSeeRNA tool [82] proposed to predict long intergenic ncRNAs. Recently, Achawanantakun et al. have proposed LncRNA-ID, a tool for long ncRNA identification based on sequence, ribosome interaction, and protein conservation features and by using random forest model [83]. Lu et al. proposed PLEK tool [84] that is based on an improved k-mer scheme and an SVM designed mainly to distinguish lncRNAs from mRNAs in RNA-seq transcriptomes of species lacking reference genomes.

6 An Example of ncRNA: MicroRNA

6.1 Background

MicroRNAs, also called miRNAs, are an abundant class of small single-stranded noncoding RNAs of about 21–22 nt, found in animals, plants, and viruses, which play important roles in many biological processes through regulating gene expression either by translational inhibition or message degradation [85, 86]. The first miRNA, described in 1993, was *lin-4*, which is essential for the temporal control of development in the model organism *C. elegans* [87, 88]. Ambos's group identified its 61-nt precursor and the mature transcript of 22 nt, and then found that *lin-4* sequence did not encode a protein and was partially complementary to multiple sequences in the 3′-untranslated region (UTR) of *lin-14* gene [87]. Meanwhile, Ruvkun's group discovered that the synthesis of LIN-14 protein is regulated posttranscriptionally by *lin-4* [88]. However, the function as gene expression regulators of miRNAs was not generally recognized until the discovery of *let-7* in the early 2000s [89–93]. This miRNA was also found to be conserved in several species, including humans [90]. Follow-up research in the domain reported various miRNAs in different cell types and tissues, as well as their functions in development, cell differentiation, cell death, apoptosis, cell proliferation, metabolism, etc. [86, 94–101]. Over the last years, miRNAs have also been detected in a number of viruses, such as Epstein–Barr virus [102], Bovine herpesvirus1 [103], Herpes B virus [104], and others [105–107]. With the ever-increasing amount of data coming from new

sequencing technologies, miRNAs have been deposited abundantly in the miRBase database (www.mirbase.org) with 24,521 miRNA loci from 206 species, processed to produce 30,424 mature miRNA products in the latest release (v20, 2013), and continue to increase [108–112].

Generally, miRNA genes are transcribed by the RNA polymerase II and III as long primary miRNAs (pri-miRNAs) of several hundreds nt. The latter is subsequently cleaved by the Drosha RNase III endonuclease to generate ~70-nt precursor miRNAs (pre-miRNA), which are characterized by a stem-loop structure. The precursors, after being transported to the cytoplasm, are processed by the Dicer RNase III endonuclease to form mature miRNAs of 21–22 nt. These mature miRNAs then bind to the 3′-UTR mRNA complementary region, forming the RNA-induced silencing complex (RISC), which induces the degradation or inhibits the translation of the mRNA, resulting in gene silencing [86, 113].

The counterpart, or the complementary sequence, of the mature miRNA is denoted as miRNA* (star-miRNA). It is often degraded after release of the mature strand, and thus usually has much lower expression. This property is much exploited when identifying miRNA from sequencing data. In some cases, distinct mature miRNAs can be formed from the same precursor transcript [114, 115]. They can contain a small sequence shift or additional nucleotides (miRNA isoforms) [116], or nonoverlapping (miRNA siblings) [114].

Although both plant and animal miRNAs have diverse activities in development and physiology, they are substantially different in their primary modes of action [117]: (1) the loci producing miRNAs have distinct genomic arrangements; (2) mature miRNAs from precursors are generated by different pathways and in different subcellular compartments; (3) plant miRNAs and their mRNA targets have extensive complementarity, inducing gene repression through cleavage of the targets [97], whereas an animal miRNA might have many more RNA targets and a target might be regulated by multiple miRNAs as animal miRNAs can recognize their targets with the complementary at the miRNA seed of only 6–8 nt [118–120].

6.2 Secondary Structure

Pre-miRNA has a hairpin secondary structure, where the precursor sequence is folded in one or more stems, symmetric and asymmetric internal loops, bulges, and a terminal loop (Fig. 6). The longest exact stem usually has more than 5 bp [121]. Most structures are nearly symmetric as they possess few bulges and asymmetric internal loops or these are compensated between each other, resulting in the two pre-miRNA strands of similar length. The irregularity of the hairpin structure with mismatches gives rise to several potential structures with little difference when predicting the structure of

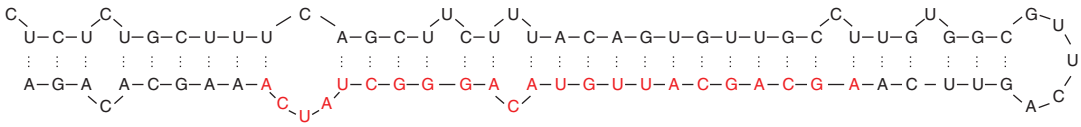


Fig. 6 Example of a pre-miRNA secondary structure (hsa-mir-107), with the mature miRNA shown in red

lowest free energy [10]. Several features based on sequence and secondary structure have been investigated with the aim of a more precise pre-miRNA identification: percentage mono-, di-, trinucleotides, asymmetry, size, position of substructures, and free energy [122].

6.3 MiRNA Prediction

Due to their great significance, the identification of miRNAs is of key importance for both biological and medical sciences. It is nevertheless difficult to detect them in cells with experimental techniques because of their small size and low abundance [91, 123]. With the help of computational approaches, one can identify the mature miRNAs and their mRNA targets from genomic sequences or deep sequencing data. A primary step, as designed by the majority of existing tools, is to predict the precursors thanks to their high conservation and stem-loop structure. Such a hairpin structure is however insufficient to correctly determine a precursor, as other ncRNAs may also possess a similar form. The prediction output would be subject to experimental validation afterward. The facility for obtaining high-throughput sequencing data nowadays allows for alternative predictions as well as deeper studies on the activities of miRNAs in the interaction with their targets.

Three main approaches have been exploited for pre-miRNA prediction from genomic sequences: comparative genomics, homology, and ab initio. In comparative genomics methods, genomes of related species are compared to detect conserved pre-miRNAs with the help of multiple sequence alignment. Such methods include miRSeeker [123], miRFinder [124], MiRScan [125], miRRim [126], RNAmicro [127], and SMIRP [128]. With a close concept, homology-based methods can predict new pre-miRNAs that are homologous to existing ones in terms of sequences and secondary structures, as proposed in ERPIN [129] and miRAlign [130]. Nevertheless, it is unlikely that these two approaches would be beneficial when a new sequence without a known homolog or cross-species sequence conservation is investigated.

The ab initio approaches, which may help to avoid the problem of homology and conservation, can be classified into three categories. The first one, considered completely ab initio, searches for pre-miRNAs occurring in given genomic sequences by means

of intrinsic properties of sequence and structure of pre-miRNAs, as in miRNAFold [121], CID-miRNA [131], and NOVOMIR [132]. The second one predicts pre-miRNAs with some additional information, for instance, positions or neighbors of candidates in genomic sequences in miR-abela [133] and MIReNA [134], or targets of candidates in Semirna [135]. These two categories can be applied as primitive filters for pre-miRNA candidates, which might be subsequently refined with other techniques. Following those primary filters, the third category will classify possible pre-miRNAs as real or pseudo ones. This classification problem could be efficiently solved with machine learning in general. Various machine learning techniques have then been introduced, such as Bayesian networks in BayesMiRNAfinder [136], genetic programming in miRPred [137], hidden Markov models in CSHMM [138], HHMMiR [139], and proMIR [140], kernel density estimator in mir-KDE [141], random forests in MiPred [69], random walk in miRank [142], and SVM in triplet-SVM [68], MaturePred [143], miPred [144], miRNA-deKmer [145], miRPara [146], RAmiRNA [147], and ViralmiR [148].

However, the predictive performance of machine learning methods depends much on training data, in which the negative data, or the pseudo pre-miRNAs, have a high impact. These can be abundantly determined via hairpin-like structures from exonic regions of protein-coding genes or from other noncoding RNAs (siRNA, snRNA, snoRNA, and tRNA) [122], yet this dataset could be inadequately representative. MiRNApre [149] and iMiRNA-SSF [150] tried to establish high-quality negative datasets to improve the identification of pre-miRNAs. Meanwhile, other methods were developed to deal with the imbalance in the training data, where the negative set is much larger than the positive one. This imbalance makes traditional learning-based classifiers, for instance standard SVM, unsuitable as these will tend to classify all given samples into the more prevalent class in the training data [151], thus pseudo pre-miRNA. Such methods include miRBoost [122], HeteroMirPred [152], HuntMi [153], microPred [154], MiRenSVM [155], and mirExplorer [156], in which miRBoost has recently shown its outperformance of the others with a good compromise between prediction accuracy and execution time, using a boosting algorithm on SVM classifiers.

Most of those tools were designed for different species altogether with common features of miRNAs. Besides, some made use of specific properties of miRNAs from a particular species, such as NOVOMIR [132], Semirna [135], and MaturePred [143] for plants, CID-miRNA [131] for human, and ViralmiR [148] for viruses.

References

1. Doty P, Boedtker H, Fresco JR et al (1959) Secondary structure in ribonucleic acids. Proc Natl Acad Sci U S A 45:482–499
2. Holley RW, Apgar J, Everett GA et al (1965) Structure of a ribonucleic acid. Science (New York, NY) 147:1462–1465
3. Lucks JB, Mortimer SA, Trapnell C et al (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci U S A 108:11063–11068
4. Leontis NB, Westhof E (1998) A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. J Mol Biol 283:571–583
5. Pleij CW, Rietveld K, Bosch L (1985) A new principle of RNA folding based on pseudoknotting. Nucleic Acids Res 13:1717–1731
6. Schimmel P (1989) RNA pseudoknots that interact with components of the translation apparatus. Cell 58:9–12
7. Mathews DH, Sabina J, Zuker M et al (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940
8. Nussinov R, Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A 77:6309–6313
9. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148
10. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415
11. Hofacker IL, Fontana W, Stadler PF et al (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie 125:167–188
12. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5:140
13. Dumas JP, Ninio J (1982) Efficient algorithms for folding and comparing nucleic acid sequences. Nucleic Acids Res 10:197–206
14. Martinez HM (1988) An RNA secondary structure workbench. Nucleic Acids Res 16:1789–1798
15. James BD, Olsen GJ, Pace NR (1989) Phylogenetic comparative analysis of RNA secondary structure. Methods Enzymol 180:227–239
16. Larsen N, Olsen GJ, Maidak BL et al (1993) The ribosomal database project. Nucleic Acids Res 21:3021–3023
17. Noller HF, Woese CR (1981) Secondary structure of 16S ribosomal RNA. Science 212:403–411
18. Noller HF (1984) Structure of ribosomal RNA. Annu RevBiochem 53:119–162
19. Gutell RR, Weiser B, Woese CR et al (1985) Comparative anatomy of 16-S-like ribosomal RNA. Prog Nucleic Acid Res Mol Biol 32:155–216
20. Han K, Kim HJ (1993) Prediction of common folding structures of homologous RNAs. Nucleic Acids Res 21:1251–1257
21. Grate L (1995) Automatic RNA secondary structure determination with stochastic context-free grammars. In: Proceedings of the third international conference on intelligent systems for molecular biology. AAAI Press, Cambridge, UK, 136–144
22. Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 31:3423–3428
23. Engelen S, Tahi F (2010) Tfold: efficient in silico prediction of non-coding RNA secondary structures. Nucleic Acids Res 38:2453–2466
24. Gorodkin J, Heyer LJ, Stormo GD (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res 25:3724–3732
25. Mathews DH, Turner DH (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317:191–203
26. Harmanci AO, Sharma G, Mathews DH (2008) PARTS: probabilistic alignment for RNA joinT secondary structure prediction. Nucleic Acids Res 36:2406–2417
27. Do CB, Foo C-S, Batzoglou S (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics 24:i68–i76
28. Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
29. Engelen S, Tahi F (2007) Predicting RNA secondary structure by the comparative approach: how to select the homologous sequences. BMC Bioinformatics 8:464

30. Shapiro BA, Wu JC (1997) Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm. Comput Appl Biosci 13:459–471
31. Rivas E, Eddy SR (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 285:2053–2068
32. Reeder J, Giegerich R (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5:104
33. Dawson W, Fujiwara K, Kawai G et al (2006) A method for finding optimal RNA secondary structures using a new entropy model (vsfold). Nucleosides Nucleotides Nucleic Acids 25:171–189
34. Poolsap U, Kato Y, Akutsu T (2009) Prediction of RNA secondary structure with pseudoknots using integer programming. BMC Bioinformatics 10(Suppl 1):S38
35. 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
36. Bellaousov S, Mathews DH (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA 16:1870–1880
37. Abrahams JP, van den Berg M, van Batenburg E et al (1990) Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. Nucleic Acids Res 18:3035–3044
38. Ruan J, Stormo GD, Zhang W (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics 20:58–66
39. Jabbari H, Condon A, Zhao S (2008) Novel and efficient RNA secondary structure prediction using hierarchical folding. J Comput Biol 15:139–163
40. Ren J, Rastegari B, Condon A et al (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. RNA 11:1494–1504
41. Sperschneider J, Datta A (2008) KnotSeeker: heuristic pseudoknot detection in long RNA sequences. RNA 14:630–640
42. Jabbari H, Condon A (2014) A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. BMC Bioinformatics 15:147
43. Hofacker IL, Fekete M, Flamm C et al (1998) Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res 26:3825–3836
44. Achawanantakun R, Sun Y (2013) Shape and secondary structure prediction for ncRNAs including pseudoknots based on linear SVM. BMC Bioinformatics 14(Suppl 2):S1
45. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244:48–52
46. Mathews DH, Turner DH, Zuker M (2007) RNA secondary structure prediction. In: Beaucage SL, Bergstrum DE, Glick GD et al (eds) Current protocols in nucleic acid chemistry. John Wiley & Sons, Inc., New York, pp. 11.2.1–11.2.17
47. Wuchty S, Fontana W, Hofacker IL et al (1999) Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49:145–165
48. Deigan KE, Li TW, Mathews DH et al (2009) Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A 106:97–102
49. Machado-Lima A, del Portillo HA, Durham AM (2008) Computational methods in noncoding RNA research. J Math Biol 56:15–49
50. Fontana W, Konings DA, Stadler PF et al (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404
51. Shapiro BA, Zhang KZ (1990) Comparing multiple RNA secondary structures using tree comparisons. Comput Appl Biosci 6:309–318
52. Shapiro BA (1988) An algorithm for comparing multiple RNA secondary structures. Comput Appl Biosci 4:387–393
53. Herrbach C (2007) Etude algorithmique et statistique de la comparaison des structures secondaires d'ARN.http://www.theses.fr/2007BOR13432
54. Blin G, Denise A, Dulucq S et al (2010) Alignments of RNA structures. IEEE/ACM Trans Comput Biol Bioinform 7:309–322
55. Höchsmann M, Töller T, Giegerich R et al (2003) Local similarity in RNA secondary structures. In: Proceedings of the IEEE computer society bioinformatics conference. IEEE, Stanford: 159–168
56. LinG-H, MaB, and ZhangK (2001) Edit distance between two RNA structures. In:Proceedings of the fifth annual international conference on computational biology. ACM, New York, pp211–220
57. Allali J, Sagot M-F (2005) A new distance for high level RNA secondary structure comparison. IEEE/ACM Trans Comput Biol Bioinform 2:3–14
58. Agius P, Bennett KP, Zuker M (2010) Comparing RNA secondary structures using a relaxed base-pair score. RNA 16:865–878

59. Dulucq S, Tichit L (2003) RNA secondary structure comparison: exact analysis of the Zhang–Shasha tree edit algorithm. Theor Comput Sci 306:471–484
60. Zhang K, Shasha D (1989) Simple fast algorithms for the editing distance between trees and related problems. SIAM J Comput 18:1245–1262
61. Smith C, Heyne S, Richter AS et al (2010) Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. Nucleic Acids Res 38:W373–W377
62. Siebert S, Backofen R (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics 21:3352–3359
63. Höchsmann M, Voss B, Giegerich R (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM Trans Comput Biol Bioinform 1:53–62
64. Ji Y, Xu X, Stormo GD (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. Bioinformatics 20:1591–1602
65. Rivas E, Klein RJ, Jones TA et al (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol 11:1369–1373
66. Fu Y, Xu ZZ, Lu ZJ et al (2015) Discovery of novel ncRNA sequences in multiple genome alignments on the basis of conserved and stable secondary structures. PLoS One 10:e0130200
67. Xu Z, Mathews DH (2011) Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. Bioinformatics 27:626–632
68. Xue C, Li F, He T et al (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 6:310
69. Jiang P, Wu H, Wang W et al (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 35:W339–W344
70. Karklin Y, Meraz RF, Holbrook SR (2005) Classification of non-coding RNA using graph representations of secondary structure. Pac Symp Biocomput 2005:4–15
71. Childs L, Nikoloski Z, May P et al (2009) Identification and classification of ncRNA molecules using graph properties. Nucleic Acids Res 37:e66
72. Nawrocki EP, Burge SW, Bateman A et al (2015) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43:D130–D137
73. Panwar B, Arora A, Raghava GPS (2014) Prediction and classification of ncRNAs using structural information. BMC Genomics 15:127
74. Rivas E, Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16:583–605
75. Brayet J, Zehraoui F, Jeanson-Leh L et al (2014) Towards a piRNA prediction using multiple kernel fusion and support vector machine. Bioinformatics 30:i364–i370
76. Lu ZJ, Yip KY, Wang G et al (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res 21:276–285
77. Lertampaiporn S, Thammarongtham C, Nukoolkit C et al (2014) Identification of non-coding RNAs with a new composite feature in the hybrid random forest ensemble algorithm. Nucleic Acids Res 42:e93
78. Wang K, Liang C, Liu J et al (2014) Prediction of piRNAs using transposon interaction and a support vector machine. BMC Bioinformatics 15:419
79. Kung JTY, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193:651–669
80. Iwakiri J, Hamada M, Asai K (2016) Bioinformatics tools for lncRNA research. Biochim Biophys Acta 1859:23–30
81. Sun L, Zhang Z, Bailey TL et al (2012) Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. BMC Bioinformatics 13:331
82. Sun K, Chen X, Jiang P et al (2013) iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics 14(Suppl 2):S7
83. Achawanantakun R, Chen J, Sun Y et al (2015) LncRNA-ID: long non-coding RNA identification using balanced random forests. Bioinformatics 31:3897–3905
84. Li A, Zhang J, Zhou Z (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics 15:311
85. Ambros V (2004) The functions of animal microRNAs. Nature 431:350–355
86. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297

87. Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854
88. Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 75:855–862
89. Reinhart BJ, Slack FJ, Basson M et al (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403:901–906
90. Pasquinelli AE, Reinhart BJ, Slack F et al (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408:86–89
91. Lagos-Quintana M, Rauhut R, Lendeckel W et al (2001) Identification of novel genes coding for small expressed RNAs. Science 294:853–858
92. Lau NC, Lim LP, Weinstein EG et al (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294:858–862
93. Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. Science 294:862–864
94. Lagos-Quintana M, Rauhut R, Yalcin A et al (2002) Identification of tissue-specific microRNAs from mouse. Curr Biol 12:735–739
95. Brennecke J, Hipfner DR, Stark A et al (2003) Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell 113:25–36
96. Wienholds E, Kloosterman WP, Miska E et al (2005) MicroRNA expression in zebrafish embryonic development. Science 309:310–311
97. Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol 57:19–53
98. Cuellar TL, McManus MT (2005) MicroRNAs and endocrine biology. J Endocrinol 187:327–332
99. Poy MN, Eliasson L, Krutzfeldt J et al (2004) A pancreatic islet-specific microRNA regulates insulin secretion. Nature 432:226–230
100. Chen C-Z, Li L, Lodish HF et al (2004) MicroRNAs modulate hematopoietic lineage differentiation. Science 303:83–86
101. Wilfred BR, Wang W-X, Nelson PT (2007) Energizing miRNA research: a review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways. Mol Genet Metab 91:209–217
102. Pfeffer S, Zavolan M, Grässer FA et al (2004) Identification of virus-encoded microRNAs. Science 304:734–736
103. Glazov EA, Horwood PF, Assavalapsakul W et al (2010) Characterization of microRNAs encoded by the bovine herpesvirus 1 genome. J Gen Virol 91:32–41
104. Besecker MI, Harden ME, Li G et al (2009) Discovery of herpes B virus-encoded microRNAs. J Virol 83:3413–3416
105. Gottwein E (2013) Roles of microRNAs in the life cycles of mammalian viruses. Curr Top Microbiol Immunol 371:201–227
106. Li S-C, Shiau C-K, Lin W (2008) Vir-Mir db: prediction of viral microRNA candidate hairpins. Nucleic Acids Res 36:D184–D189
107. Qureshi A, Thakur N, Monga I et al (2014) VIRmiRNA: a comprehensive resource for experimentally validated viral miRNAs and their targets. Database 2014:bau103
108. Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32:D109–D111
109. Griffiths-Jones S, Grocock RJ, van Dongen S et al (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–D144
110. Griffiths-Jones S, Saini HK, van Dongen S et al (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158
111. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157
112. Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68–D73
113. Lee Y, Jeon K, Lee J-T et al (2002) MicroRNA maturation: stepwise processing and subcellular localization. EMBO J 21:4663–4670
114. Zhang W, Gao S, Zhou X et al (2010) Multiple distinct small RNAs originate from the same microRNA precursors. Genome Biol 11:R81
115. Vazquez F, Blevins T, Ailhas J et al (2008) Evolution of Arabidopsis MIR genes generates novel microRNA classes. Nucleic Acids Res 36:6429–6438
116. Ebhardt HA, Fedynak A, Fahlman RP (2010) Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity. Silence 1:12
117. Axtell MJ, Westholm JO, Lai EC (2011) Vive la différence: biogenesis and evolution of microRNAs in plants and animals. Genome Biol 12:221

118. Krek A, Grün D, Poy MN et al (2005) Combinatorial microRNA target predictions. Nat Genet 37:495–500
119. Rajewsky N (2006) MicroRNA target predictions in animals. Nat Genet 38:S8–S13
120. Friedman RC, Farh KK-H, Burge CB et al (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19:92–105
121. Tempel S, Tahi F (2012) A fast ab-initio method for predicting miRNA precursors in genomes. Nucleic Acids Res 40:e80
122. Tran VDT, Tempel S, Zerath B et al (2015) miRBoost: boosting support vector machines for microRNA precursor classification. RNA 21:775–785
123. Lai EC, Tomancak P, Williams RW et al (2003) Computational identification of drosophila microRNA genes. Genome Biol 4:R42
124. Huang T-H, Fan B, Rothschild MF et al (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics 8:341
125. Lim LP, Lau NC, Weinstein EG et al (2003) The microRNAs of caenorhabditis elegans. Genes Dev 17:991–1008
126. Terai G, Komori T, Asai K et al (2007) miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. RNA 13:2081–2090
127. Hertel J, Stadler PF (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics 22:e197–e202
128. Peace RJ, Biggar KK, Storey KB et al (2015) A framework for improving microRNA prediction in non-human genomes. Nucleic Acids Res 43:e138
129. Legendre M, Lambert A, Gautheret D (2005) Profile-based detection of microRNA precursors in animal genomes. Bioinformatics 21:841–845
130. Wang X, Zhang J, Li F et al (2005) MicroRNA identification based on sequence and structure alignment. Bioinformatics 21:3610–3614
131. Tyagi S, Vaz C, Gupta V et al (2008) CID-miRNA: a web server for prediction of novel miRNA precursors in human genome. Biochem Biophys Res Commun 372:831–834
132. Teune J-H, Steger G (2010) NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome. J Nucleic Acids 2010:495904
133. Sewer A, Paul N, Landgraf P et al (2005) Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 6:267
134. Mathelier A, Carbone A (2010) MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics 26:2226–2234
135. Muñoz-Mérida A, Perkins JR, Viguera E et al (2012) Semirna: searching for plant miRNAs using target sequences. OMICS 16:168–177
136. Yousef M, Nebozhyn M, Shatkay H et al (2006) Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier. Bioinformatics 22:1325–1334
137. Brameier M, Wiuf C (2007) Ab initio identification of human microRNAs based on structure motifs. BMC Bioinformatics 8:478
138. Agarwal S, Vaz C, Bhattacharya A et al (2010) Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). BMC Bioinformatics 11(Suppl 1):S29
139. Kadri S, Hinman V, Benos PV (2009) HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC Bioinformatics 10(Suppl 1):S35
140. Nam J-W, Kim J, Kim S-K et al (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res 34:W455–W458
141. Chang DT-H, Wang C-C, Chen J-W (2008) Using a kernel density estimation based classifier to predict species-specific microRNA precursors. BMC Bioinformatics 9(Suppl 12):S2
142. Xu Y, Zhou X, Zhang W (2008) MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics 24:i50–i58
143. Xuan P, Guo M, Huang Y et al (2011) MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. PLoS One 6:e27422
144. Ng KLS, Mishra SK (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23:1321–1330
145. Liu B, Fang L, Wang S et al (2015) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J Theor Biol 385:153–159
146. Wu Y, Wei B, Liu H et al (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. BMC Bioinformatics 12:107
147. Tyagi V, Prasad CS (2012) RAmiRNA: software suite for generation of SVMbased pre-

diction models of mature miRNAs. Bioinformation 8:581–585

148. Huang K-Y, Lee T-Y, Teng Y-C et al (2015) ViralmiR: a support-vector-machine-based method for predicting viral microRNA precursors. BMC Bioinformatics 16(Suppl 1):S9
149. Wei L, Liao M, Gao Y et al (2014) Improved and promising identification of human microRNAs by incorporating a high-quality negative set. IEEE/ACM Trans Comput Biol Bioinform 11:192–201
150. Chen J, Wang X, Liu B (2016) iMiRNA-SSF: improving the identification of microRNA precursors by combining negative sets with different distributions. Sci Rep 6:19062
151. Wu G, Chang EY (2003) Class-boundary alignment for imbalanced dataset learning. In: ICML workshop on learning from imbalanced data sets. AAAI Press, Washington DC, 49–56
152. Lertampaiporn S, Thammarongtham C, Nukoolkit C et al (2013) Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. Nucleic Acids Res 41:e21
153. Gudyś A, Szcześniak MW, Sikora M et al (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinformatics 14:83
154. Batuwita R, Palade V (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics 25:989–995
155. Ding J, Zhou S, Guan J (2010) MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics 11(Suppl 11):S11
156. Guan D-G, Liao J-Y, Qu Z-H et al (2011) mirExplorer: detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. RNA Biol 8:922–934

Chapter 8

Computational Prediction of RNA-Protein Interactions

Carla M. Mann, Usha K. Muppirala, and Drena Dobbs

Abstract

Experimental methods for identifying protein(s) bound by a specific promoter-associated RNA (paRNA) of interest can be expensive, difficult, and time-consuming. This chapter describes a general computational framework for identifying potential binding partners in RNA-protein complexes or RNA-protein interaction networks. Protocols for using three web-based tools to predict RNA-protein interaction partners are outlined. Also, tables listing additional webservers and software tools for predicting RNA-protein interactions, as well as databases that contain valuable information about known RNA-protein complexes and recognition sites for RNA-binding proteins, are provided. Although only one of the tools described, lncPro, was designed expressly to identify proteins that bind long noncoding RNAs (including paRNAs), all three approaches can be applied to predict potential binding partners for both coding and noncoding RNAs (ncRNAs).

Key words RNA-protein interactions, Computational prediction, RPISeq, catRAPID, lncPRO, RNA-protein databases, Machine learning, ncRNA

1 Introduction

Our understanding of RNA-protein interactions has advanced dramatically over the past decade due to exciting advances in experimental technologies for identifying binding partners in RNA-protein complexes and RNA-protein interaction networks [1, 2]. These include high-throughput CHIP and RNASeq-based methods that can identify RNAs bound by specific proteins in vivo [3–5], methods that can identify RNA binding proteins, their target RNAs, and their RNA binding sites on a genome-wide scale [6–9], and integrated biochemical and bioinformatics approaches that can identify the specific recognition sequences for RNA binding proteins [10]. A major motivation for these studies has been the search for cellular and molecular functions for noncoding RNAs (ncRNAs), many of which have been shown to play important roles in disease as well as in normal development [11]. In particular, promoter-associated RNAs (paRNAs), the focus of this volume, not only regulate transcription [12], but also serve as

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_8,

epigenetic modulators that affect cellular differentiation [13] (e.g., protein localization [14], and gene regulation [15]). Genetic regulation by paRNAs or other ncRNAs is often mediated through interactions of the RNA with specific RNA binding proteins; thus, identifying the binding partner of a newly discovered paRNA (or any ncRNA) can provide important clues to its function [16].

Despite the technical advances mentioned above, the experimental time, effort, and expense required to identify biologically relevant protein binding partners for a specific RNA (or vice versa) have created a demand for computational methods that can predict the most likely binding partners in RNA-protein complexes and/or identify novel candidate interaction partners in RNA-protein interaction networks. The goal of the chapter is to provide step-by-step protocols to assist molecular biologists and other experts in accessing and utilizing available computational resources that provide access to existing information about specific RNA-protein interactions, as well as software for predicting potential RNA-protein binding partners when experimental information is not available. For additional background and details regarding these and other published approaches, we refer the reader to recent reviews [17, 18]. The methods outlined in this chapter are generally applicable to any RNA, coding or noncoding, small or large; thus, they can be valuable for quickly identifying potential protein binding partners for any specific paRNA.

In this chapter, we focus on currently available *web-based* computational tools for *partner* prediction, i.e., predicting which protein binds to a specific RNA of interest in an RNA-protein complex or RNA-protein interaction network. Several available tools are also capable of predicting the converse, i.e., which RNA(s) bind to a specific protein of interest. Software and servers for *interface* prediction, i.e., predicting which specific amino acid residues and/or ribonucleotides are involved in recognition and binding are not described here, but have been reviewed elsewhere [19–21].

The protocol involves two major **steps 1** and **2** (illustrated in Fig. 1):

Step 1: Determine whether experimental data regarding the binding partner(s) of the query ncRNA or putative RNA-binding protein are already available. This step is described in Methods, Subheading 3.1, which outlines strategies for exploiting available online databases and servers (provided in Table 1 below) that focus on ncRNA or RNA-protein interactions, or provide sequence and/or structural data regarding RNA-protein complexes.

Step 2: If known or potential interaction partners cannot be identified using available resources, or if the user wishes to identify additional potential binding partners, use one (or preferably, all three) of the following web-based tools for predicting RNA-protein interaction partners:

- **RPISeq** (Subheading 3.2)—a machine learning-based approach developed by our group [30], which requires only

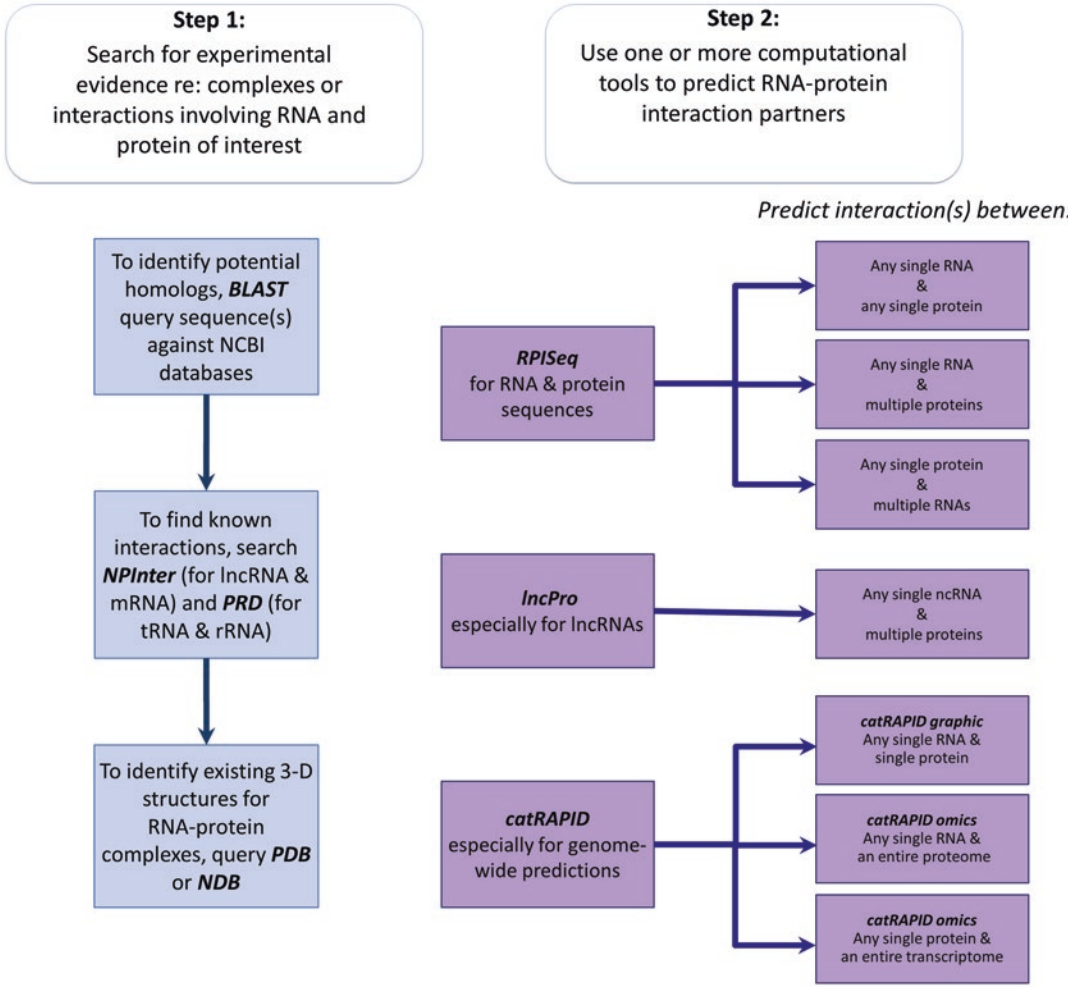


Fig. 1 Flowchart for identifying potential RNA-protein interaction partners

sequence information for the RNA(s) and protein(s) of interest. This method was not specifically designed for predicting partners of promoter-associated RNAs (or ncRNAs), but it can readily predict these interactions.

- **lncPro** (Subheading 3.3)—a method developed by the group of Tingting Li [31], specifically for predicting the likelihood that a specific long noncoding RNA (lncRNA) interacts with one or more candidate protein sequences.
- **catRAPID** (Subheading 3.4)—a suite of programs developed by the group of Gian Gaetano Tartaglia, including algorithms for estimating the binding propensity for individual RNA-protein pairs (catRAPID graphic) [32] and a server for large-scale interactome predictions, e.g., for the interaction of a single RNA with an entire proteome or a single protein with an entire transcriptome (catRAPID omics) [33].

Table 1
Databases of RNA-protein interactions and interfaces

Database	URL	Description
BioGRID [22]	http://thebiogrid.org/	Manually curated protein and genetic interactions for major model organisms
IntAct [23]	http://www.ebi.ac.uk/intact/	Manually curated molecular interactions, including comprehensive data about their source experiments
NDB [24, 25]	http://ndbserver.rutgers.edu/	Nucleic acid and DNA/RNA-protein complex structures, including derived data for nucleic acids
NPInter [26]	http://www.bioinfo.org/NPInter/	Functional interactions of ncRNAs and protein-related biomolecules, classified into categories based on interaction type
PDB [27]	http://www.rcsb.org/pdb/home/home.do	Experimentally determined three-dimensional structures
PRD [28]	http://pri.hgc.jp/	RPIs from 22 species, focusing on gene-level information
RBPDB [29]	http://rbpdb.ccbr.utoronto.ca/	Experimental data on binding preferences and specificities of RBPs
RPIntDB	http://pridb.gdcb.iastate.edu/RPISeq/RPIntDB.html/	RPIs from databases and high-throughput experiments in literature

Note: PRIDB [35], which was developed by our group, is not included in this list of recommended databases because it is no longer maintained

The user is strongly encouraged to try all three web-based tools because the underlying algorithms and datasets used for training and performance evaluation are different in each case. Direct performance comparisons of the methods on various benchmark datasets indicate that the methods have different strengths and weaknesses [17, 18, 34]. The user should, of course, interpret all prediction results with caution: although each of these tools has been shown to perform well "on average" in predicting RNA-protein interactions, a highly accurate prediction for any given RNA-protein pair cannot be guaranteed.

2 Materials

2.1 Databases of Experimentally Validated RNA-Protein Complexes and Interactions

Before making computational predictions, the user is advised to search for existing experimental evidence regarding the specific RNA-protein interaction of interest, both in published literature and in relevant specialized databases. At present, the availability of information regarding validated RNA-protein interactions is increasing rapidly as new experimental data are incorporated into web-based resources. These include databases containing evidence

for physical or genetic interactions obtained from both "low" and "high" throughput experiments (e.g., NPInter or PRD), as well as databases of high resolution structural information regarding both the components and the interfaces in RNA-protein complexes (e.g., PDB or NDB). Table 1 provides an alphabetical listing of several valuable databases that contain information about RNA-protein complexes and interaction networks. A suggested strategy for utilizing selected resources from this list is provided in Subheading 3.1.

2.2 Servers and Software for Predicting RNA-Protein Partners

At present, there are only a few published methods for predicting partners in RNA-protein complexes or interaction networks. Subheadings 3.2–3.4 focus on three published methods (RPISeq, catRAPID, lncPro) that are freely available on web-based servers. Table 2 below includes several additional methods for which web servers and software may not be available.

3 Methods

3.1 Search the Literature and Databases for Existing Experimental Evidence

Before using computational approaches to predict potential interactions, the user should search published literature and existing databases for experimental evidence regarding interactions involving the RNA or protein of interest (*see* **Note 1**). If the sequence of interest corresponds to a protein or RNA of unknown function, potential homologs can be identified via a BLAST search. As outlined below, both the original query sequence and its homologs can be used to search databases of known RNA-protein interactions, such as those listed in Table 1.

1. Run the sequence or sequences through **NCBI's BLAST server**, available at http://blast.ncbi.nlm.nih.gov/Blast.cgi [39], or use similar genomics resources elsewhere (*see* **Note 2**). BLAST, or Basic Local Alignment Search Tool, finds highly similar sequences in the NCBI or ENSEMBL databases. If the query sequence has been previously identified and/or analyzed, clues to its involvement in specific RNA-protein interactions may be found in the NCBI "Gene" or "Protein" pages corresponding to the sequence (*see* **Note 3**). If the query sequence itself is not available in one of the NCBI databases, potential homologs identified by BLAST can be used as a starting point for subsequent searches in the databases listed in **steps 2–5** below.
2. Query **NPInter**, available at http://www.bioinfo.org/NPInter/index.htm [26]. NPInter v3.0 is the largest curated database of experimentally validated biomolecular interactions involving ncRNAs extracted from the literature. NPInter currently contains more than 900,000 ncRNA interactions, including interactions with proteins, as well as with DNA and both ncRNAs and mRNAs. Note, however, that tRNA and

Table 2
Computational methods for predicting RNA-protein interaction partners

Method	Training dataset	Features	Description
lncPro Lu et al [31]	322 interacting and 327 non-interacting pairs of RNA-protein chains from 18 RNA-protein complexes from PDB	Physicochemical properties consisting of RNA secondary structure propensities, hydrogen-bonding propensities, and van der Waals interaction propensities	Propensities are calculated for the protein and RNA sequence and fed through a complex equation to generate a probability score ranging from 0 to 100 (http://bioinfo.bjmu.edu.cn/lncpro/)
catRAPID Bellucci et al [32]	410 interacting pairs from 858 RNA-protein complexes from PDB	Physicochemical properties including secondary structure propensities, hydrogen-bonding propensities, and van der Waals interaction propensities	Propensities are calculated for each amino acid and ribonucleotide to generate an interaction profile (http://service.tartaglialab.com/page/catrapid_group)
RPISeq Muppirala et al [30]	2241 interacting pairs from 943 RNA-protein complexes from PRIDB (RPI2241)	Sequence composition of proteins, represented as conjoint triads, and RNAs, represented as tetrads	Protein and RNA sequences encoded sequence-composition-based features are used to train SVM and RF classifiers (http://pridb.gdcb.iastate.edu/RPISeq)
Wang et al [36]	RPI 2241 generated by Muppirala et al. and 367 interacting pairs from NPInter	Sequence composition of protein and RNA	Input to NB and ENB classifiers is a combination of protein triads and RNA triad features similar to those used in RPISeq
Pancaldi and Bähler [37]	5166 mRNA-protein interacting pairs from immunopurification experiments	Predicted protein secondary structure, localization, protein physical properties, gene physical properties, UTR properties, genetic interactions	Protein and RNA sequences encoded using >100 features are used to train SVM and RF classifiers
RPI-Pred [38]	1807 interacting and 1436 non-interacting RNA-protein chain pairs from PDB	1D protein block representation of predicted or actual 3D structure of RNA and protein combined with RNA and protein sequence	Protein and RNA sequences encoded with 112 protein and 20 RNA vectors are used to train SVM classifier (http://ctsb.is.wfubmc.edu/projects/rpi-pred/)

rRNA interactions are excluded (*see* **Note 4**). NPInter includes data from 22 different organisms and provides an integrated USCS Genome Browser to assist the user in locating binding sites in the human, mouse, and yeast genomes. The database is searchable by molecule name, molecule type, or database ID and provides access to software and servers, including **lncPro** (described in Subheading 3.3 below) for predicting lncRNA-protein interactions and the **iRNA** server for predicting RNA-RNA interactions (*see* **Note 5**).

3. Query the **Protein-RNA Interaction Database (PRD)**, available at http://pri.hgc.jp/ [28]. The PRD is smaller than NPInter, containing 10,817 experimentally validated RNA-protein interactions, but is valuable because it includes both rRNA and tRNA interactions, which are excluded from NPInter. The database offers flexible keyword searches (*see* **Note 6**).
4. In every case, the user should search the **Protein Data Bank (PDB)**, available at www.rcsb.org [27]. The PDB contains over 1600 three-dimensional structures of RNA-protein complexes determined using experiments such as X-ray crystallography, nuclear magnetic resonance (NMR) imaging, and cryo-electron microscopy. The PDB has a powerful search engine that allows the database to be queried in a variety of ways, e.g., by RNA or protein name, sequence, or GO terms. The PDB also provides excellent structure visualization tools as well as links to valuable third-party resources for visualizing and analyzing the structures of macromolecules (*see* **Note 7**).
5. In the same vein, the **Nucleic Acid Database (NDB)**, available at http://ndbserver.rutgers.edu [24, 25], is another valuable resource that focuses on experimentally determined three-dimensional structures of nucleic acids, including DNA-protein and RNA-protein complexes. Although the NDB contains only a subset of structures in the PDB, NDB makes it easier for the user to focus on structures that contain RNA-RNA, RNA-protein, or RNA-drug interactions. Also, the NDB provides convenient access to a wide variety of tools and software specifically designed for analyzing RNA sequences and structures (*see* **Note 8**).

3.2 RPISeq–to Predict Binding Partner(s) for Any Known RNA or Protein Sequence

The RPISeq webserver implements the RPISeq method developed by Muppirala et al [30]. RPISeq uses two types of machine learning classifiers, Random Forest (RF) and Support Vector Machine (SVM), to predict RNA-protein interactions using only sequence information. RPISeq can be used to predict the interaction probabilities of any type of RNA (not just ncRNAs) with any protein(s) of known sequence (*see* **Note 9**).

1. Access the RPISeq Webserver available at http://pridb.gdcb.iastate.edu/RPISeq/.

A stand-alone version of RPISeq is also freely available (*see* **Note 10**).

2. For **single RNA–single protein predictions**: The simplest function of the RPISeq server is to predict whether a specific known RNA interacts with a specific known protein. In this case, the user must enter the protein and RNA sequences (in FASTA format) in the appropriate text boxes on the homepage and click "Submit."
3. The RPISeq results, which are typically returned a few seconds after submission, include a display of the submitted input sequences along with the interaction probabilities predicted using both the RF and SVM classifiers. A probability greater than 0.50 is usually considered to be a positive prediction, although more stringent thresholds can be chosen.
4. For **single RNA–multiple protein predictions**: To predict the interaction probabilities for a single RNA with multiple potential protein partners, go to http://pridb.gdcb.iastate.edu/RPISeq/batch-rna.html.
5. Enter the RNA sequence and click "Choose File" to upload a file of protein sequences in FASTA format (limited to 100 Mb) (*see* **Note 10**). Click "Submit."
6. The results are returned as a table listing the interaction probabilities for the input RNA with each protein in the FASTA input file. Probabilities greater than 0.50 are usually considered to be positive predictions. The results may take several minutes to display, depending on the number of protein sequences submitted.
7. For **multiple RNA–single protein predictions**: To predict the interaction probabilities for a single protein with multiple potential RNA partners, go to http://pridb.gdcb.iastate.edu/RPISeq/batch-prot.html.
8. Enter the protein sequence in the text box and click "Choose File" to upload a file of RNA sequences in FASTA format (limited to 100 Mb) (*see* **Note 10**). Click "Submit."
9. The results are returned as a table listing the interaction probabilities for the input protein with each RNA sequence in the FASTA input file. Again, probabilities greater than 0.50 are usually considered to be positive predictions. The results may take several minutes to display if a large number of RNAs were submitted.

3.3 lncPro- to Predict Protein Binding Partner(s) for Any Known lncRNA

The lncPro webserver implements the lncPro method developed by Lu et al. [31]. lncPro is designed to predict whether a specific long noncoding RNA(lncRNA) interacts with one or more user-provided protein sequences. The method uses the hydrogen bonding and van der Waals propensities of the RNA and protein

sequences, in addition to the predicted secondary structure of the RNA, to calculate the probability that a specific lncRNA and protein will interact with one another (*see* **Note 11**).

1. Access the lncPro server homepage at bioinfo.bjmu.edu.cn/lncpro/.

   A stand-alone version of lncPro is also freely available for download (*see* **Note 12**).
2. For **single RNA—multiple protein predictions**: On the lncPro homepage, click on the "Predict" tab. The lncPro server takes as input a single RNA sequence in FASTA format and a file of multiple protein sequences in FASTA format.
3. Enter the RNA sequence and click "Choose File" to upload a file of protein sequences in FASTA format. Click "Submit."
4. The results are returned as a table listing the interaction probabilities for the input lncRNA sequence with each protein sequence in the FASTA input file. Probabilities greater than 0.50 are usually considered to be positive predictions.

3.4 catRAPID—to Predict Either Individual or Transcriptome/Proteome Wide Interactions

The catRAPID suite of RNA-protein interaction predictors includes methods for predicting the interaction propensity for individual RNA and protein partners (catRAPID graphic) [32]); identifying segments of RNA and protein that most likely participate in forming the RNP interface (catRAPID fragments); estimating the interaction strength of an RNA-protein pair in comparison to a reference set (catRAPID strength); and identifying the most probable interactions between a specific protein and a complete transcriptome, or between a specific RNA and a complete proteome, for eight model organisms (catRAPID omics [33]). Additional catRAPID modules can predict pairs of coexpressed proteins and RNAs in human tissues (catRAPID express [40]) and whether a protein is likely to bind RNA (catRAPID signature [41]) (*see* **Note 13**).

1. The suite website is located at http://service.tartaglialab.com/page/catrapid_group; extensive documentation and tutorials for all tools in the catRAPID suite are provided at: http://service.tartaglialab.com/static_files/shared/tutorial.html.
2. For **single RNA–single protein predictions**: Navigate to the catRAPID group page at http://service.tartaglialab.com/page/catrapid_group and select "catRAPID graphic."
3. Enter the protein and RNA sequences (in FASTA format) in the text boxes provided. Note: The input protein sequence must be between 50 and 750 amino acids in length; the input RNA sequence must be between 50 and 1200 ribonucleotides in length.

4. If desired, enter a name in the optional "Submission label" box and enter your email address in the optional "Email address" box for notification when results are available. Click "Submit."
5. The results are returned as a graphical "heat map" representing the interaction score for individual amino acid and ribonucleotide pairs; this interaction score is derived from the interaction propensity, which is also reported. The interaction scores range from −3 to +3, with higher values (red) corresponding to a higher probability of interaction. catRAPID graphic also generates a discriminative power (DP) score, which is a confidence metric; a DP score greater than 50 %, coupled with a positive interaction score, indicates that an interaction is likely to occur. A DP score greater than 75 % is a very high-confidence prediction. The results page also provides the server-generated ID for the job, the date and timestamp for the run, links for downloading the protein and RNA sequences submitted by the user, and the interaction heat map in .png format.
6. For **single RNA–proteome predictions**: Navigate to the catRAPID group page at http://service.tartaglialab.com/page/catrapid_group and select "catRAPID omics."
7. Select the second option: "catRAPID omics [transcript vs. nucleotide-binding proteome]" to open the submission page. The first text box is an optional box for a submission label, which should include the RNA name or other identifiers for easy identification of results. Enter the query RNA sequence (which must be longer than 50 ribonucleotides) in FASTA format.
8. The user is then presented with nine radial buttons under the heading: "Which library would you like to analyze?" Select the proteome of the model organism of interest—ideally, the source organism for the RNA query sequence (or its closest relative)—from the eight organism libraries available (*see* **Note 14**). The user can also create a custom proteome library (*see* **Note 15**).
9. The user is then asked whether she/he would like to use nucleic acid binding domains. The default "No" option will query all proteins in the proteome (<750 amino acids long), regardless of whether they possess a recognized RNA-binding domain. The "Yes" option will specifically examine and make predictions between the query RNA and proteins in the selected proteome that possess an RNA-binding domain.
10. Select whether predictions should be made against RNA-binding proteins only or against both RNA- and DNA-binding proteins, and whether disordered proteins should be included in the prediction; the latter is recommended because disordered protein regions frequently bind RNA.

11. A valid email address can be entered into the next text box for notification of when the job is completed. The user should then click "Submit query."
12. catRAPID omics results may take a few hours to be generated, depending on the size of the selected proteome. The results page contains a section summarizing the input parameters and a pie chart showing the distribution of rankings (on a scale of 1–3) of the possible RNA-protein interactions. Red indicates a likely interaction; orange indicates a moderately likely interaction, and yellow indicates an unlikely interaction. The next section consists of a 9-column table (*see* Fig. 2), in which the interactions are listed in order of highest to lowest scores (first column). The "Ranking" displayed in the last column is a metric of the probability of interaction: three stars indicates a strong interaction probability; 0 stars indicates an unlikely interaction. Protein-RNA pairs with both a high star ranking (3) and high discriminative power (>75 %) are predicted to interact. High star rankings with low discriminative power correspond to low-confidence predictions; as an example, note that in Fig. 2, the highest-scoring interaction has a very low discriminative power and is therefore considered to be unreliable (*see* **Note 16**).
13. For **single-protein—transcriptome predictions**: Navigate to the catRAPID group page at http://service.tartaglialab.com/page/catrapid_group and select "catRAPID omics."
14. On the subsequent page, select the first option: "catRAPID omics [transcript vs. coding and noncoding transcriptome]," which will open a submission page. The first text box is an optional box for a submission label, which should include the query protein name for easy identification of results. Enter the query protein sequence (which must be longer than 50 amino acids) in FASTA format.

#	Protein ID	RNA ID	Z-score?	Discriminative Power (%)?	Interaction Strength (%)?	Domain?	Motif?	Ranking?
1	PCBP3_MOUSE	RepA_1	-0.98	10	3	yes	yes	★★★
2	SKI_MOUSE	RepA_1	1.93	99	98	yes	no	★★★
3	SUZ12_MOUSE	RepA_1	1.56	99	99	yes	no	★★★
4	HDX_MOUSE	RepA_1	1.46	98	96	yes	no	★★★
5	SSRP1_MOUSE	RepA_1	1.38	98	93	yes	no	★★★
6	RN5A_MOUSE	RepA_1	1.28	98	94	yes	no	★★★
7	RTF1_MOUSE	RepA_1	1.28	98	91	yes	no	★★★
8	RFX2_MOUSE	RepA_1	1.25	98	91	yes	no	★★★
9	ZMAT1_MOUSE	RepA_1	1.15	98	89	yes	no	★★★
10	THOC1_MOUSE	RepA_1	1.12	97	97	yes	no	★★★

Fig. 2 Example of catRAPID omics results for the Xist repeat A region in *Mus musculus*

15. The user is then presented with nine radial buttons under the heading "Which library would you like to analyze?" Select the transcriptome of the model organism of interest, or its closest relative. The user can generate a custom transcriptome library, if desired (*see* **Note 15**).
16. The user is then asked whether she/he would like to use nucleic acid binding domains. If "No" is selected, the query protein must be <750 amino acids in length. In this case, the method will utilize the whole protein sequence instead of focusing on RNA- or DNA-binding domains. The "Yes" option will reveal two new sets of radial buttons; the first group allows the user to specify whether only RNA-binding domains should be selected, or whether both RNA- and DNA-binding domains should be examined. The second group allows the user to include predicted disordered regions in the query protein as part of the calculation (recommended).
17. The last group of radial buttons allows the user to specify whether she/he wishes to query partners from the coding (mRNA) or noncoding (tRNA, rRNA, ncRNA, etc.) transcriptome. A valid email address can be entered for notification of job completion. The user should then click "Submit query."
18. catRAPID omics results may take a few hours to be generated, depending on the size of the selected transcriptome. The results page is identical to that returned for the RNA vs. proteome predictions described in **step 12** above, and results are interpreted the same way (*see* **Note 16**).

4 Notes

1. At present, none of the available computational tools for predicting RNA-protein interaction partners reports whether experimental evidence for a specific interaction is available (i.e., even when an interaction partner is *known*, the software will make a *prediction*, which may or may not correspond to the experimentally validated interaction partner). Thus, as a first step, the user should always search published literature (via search engines such as **NCBI/PubMed** (http://www.ncbi.nlm.nih.gov/) or **Google Scholar** (http://scholar.google.com) and relevant databases (*see* Subheading 3.1) for existing experimental data regarding a specific RNA-protein interaction of interest. In addition to the resources described in Subheading 3.1 and Table 1, many additional valuable databases and servers that provide extensive information regarding in vivo RNA-protein complexes, RNA binding proteins and their recognition sites, RNA-protein complexes, and RNA-protein interaction networks are becoming available.

OMICtools (http://omictools.com) provides an extensive and up-to-date directory of these resources [42]).

2. According to OmicsTools (http://omictools.com)[42], the ENA (European Nucleotide Archive) Sequence Search tool (http://www.ebi.ac.uk/ena) [43], hosted by the EMBL-EBI, is a "nucleotide search tool which is far faster than BLAST for large datasets, with only a marginal loss in search sensitivity" (http://omictools.com/ena-sequence-search-s2042.html).
3. If the query sequence corresponds to a known protein or RNA, the **NCBI "Gene" database** (http://www.ncbi.nlm.nih.gov/gene) is an excellent starting point for investigating whether potential binding partners of the query have been previously identified. (Tip: Because proteins and RNAs from humans are usually better annotated than those from other organisms, valuable information can be obtained by visiting the **Gene** page for the human homolog of a query sequence.) On the sidebar of the **Gene** full report page, the **Table of Contents** may include links to Pathways from BioSystems (for a protein query), and Interactions (for both protein and RNA queries). In addition, the General Gene Information link provides a list of GO annotations, such as "RNA binding," or under the GO Component heading, a list of specific macromolecular complexes with which the protein or RNA has been associated. Finally, the GenRifs section can provide direct access to the most relevant literature regarding RNAs bound by the query protein.
4. Because **NPInter** [26] specifically excludes protein-ncRNA interactions that involve tRNA or rRNA, the user interested in such interactions should consult the **PRD** (*see* Subheading 3.1, **step 3**), as well as the PDB and NDB (*see* Subheading 3.1, **step 4**) because these databases contain many such interactions.
5. **NPInter** [26] provides tools for: (a) BLASTing a given protein or RNA sequence against every RNA or protein sequence in NPInter (http://www.bioinfo.org/NPInter/blast/blast_link.cgi); (b) predicting whether or not a specific lncRNA-protein interaction is likely, using **lncPro** (http://www.bioinfo.org/NPInter/lncPro.htm) (*see* Subheading 3.3); and (c) predicting whether two specific RNAs are likely to interact, using **RIsearch** (http://www.bioinfo.org/NPInter/RISearch.htm).
6. The **PRD** [28] contains 10,817 documented physical interactions between RNA and proteins extracted from BioGRID [22], IntAct [22], and the PDB [27], including many interactions that involve tRNA or rRNA.
7. The **PDB Advanced Search** (http://www.rcsb.org/pdb/search/advSearch.do?search=new) is a powerful tool that allows the user to BLAST a sequence of interest against all

structures in the database, to identify GO annotations, citations in publications, etc. In addition, the PDB offers several built-in visualization tools (http://www.rcsb.org/pdb/secondary.do?p=v2/secondary/visualize.jsp - RCSBviewer), as well as links to additional resources and software for analyzing macromolecular structures (http://www.rcsb.org/pdb/static.do?p=general_information/web_links/index.html).

8. The **NDB** [24, 25] focuses on structures that contain either DNA or RNA and provides links to many valuable RNA sequence and structure analysis tools (http://ndbserver.rutgers.edu/ndbmodule/services/index.html) as well as software for identifying RNA motifs and for predicting secondary and tertiary structures of RNA molecules (http://ndbserver.rutgers.edu/ndbmodule/services/softwares.html).
9. **RPISeq** [30] consists of Random Forest (RF) and Support Vector Machine (SVM) machine-learning classifiers that predict the probability of interaction between an RNA and a protein based solely on their primary sequences. In this method, RNA sequences are encoded as normalized frequencies of RNA tetrads, and protein sequences are encoded using a conjoint triad feature (CTF) method originally proposed by Shen et al for predicting protein-protein interactions [44]. Based on the propensity of the observed conjoint triads to bind the observed RNA tetrads, RPISeq outputs the probability that the submitted RNA and protein will interact. In performance evaluation experiments using10-fold cross-validation on RPI2241 (a nonredundant dataset including 2241 RNA-protein pairs derived from PRIDB [35]), the RPISeq SVM classifier achieved an accuracy of 87.1 %, and the Random Forest classifier achieved an accuracy of 89.6 %. Additional performance metrics and comparisons with other methods are provided in Muppirala et al [17, 30].
10. The **RPISeq** webserver is currently capable of returning predictions for up to 100 sequences (or up to 100 Mb) in a single run. For larger datasets, a stand-alone version of RPISeq is available upon request to the author (instructions available at http://pridb.gdcb.iastate.edu/RPISeq/contact.php).
11. **lncPro** [31] encodes potentially interacting lncRNA and protein sequences as feature vectors of identical dimensions (based on secondary structure, hydrogen-bonding, and van der Waal's interaction propensities observed in 41 RNP complexes from the PDB) and uses matrix multiplication to generate an interaction score for each RNA-protein pair. The algorithm was trained and tested on a dataset of 726 nonredundant RNA-protein pairs extracted from 18 complexes in the PDB that contain RNAs longer than 100 nts. In 4-fold cross-validation experiments, the method obtained a

Discriminative Power (DP) value of 90.3 %. Additional performance metrics and comparisons with other methods (including catRAPID) are provided in [31].

12. A stand-alone version of lncPro is available at: http://bioinfo.bjmu.edu.cn/lncpro/#fragment-3.
13. The original **catRAPID graphic** algorithm [32] generates predictions using interaction profiles of the query protein and RNA sequences, which are based on several physicochemical properties, including predicted secondary structure, hydrogen bonding, and van der Waals interaction propensities. On a nonredundant dataset of 858 RNA-protein complexes from the PDB, the reported discriminative power was 78 %. Additional performance metrics and comparisons with other methods are provided in [18].
14. Currently eight proteome libraries are available, from: *C. elegans*, zebrafish, fruit fly, human, mouse, brown rat, yeast, and western clawed frog. The user also has the option of submitting a custom sequence library.
15. To generate a custom library, the user should select the ninth radial button in the "library" section. This will cause a text box to appear along with a link to generate a library. Select the red-highlighted "Generate" text to navigate to the library generation tool. (Note that simply clicking on the link will open the library generation tool in the current tab, which may cause data loss. It is highly recommended that the user right-click or command-click the link to open the library generation tool in a new window.) This takes the user to a library submission page, where she/he can label the library (with a descriptive name, including the source organism) and submit an email address for notification purposes. Select the button to upload a file of FASTA formatted sequences (≤500 sequences), click "Submit query" at the bottom of the page, and wait for the library to finish processing. The user will be provided with an ID reference for the library.
16. The first column of the catRAPID omics result table (Fig. 2) lists the numerical rankings assigned to the protein-RNA interactions, from most to least probable. The first row of the table corresponds to the most highly ranked RNA-protein pairing. The second column contains the ID for protein being analyzed (with a clickable link to the protein's ENSEMBL entry). The third column contains the ID for the query RNA sequence (linked to its sequence). The fourth column contains the normalized interaction propensity (*Z*-score), with higher values indicating a more likely interaction. The fifth column contains the discriminative power (%) score. The sixth column contains the interaction strength, which is an indicator of the specificity

of the reaction; a low value for the interaction strength may indicate that the protein binds the RNA nonspecifically. The seventh column indicates whether the protein possesses a known RNA-binding domain; the eight indicates whether the RNA has any recognized protein-binding motifs. The ninth column is a "star ranking" of the results. The "star rank" of an interaction is a value from 0 to 3, calculated based on three criteria: (a) whether the protein has an RNA-binding domain, or both a DNA-binding domain *and* a disordered region: if both are present, 1 is added to the star rank; if the protein has only a DNA-binding domain or only a disordered region, 0.5 is added to the star rank. If the protein has neither an RNA- or DNA-binding domain and no disordered regions, 0 is added to the rank score; (b) whether the protein has any RNA-binding motifs: if so, 1 is added to the score, and 0 otherwise; (c) the predicted interaction propensity (which is normalized on a scale of 0–1) and added to the scores from (a) and (b).

Acknowledgments

This work was supported by NIH grant GM066387 and a Presidential Initiative for Interdisciplinary Research (PIIR) award from Iowa State University to DD. We thank Rasna Walia for valuable discussions and suggestions.

References

1. Rinn JL, Ule J (2014) Oming in on RNA-protein interactions. Genome Biol 15(1):401
2. Mattick JS, Rinn JL (2015) Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol 22(1):5–7
3. Sutandy FX, Hsiao FS, Chen CS (2015) High throughput platform to explore RNA-protein interactomes. Crit Rev Biotechnol 36(1):11–19
4. Silverman IM, Li F, Alexander A et al (2014) RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol 15(1):R3
5. Buenrostro JD, Araya CL, Chircus LM et al (2014) Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nat Biotechnol 32(6):562–568
6. Ray D, Kazan H, Cook KB et al (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499(7457):172–177
7. Castello A, Horos R, Strein C et al (2016) Comprehensive identification of RNA-binding proteins by RNA interactome capture. Methods Mol Biol 1358:131–139
8. Strein C, Alleaume AM, Rothbauer U et al (2014) A versatile assay for RNA-binding proteins in living cells. RNA 20(5):721–731
9. Kramer K, Sachsenberg T, Beckmann BM et al (2014) Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. Nat Methods 11(10):1064–1070
10. Dieterich C, Stadler PF (2013) Computational biology of RNA interactions. Wiley Interdiscip Rev RNA 4(1):107–120
11. Chu C, Spitale RC, Chang HY (2015) Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol 22(1):29–35
12. Goodrich JA, Kugel JF (2006) Non-coding-RNA regulators of RNA polymerase II transcription. Nat Rev Mol Cell Biol 7(8):612–616
13. Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15(1):7–21
14. Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises

from the RNA world. Genes Dev 23(13): 1494–1504
15. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81:145–166
16. Yang Y, Wen L, Zhu H (2015) Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein. Cell Biosci 5:59
17. Muppirala U, Lewis BA, Dobbs D (2013) Computational tools for investigating RNA-protein interaction partners. J Comput Sci Syst Biol 6:182–187
18. Cirillo D, Livi CM, Agostini F et al (2014) Discovery of protein-RNA networks. Mol Biosyst 10(7):1632–1642
19. Walia RR, Caragea C, Lewis BA et al (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinform 13(1):89
20. Puton T, Kozlowski L, Tuszynska I et al (2012) Computational methods for prediction of protein-RNA interactions. J Struct Biol 179(3): 261–268
21. Yan J, Friedrich S, Kurgan L (2015) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. Brief Bioinform 17(1):88–105
22. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R et al (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43 (Database Issue):D470–D478
23. Orchard S, Ammari M, Aranda B et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42 (Database Issue):D358–D363
24. Coimbatore Narayanan B, Westbrook J, Ghosh S et al (2014) The Nucleic Acid Database: New features and capabilities. Nucleic Acids Res 42 (Database issue):D114–D122
25. Berman HM, Olson WK, Beveridge DL et al (1992) The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 63:751–759
26. Yuan J, Wu W, Xie C et al (2014) NPInter v2.0: an updated database of ncRNA interactions. Nucleic Acids Res 42(Database issue): D104–D108
27. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
28. Fujimori S, Hino K, Saito A et al (2012) PRD: A protein-RNA interaction database. Bioinformation 8(15):729–730
29. Cook KB, Kazan H, Zuberi K et al (2011) RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 39(Database Issue): D301–D308
30. Muppirala UK, Honavar VG, Dobbs D (2011) Predicting RNA-protein interactions using only sequence information. BMC Bioinform 12:489
31. Lu Q, Ren S, Lu M et al (2013) Computational prediction of associations between long noncoding RNAs and proteins. BMC Genom 14:651
32. Bellucci M, Agostini F, Masin M et al (2011) Predicting protein associations with long noncoding RNAs. Nat Methods 8(6):444–445
33. Agostini F, Zanzoni A, Klus P et al (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. Bioinformatics 29(22):2928–2930
34. Cirillo D, Agostini F, Tartaglia GG (2013) Predictions of protein-RNA interactions. Wiley Interdiscip Rev RNA 3(2):161–175
35. Lewis BA, Walia RR, Terribilini M et al (2011) PRIDB: a Protein-RNA interface database. Nucleic Acids Res 39 (Database Issue):D277–D282
36. Wang Y, Chen X, Liu ZP et al (2013) De novo prediction of RNA-protein interactions from sequence information. Mol Biosyst 9(1):133–142
37. Pancaldi V, Bahler J (2011) In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Res 39(14):5826–5836
38. Suresh V, Liu L, Adjeroh D et al (2015) RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res 43(3):1370–1379
39. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
40. Cirillo D, Marchese D, Agostini F et al (2014) Constitutive patterns of gene expression regulated by RNA-binding proteins. Genome Biol 15(1):R13
41. Livi CM, Klus P, Delli Ponti R et al (2015) catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics 32(5):773–775
42. Henry VJ, Bandrowski AE, Pepin AS et al (2014) OMICtools: an informative directory for multi-omic data analysis. Database (Oxford) doi: 10.1093/database/bau069.bau069
43. Leinonen R, Akhtar R, Birney E et al (2011) The European Nucleotide Archive. Nucleic Acids Res 39 (Database issue):D28–D31
44. Shen J, Zhang J, Luo X et al (2007) Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci U S A 104(11):4337–4341

Chapter 9

Isolation of Nuclear RNA-Associated Protein Complexes

Ranveer Singh Jayani, Amanjot Singh, and Dimple Notani

Abstract

Noncoding RNAs, being at the center stage of the organismal development and homeostasis, warrant a detailed analysis to utilize their full therapeutic potential. They form complexes with various proteins that enable the noncoding RNAs to acquire specific cellular functions such as the transcriptional outcomes that are controlled in a spatio-temporal manner. In this protocol, we describe a method to isolate such known (and unknown) protein complexes bound to a nuclear noncoding RNA.

Key words In vitro transcription, Biotin-labeling, RNA-pull-down, Mass-spectrometry, Long noncoding RNA (LncRNA)

1 Introduction

The information obtained from next-generation sequencing provides an unrivaled resource of genetic information that can facilitate discoveries of medical and biological importance. Advancements of sequencing techniques within the last decade have made it possible to sequence large DNA/RNA sequences in an affordable time and cost manner [1–3]. These advancements have revolutionized the genomics field and have presented a surprising finding that majority of genome transcribes into pervasive transcripts [4]. Only a fraction of this vast transcriptome comprises coding transcripts, which ultimately translate into protein molecules. Most of the remaining transcripts are non-protein coding, but nevertheless, are important regulators of the biological processes [5]. Studies have defined and characterized the properties of these noncoding transcripts, which include lncRNA (long noncoding RNA) [6, 7], miRNA (micro RNA) [8], rRNA (ribosomal RNA), piRNAs (PIWI-interacting RNAs) [9], promoter-associated RNAs [10], and the recently discovered eRNAs (enhancer-RNA) [11]. These noncoding entities regulate the transcriptional, posttranscriptional, and translational outcomes of coding genes [12, 13].

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_9,

The noncoding RNAs (ncRNAs) control various aspects of cellular functions by either directly binding the target RNA for degradation, e.g., miRNA [14]; or by recruiting the regulatory complexes; that in turn form the crucial regulatory components of the genome [12]. The primary sequence together with secondary structure of these noncoding RNAs plays a central role in determining the type and nature of cofactors/complexes bound to these lncRNA, which ultimately determines the regulatory output of such interactions [6]. Therefore, the study of the proteins associated with these regulatory RNAs becomes of utmost importance in understanding the regulatory mechanisms utilized by the noncoding RNAs.

The protocol we describe here can be used for the identification of protein complexes that are bound to the nuclear RNA of interest. In this method, an in vitro transcribed RNA immobilized on streptavidin beads acts as bait for capturing the interacting proteins. The captured proteins can then be identified by mass spectrophotometry or subjected to immunoblotting for the identification of proteins in an open-ended and known target-specific manner, respectively.

Briefly, using the biotin-labeling system, RNA of interest is in vitro transcribed from a vector backbone harboring DNA template corresponding to the transcribed RNA, which is under the control of bacteriophage promoters such as T7, T3, or SP6. The plasmid is then linearized to avoid transcription of RNA from the vector backbone. The RNA integrity and quality is ascertained by gel electrophoresis. Subsequently, purified RNA is allowed to fold into its native secondary structure under physiologically relevant conditions. Folded RNA is then incubated with nuclear lysate to allow the formation of specific RNA-protein complexes under stringent buffer conditions. Following this, RNA-protein complexes are immobilized on streptavidin magnetic beads and washed multiple times in stringent conditions to avoid capturing nonspecific proteins. These complexes are then eluted in solution by competitive displacement with avidin protein that has more affinity for streptavidin. The protein is then subjected to mass spectrophotometric analysis or denatured in SDS buffer for immunoblotting.

2 Materials

The materials required are divided into the subsections based on the various steps of the protocol. All the stock solutions and buffers should be prepared in RNase-free water (*see* **Notes 1** and **2**).

2.1 In Vitro Transcription Components

1. Biotin RNA labeling mix (10×) (RocheLifeScience, catalog # 11685597910) for in vitro transcription reactions. The mix contains 10 mM ATP, 10 mM CTP, 10 mM GTP, 6.5 mM UTP, 3.5 mM, Biotin-16-UTP (pH 7.5).

2. T7 RNA polymerase.
3. Transcription buffer (5×): 200 mM Tris–HCl (pH 7.9), 50 mM NaCl, 30 mM $MgCl_2$, 10 mM spermidine (Promega, catalog # P1181).
4. Dithiotreitol (DTT), 100 mM.
5. DNase I.
6. EDTA 0.2 M (pH 8.0).
7. Ammonium acetate 5 M.

2.2 Nuclei Extract Preparation Components

1. HeLa cells.
2. Phosphate-buffered saline (PBS, 1×, pH 7.4): 10 mMPO_4^{3-}, 137 mM NaCl, 2.7 mM KCl.
3. Nuclei isolation buffer (NIB): 40 mM Tris–HCl (pH 7.5), 20 mM $MgCl_2$, 1.28 M sucrose, 4 % Triton X-100, 1 mM PMSF, protease inhibitors, and 20 U/ml SUPERase inhibitor (Thermo Fisher Scientific, catalog # AM2694) (*see* **Note 3**).

2.3 RNA Pull-Down Components

1. RNA structure buffer (RSB): 10 mM Tris–HCl (pH 7.0), 100 mM KCl, 10 mM $MgCl_2$, 1 mM PMSF, protease inhibitors, and 20 U/ml SUPERase inhibitor.
2. RNA immuno-precipitation (RIP) buffer: 25 mM Tri–HCl (pH 7.4), 150 mM KCl, 0.5 mM DTT, 0.5 % NP40, 1 mM PMSF, protease inhibitors, and 20 U/ml SUPERase inhibitor.
3. DynabeadsMyOne Streptavidin beads (Thermo Fisher Scientific, catalog # 65001).
4. Dounce homogenizers (2 ml).

3 Methods

3.1 Cloning of RNA Sequence in a Plasmid

Primer design, PCR amplification, and subsequent cloning of DNA fragments have been described previously in multiple protocols [15, 16]. Hence, these steps have not been described in detail in this chapter. Basically, the RNA should be cloned in a plasmid containing bacteriophage promoters (e.g., T3, T7, and SP6). It is crucial to clone the correct RNA transcribing sequence corresponding to either sense or anti-sense RNA of interest. For example, if the promoter-associated anti-sense RNA is being studied, then the corresponding DNA should be cloned in reverse orientation such that anti-sense transcript is transcribed from the plasmid.

3.2 In Vitro Transcription of Biotin-Labeled RNA

1. Linearize the plasmid with the cloned DNA sequence using appropriate restriction enzyme such that the backbone is cut immediately after the transcribing region of interest while bacteriophage promoter is intact (*see* **Notes 4–6**).

2. Run the digested product on gel to purify the linearized fragment of DNA.
3. Purify the linearized plasmid using gel purification columns.
4. Prepare the in vitro transcription mix by assembling the following components in a microcentrifuge tube (*see* **Note 7**).
   (a) 1 μg (*x* μl) of linearized plasmid as template
   (b) 2 μl Biotin RNA labeling mix (10×)
   (c) 4 μl Transcription buffer (5×)
   (d) 2 μl DTT (100 mM)
   (e) *x* μl RNase-free water to make the final volume of 20 μl
   (f) 2 μl appropriate RNA polymerase (20 U) (T7 polymerase in this case)
5. Mix the contents of the tube by tapping gently, centrifuge briefly and incubate at 37 °C for 2 h (*see* **Note 8**).
6. Add 1 μl of DNase I (10 U) and incubate at 37 °C for 15 min to digest the template DNA.
7. Stop the reaction by adding 2 μl EDTA (0.2 M, pH 8.0).
8. Bring the volume to 100 μl by adding RNAse-free water.
9. To precipitate RNA, add 1 volume (100 μl) of 5 M ammonium acetate.
10. Incubate on ice for 10–15 min and centrifuge at high speed (13,000 × *g*) for 12 min at 4 °C.
11. Decant the supernatant and wash the pellet in 70 % ethanol. Air-dry the pellet briefly and resuspend in RNase-free water (*see* **Notes 9** and **10**).

3.3 Preparation of Nuclei

1. Grow HeLa cells (or any other mammalian cell-line of interest) to a confluency of 90–95 % in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10 % FBS and penicillin/streptomycin, under 5 % CO_2 atmosphere.
2. Harvest 1×10^7 cells by scrapping and wash the cells once with ice-cold PBS and gently resuspend in 2 ml PBS.
3. To this, add 2 ml nuclear isolation buffer (NIB) and gently resuspend the cell pellet.
4. Add 6 ml of distilled water, mix well and keep on ice for 20 min, gently shaking intermittently (*see* **Note 11**).
5. Pellet the nuclei by centrifugation at 2500 × *g* for 15 min and proceed for preparing the nuclei extract (*see* **Note 12**).

3.4 RNA Pull-Down from Nuclei Extract

1. Gently resuspend the nuclei pellet in 1 ml RIP buffer and incubate on ice for 5 min.
2. Transfer the nuclei solution to a prechilled dounce homogenizer and shear the nuclei mechanically with 20 strokes using

the pestle A keeping it on ice. Repeat the same procedure with pestle B (*see* **Notes 13** and **14**).

3. Pellet down the nuclear membranes and other debris by centrifugation at 13,000 × *g* for 10 min at 4 °C.
4. Quantify the protein in nuclear extract using the colorimetric protein assay.
5. Incubate 1 μg of RNA with 20 μl of RNA structure buffer to incubate at room temperature for 5 min. Mix the folded 1 μg RNA (from Subheading 3.2) with 1 mg of nuclear extract in 1 ml RIP buffer and incubate at room temperature for 1 h, with intermittent mixing (*see* **Notes 15–17**).
6. During the incubation, wash 60 μl of DynabeadsMyOne Streptavidin beads with 1 ml PBS followed by another wash with 1 ml RIP buffer.
7. Add the beads to the nuclei extract-RNA binding reaction and incubate for 1 h with intermittent mixing.
8. Wash the beads five times with RIP buffer by mixing and collecting the beads using a magnetic stand.
9. Boil the RNA-protein complex bound streptavidin beads in 2× SDS loading buffer or elute the complexes from streptavidin beads by adding free biotin.

3.5 Down-Stream Processing

The RNA-protein complexes retrieved can be used for various downstream processing depending on the desired results. The simplest way to identify candidate proteins is to boil the eluates with 2× loading dye and perform immunoblotting with antibodies against probable candidate proteins. For a detailed knowledge of the proteome pulled down by the RNA, the eluted samples could be subjected to mass-spectrometric analysis (*see* **Note 18**).

4 Notes

1. RNA is highly unstable and can be degraded by trace amounts of RNase in the buffers and plasticware used. All buffers should be made in DEPC-treated water. Wear gloves and clean all pipettes and bench surfaces with RNase inhibitor solution.
2. Di-ethyl pyro-carbonate (DEPC) is a nonspecific inhibitor of RNases. DEPC (0.05–0.1 %v/v) should be added to water and kept at room temperature overnight prior to autoclaving. DEPC is carcinogenic and proper precautions should be taken while handling it. A safer alternative to DEPC, di-methyl-propyl carbonate, can be used in the same way.
3. For interrogating RNA bound proteins from nuclear compartment only. For cytosolic fractionation, see other protocols.

4. We used T7 RNA polymerase for a T7 promoter-containing vector. Choose either T3, T7, or SP6 RNA polymerase depending on the plasmid used for cloning the DNA.
5. Plasmids such as pTRIPLEscript family (Ambion) contain T3, T7, and SP6 in tandem, upstream of multiple cloning sites and provide a flexibility of using any of the three polymerases.
6. Extraneous transcripts have been reported to appear in addition to the desired transcript when templates contain 3′ overhangs. Therefore, it is recommended that plasmids should not be linearized with enzymes that leave a 3′ overhang.
7. SUPERase inhibitor can be added to prevent the action of RNases in the reaction mixture.
8. For the regions that are hard to transcribe (e.g., High G-C content, secondary structures), use DMSO (3–6 %), higher incubation temperature, or denature the linearized plasmid at 65 °C before incubation
9. Biotin labeled RNA can be used immediately or stored ethanol precipitated at −80 °C.
10. To confirm that full-length transcript was generated, run a small amount of the product on agarose gel.
11. Incubation on ice may have to be optimized depending on the cell line used for nuclei preparation.
12. To check for the quality of nuclei preparation, take 10 μl nuclei suspension on a glass slide and examine under a microscope at 20× magnification.
13. The homogenizer should be prechilled on ice to prevent temperature shock and degradation of the samples.
14. Pestle should be moved slowly during homogenization to prevent frothing in the sample.
15. Proper folding of RNA is required for the binding to interacting proteins. However, RNA has an inherent property of folding into the most favorable secondary structure that results in lowest free energy change. The RNA transcribed in Subheading 3.2 folds into secondary structure as it is transcribed.
16. As a control for nonspecific background pull down by beads, a pull-down with beads alone should be included.
17. Similarly, another RNA or the anti-sense RNA to the RNA of interest should also be used separately to conform the specificity of RNA-protein interaction.
18. Mass spectrometric analysis would require protein amounts visible with silver stain. Typical sample requirements for mass spectrometry are 10–100 fmol.

References

1. Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59
2. Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138
3. Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352
4. Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet 9:e1003569
5. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: Insights into functions. Nat Rev Genet 10:155–159
6. Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641
7. Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23: 1494–1504
8. Ebert MS, Sharp PA (2012) Roles for microRNAs in conferring robustness to biological processes. Cell 149:515–524
9. Luteijn MJ, Ketting RF (2013) PIWI-interacting RNAs: From generation to transgenerational epigenetics. Nat Rev Genet 14:523–534
10. Napoli S, Pastori C, Magistri M et al (2009) Promoter-specific transcriptional interference and c-myc gene silencing by siRNAs in human cells. EMBO J 28:1708–1719
11. Li W, Lam MT, Notani D (2014) Enhancer RNAs. Cell Cycle 13:3151–3152
12. Li W, Notani D, Ma Q et al (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498:516–520
13. Holoch D, Moazed D (2015) RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet 16:71–84
14. Sun K, Lai EC (2013) Adult-specific functions of animal microRNAs. Nat Rev Genet 14: 535–548
15. Surzycki S (2000) DNA cloning—Experimental procedures. In: Surzycki S (ed) Basic techniques in molecular biology. Springer, Heidelberg, pp. 320–373
16. Purbey PK, Jayakumar PC, Patole MS et al (2006) pC6-2/caspase-6 system to purify glutathione-S-transferase-free recombinant fusion proteins expressed in *Escherichia coli*. Nat Protoc 1:1820–1827

Part III

Functional Studies of Promoter-Associated RNAs

Chapter 10

Identification of Long Noncoding RNAs Associated to Human Disease Susceptibility

Marco Magistri and Dmitry Velmeshev

Abstract

Transcriptomic as well as in vivo studies have revealed the importance of several lncRNAs in many complex diseases including cancer, cardiovascular, and neurological disorders. In this protocol, we describe how to perform RNAseq data analysis to identify lncRNAs associated with disease states utilizing the open-source software **CANEapp** (**app**lication for **C**omprehensive automated **A**nalysis of **N**ext-generation sequencing **E**xperiments).

Key words Long noncoding RNAs (lncRNAs), RNA sequencing, CANEapp

1 Introduction

During the decade following the publication of the Human Genome, noncoding RNAs (ncRNAs) have reshaped the landscape of genome regulation. More than 70 % of mammalian genome is transcribed, mostly as ncRNAs [1] and the number of ncRNAs in eukaryotic genomes increases as a function of developmental complexity [2–4]. Furthermore, many long ncRNAs (lncRNAs), transcripts longer than 200 nt, are expressed tissue-specifically and are thought to mediate fundamental biological functions [5, 6]. LncRNAs have roles in regulating the expression and/or function of protein-coding genes [7] and in some cases the epigenetic status of entire genomic loci [8, 9].

Next-generation sequencing technologies have revolutionized the field of genetic and molecular biology. The power of RNA sequencing (RNA-seq) is in delivering high-throughput quantitative and qualitative information about all RNA species (transcriptome) expressed in the analyzed sample. Indeed, RNA-seq provides genome-wide estimate of transcripts expression and, at the same time, allows the discovery of previously non-annotated transcriptional elements, such as novel RNA transcripts and noncoding RNA genes. Thus, RNA-seq represents the ideal tool to perform

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_10,

genome-wide discovery of novel ncRNAs and to identify ncRNAs whose expression is dysregulated in a disease state. RNA-seq data analysis typically consists of a number of consecutive steps and requires the combination of multiple software packages to perform reads alignment, transcript reconstruction, and gene expression estimation. In this protocol, we describe how to perform RNAseq data analysis to identify lncRNAs associated with disease states utilizing the open-source software **CANEapp** (**app**lication for **C**omprehensive automated **A**nalysis of **N**ext-generation sequencing **E**xperiments). CANEapp is a free, open-source tool created to provide biologists with no background in bioinformatics and computational science with an easy way to perform cutting-edge analysis of large-scale RNA-seq data. One of the main features of CANEapp is the capability to perform ab initio assembly of transcripts, which does not rely on previous transcriptome annotations and thus allows the discovery of not previously annotated transcripts [10, 11]. CANEapp includes a workflow that filters single-exon transcripts that potentially originate from transcriptional noise or sequencing artifacts, filters out lowly expressed loci, and classifies novel loci into noncoding RNAs or potential novel protein-coding genes.

2 Material

CANEapp consists of a streamlined RNA-seq analysis pipeline that efficiently manages the computational resources, parallelizes computation, and automates the entire analysis. CANEapp functions through a point and click windows-based graphical user interface to easily design experiments and manage analysis setup.

To use CANEapp for RNA-seq analysis, users need to:

1. Download CANEapp package here: (http://psychiatry.med.miami.edu/research/laboratory-of-translational-rna-genomics/CANE-app) to a Mac or Windows computer with the latest version of Java installed.
2. Have access to a Linux server with Python version 2.7 or higher. Since RNA-seq analysis is computationally demanding, CANEapp needs a server with at least 30 GB of RAM. CANEapp can be used with a variety of Linux operating systems (Ubuntu, CentOS, RedHat, Fedora), a cloud server (such as Amazon EC2) or a Linux cluster using LSF job scheduling system. If using Amazon Cloud just search for CANEapp Amazon Machine Image, create a new instance based on it and use it as the server for CANEapp. If using a Linux server with administrative rights, run CANEapp as the root user and all prerequisites will be installed automatically. Users without administrative rights need to contact system administrator to install the prerequisites. All prerequisites can be

found in the misc folder of CANEapp (For CentoOS, RedHat, and Fedora: CANE_library_CentOS.sh. For Ubuntu: CANE_library_Ubuntu.sh).

3 Methods

Open the CANEapp JAR file. Analysis setup will start on the "Manage projects" tab. On this tab user can create new projects, check status of running projects, and remove existing projects.

3.1 Create a New Project

1. On the "Manage projects" tab click "Create new project" button.
2. Type in the name of the project and browse to a location on the computer where the files related to the project want to be saved.
3. Press submit. Now the project's name is displayed in the list of recent projects and user can proceed with the experimental design (Fig. 1).

3.2 Add Experimental Groups

1. Click on the tab "Add groups"
2. Type in the group's name (for instance Disease and Control)
3. Press "Add Group" (*see* **Note 1**). Groups can be removed from the list by pressing "Remove Group" button (Fig. 2).

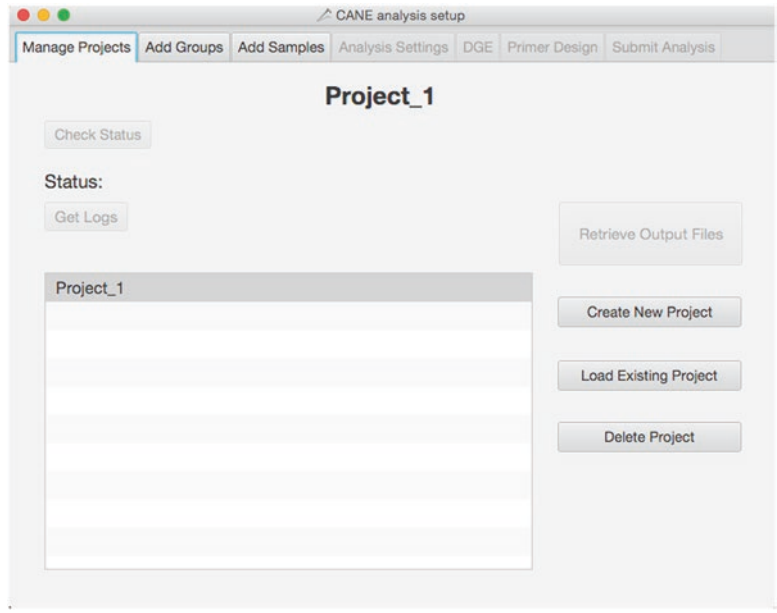


Fig. 1 In the "Manage Projects" tab of the graphical user interface, it is possible to create, load, or delete projects. Once the analysis is complete it is also possible to retrieve output files from this tab

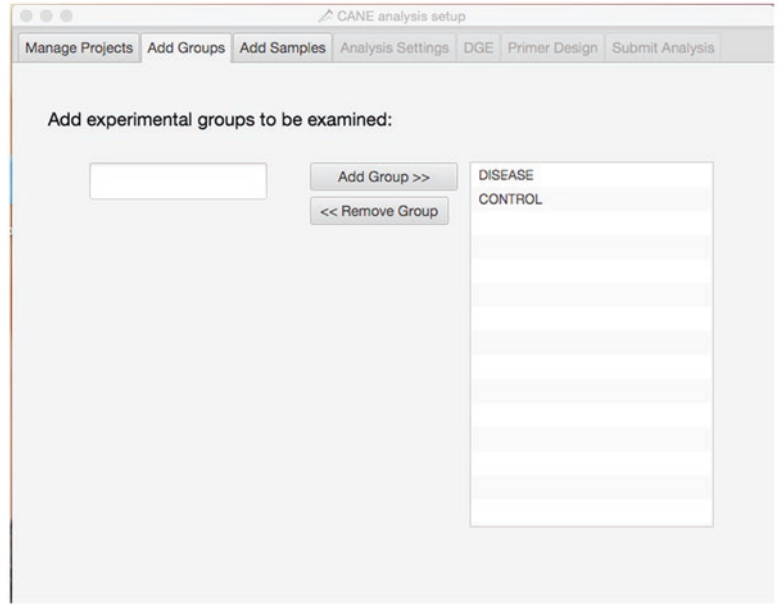


Fig. 2 The "Add Groups" tab of the graphical user interface allows users to add or remove experimental groups

Once at least one group has been added it will be possible to proceed to the next step and assign to groups the reads files of the sequenced samples.

3.3 Add Samples

1. On the "Add Samples" tab first select the experimental group from the list on the right and type in the name of the first sample. Then upload data files from a local computer or use the files that are already on the server (Fig. 3). CANEapp can be used to analyze ribo-depleted and poly-A-enriched libraries (*see* **Note 2**).
2. Specify the type of sequencing: single- or paired-end (*see* **Note 3**).
3. If data are uploaded from the local computer, browse to the raw reads file. The accepted format is fastq but tar.gz, tar, gz, or bz2-compressed fastq files, as well as SRA (NIH Short Sequence Archive) files can be uploaded. In case reads files are already on the server specify the full path to the file (including the file name) on the server (Fig. 4).
4. Specify the library preparation protocol used from a list of predesigned libraries prep or specify a custom library preparation protocol. By un-clicking "Default" it is possible to add additional information about the library that will help to increase analysis accuracy. Click "Add" and proceed with the remaining samples (Fig. 4) (*see* **Note 3**).

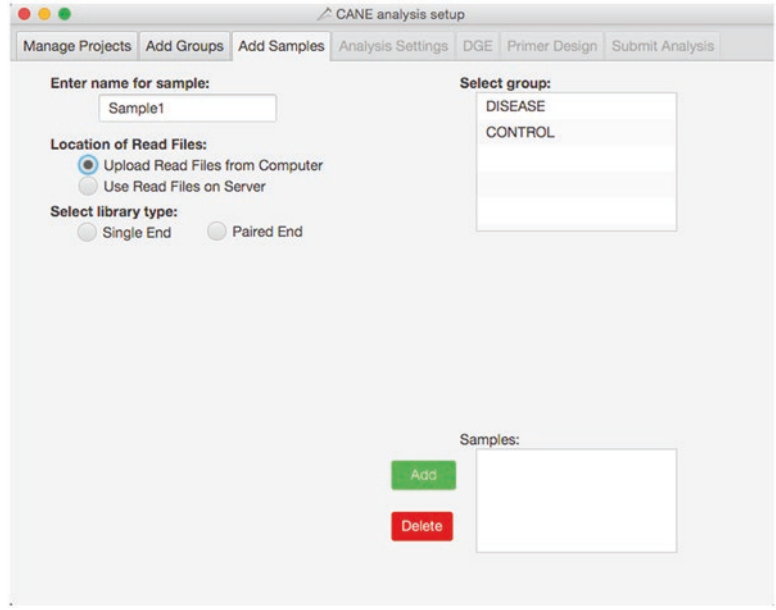


Fig. 3 In the "Add Samples" tab users can upload samples files and assign samples to experimental groups

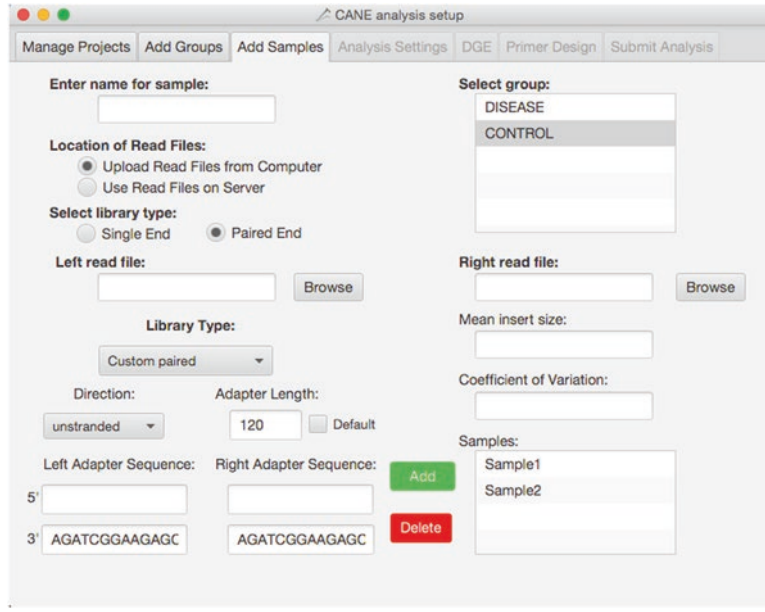


Fig. 4 The "Add Samples" tab permits users to specify the library preparation protocol used and to add additional information about the library

3.4 Specify Analysis Settings

1. Navigate to "Analysis settings."
2. By default, the pipeline will perform adaptor trimming ("Trim raw reads" option).

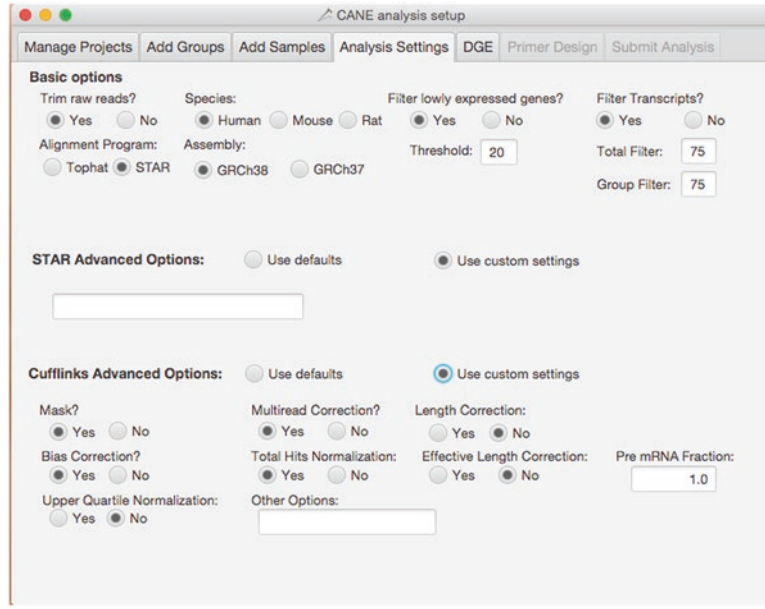


Fig. 5 In the "Analysis Settings" tab users can specify the analysis settings and choose the aligner software between STAR and Tophat. From this tab it is also possible to select and run custom options for alignment (Tophat or STAR) and assembly (Cufflinks)

3. Select the alignment program to be used between TopHat and STAR (*see* **Note 4**).
4. Cufflinks is used for transcriptome assembly (Fig. 5). By default, the pipeline will filter out single-exon transcripts ("Filter transcripts" option) based on what percentage of all samples (Total Filter) or samples from one group (Group Filter) expresses the transcript (*see* **Note 5**). The pipeline will then filter out lowly expressed genes ("Filter lowly expressed genes" option) based on minimum number of reads mapping to a transcript (*see* **Note 6**).

3.5 Set Up Differential Gene Expression Analysis

1. On the next tab, "DGE," select from three alternative workflows for differential gene expression (DGE) analysis: Cuffdiff, edgeR, or DESeq2 (Fig. 6). All three softwares can be run in parallel.
2. For Cuffdiff use default options or specify custom options (http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/).
3. For edgeR select from two different approaches to differential expression testing: Generalized Linear Models (GLM) or exact test. It is possible to use them in parallel.
4. For edgeR and DESeq2 select the pairwise combinations of the groups to compare (*see* **Note 7**).

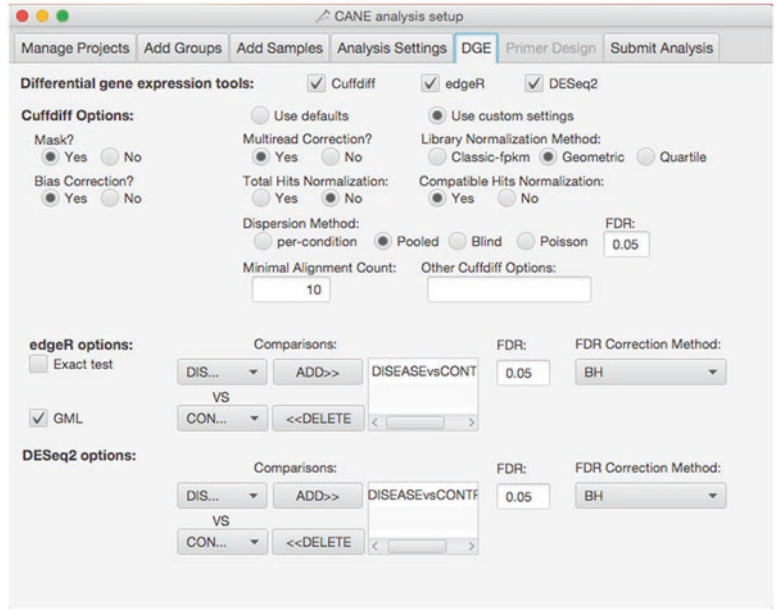


Fig. 6 In the Differential Gene Expression "DGE" tab it is possible to select which tool to use to compute differential expression analysis from Cuffdiff, edgeR, and DESeq2. Users can select all three and the pipeline will run them in parallel. From this tab it is also possible to select and run custom options for these tools

3.6 Submit the Analysis

1. Proceed to "Submit Analysis."
2. Specify user name, server address, home folder and either a password or a public key to access the server.
3. Click "Submit Analysis" button (Fig. 7). The first time CANEapp is run on a current server, it will take a minute to transfer the pipeline files to the server. After that a file transfer window with a progress bar will appear. A window will notify the files have been transferred (*see* **Notes 8–10**).

3.7 Check the Status of a Running Project

Once the project has been submitted, it is possible to check the status at any time on the "Manage Projects" tab. Select a project and click "Check Status" button.

3.8 Retrieve the Data

Once the project is completed, the Status will read "Done" and it will enable the Retrieve Output Files button. Click the button and the files will be downloaded in the local project folder.

The files will include one tab-delimited text file for each pairwise comparison between groups containing all the genes, another tab-delimited file containing only differentially expressed genes (based on FDR) and the third file containing genes filtered by both the expression and FDR. These files can be opened in Excel and contain information including the gene ID, gene name, gene classification, raw read counts for each sample (first column for each

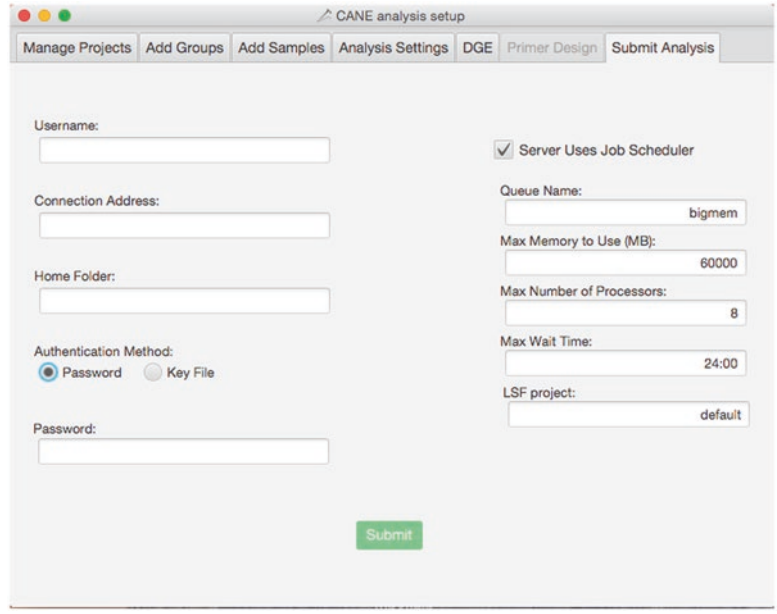


Fig. 7 In the "Submit Analysis" tab users need to specify username, served address, home folder, and either a password or a public key to access the server. If the server is a cluster using LSF system for job scheduling, it is possible to check the "Server Uses Job Scheduler" option and specify the cluster queue, amount of memory, and number of processors and time required for the analysis

sample) and FPKM (second column for each sample), log of fold change between the groups, and statistical values for differential expression (p and corrected-p value). Differentially expressed noncoding RNAs need to be prioritized for downstream functional analysis (*see* **Note 11**).

The other two output files are the GTF (Gene Transfer Format) files for all genes and for only differentially expressed genes. These files can be used to visualize reconstructed transcripts and loci on IGV (Integrated Genome Viewer).

3.9 Prioritization for Downstream Functional Analysis

The final output of CANEapp is differential expression analysis that usually results in a long list of differentially expressed genes that need to be filtered to identify noncoding RNAs that are likely to be associated with human disease and to select candidates for downstream validation and functional analysis. Unfortunately, there is not a standardized way to prioritize for candidates but there are certain filtering steps that might help with the process.

1. Set a threshold to define statistically significant differentially expressed lncRNAs. CANEapp has the potential to perform differential expression analysis utilizing three different work-

flows: Cuffdiff, EdgeR, and DEseq. By default the threshold for statistical significance is set to an adjusted p value of 0.05 but users have the possibility of decreasing this value for a more stringent analysis.

2. In the result file, together with the adjusted p values, it will also be provided the fold change (FC) difference of transcript expression between disease and control. Differentially expressed transcripts can thus be ranked by their fold change and only those having larger expression differences between disease and control should be considered for downstream analysis.
3. Prioritize based on expression level. Long noncoding RNAs are normally expressed at lower level compared to protein coding genes but some lncRNAs that are expressed at very low level may simply represent transcriptional noise due to low fidelity binding of RNA Pol2 to weak promoters. To overcome this possible problem lncRNAs can be ranked by their expression value (FPKM) and lncRNA having FPKM values lower than the 10th percentile will be discarded.
4. Prioritize differentially expressed lncRNAs based on their genomic proximity to genes relevant for the disease studied. Many lncRNAs, like promoter associated RNAs or natural antisense transcripts (NATs), have been shown to act in cis to regulate the expression/function of neighboring genes. Dysregulated expression of lncRNAs is likely to cause the altered expression/function of nearby genes.
5. Perform prediction analysis of transcription factor binding to the promoters of differentially expressed lncRNAs.
6. Consider the degree of conservation. LncRNAs have general low sequence conservation between species, maybe due to rapid evolution of lncRNAs. Nevertheless, there are many lncRNAs that are highly conserved and that are more likely to have a well-preserved function throughout evolution.

3.9.1 Primer Design

CANEapp includes a primer design tool to perform qRT-PCR validation of the gene expression estimated with RNA-seq or confirm presence of novel, previously unannotated genes.

1. Open the tab-delimited output file from CANEapp and select the genes to be validated by RT-PCR.
2. Copy the first column containing the Gene IDs (XLOC).
3. Navigate to the Primer Design tab and paste the IDs in the window.
4. Press “Submit Gene List” and wait until primers are designed.

4 Notes

1. RNAseq data from disease state and respective healthy control. In case preliminary RNA-seq data was already generated, power analysis can be performed to estimate the number of additional replicates sequenced to observe an effect of a certain size and with a set type II error (p value). For instance, the RNASeqPowerCalculator R package can be used to estimate the power based on preliminary RNA-seq count data: http://www2.hawaii.edu/~lgarmire/RNASeqPowerCalculator.htm
2. More than 90 % of total RNA consists of ribosomal RNA (rRNA) that, if sequenced, would take up the great majority of the reads; thus, rRNA needs to be removed from each sample before library preparation. One possibility is to capture polyadenylated transcripts, consisting of protein coding genes and polyA+-noncoding RNAs, and discard the non-polyadenilated fractions of the RNA before library preparation. Another way is to specifically deplete the samples of the rRNA and proceed with library preparation with the remaining RNA that will contain both PolyA+ and PolyA− transcripts. Ribosomal depletion is the preferred method for the study of long noncoding RNAs like promoter-associated RNAs, which consist of both PolyA+ and PolyA- transcripts.
3. In single-end, the library is sequenced only from one strand, thus generating reads from one end of the same DNA molecule. In paired-end, the library is sequenced from both strands, thus generating reads from both ends of the same DNA molecule. Paired-end sequencing greatly improves accuracy in the alignment of the reads and it is recommended over single-end for novel RNA transcripts discovery.
4. TopHat is a more conventional tool that is relatively slow but does not require a lot of resources, whereas STAR is an aligner with ultrafast performance but requires a lot of RAM. Then select the species and the assembly.
5. All the default analysis settings can be modified; however, CANEapp was tested with the default options and demonstrated high accuracy.
6. If you user is familiar with TopHat, STAR, and Cufflinks and want to modify default options for these tools, click "Use Custom Settings" next to one of them. For TopHat or STAR you have to put the options the same way you would use them in the command line (i.e., to change the max insertion length for TopHat alignment to 2, paste "—max-insertion-length 2" in the TopHatoption box).
7. In order to use edgeR or DESeq2 at least two replicates per experimental group are needed. Cuffdiff uses a statistics that

follows a normal distribution and a t test to perform differentially expression testing. Both edgeR and DESeq2 use a negative binominal distribution and a Fisher's exact test adopted for it.

8. If server is a cluster using LSF system for job scheduling, check the "Server Uses Job Scheduler" option and specify the cluster queue, amount of memory, and number of cores to be used for a job and max time to run a job. If using Amazon EC2 to perform the analysis, the easiest way is to use CANEapp Amazon Machine Image (AMI) to create a new instance with the amount of resources needed. Search for CANEapp AMI and create an instance with as much resources as you need. Then in the CANEapp GUI provide the public key for the instance together with the instance IP address in the GUI. Make sure the instance is running before submitting the analysis.
9. Before submitting the analysis, make sure to have at least 30 GB of RAM (or more for large projects) and enough disk space for the analysis. As a rule of thumb, free space three times the size of the raw data is needed to safely run the analysis.
10. Make sure the computer does not go to the sleep mode while the files are transferring. After the files are transferred the status of the project can be monitored and GUI can be closed. The rest of the analysis process will take place on the server side. CANEapp will utilize all available resources of the server so only one project at a time can be run. It is recommended to avoid running resource-demanding processes on the same server together with CANEapp. If using a cluster with the LSF job submitting system, it is possible to run several projects in parallel but, before starting another project, software and reference installation steps have to be completed.

References

1. Djebali S, Davis CA, Merkel A (2012) Landscape of transcription in human cells. Nature 489(7414):101–108. doi:10.1038/nature11233
2. Carninci P, Kasukawa T, Katayama S (2005) The transcriptional landscape of the mammalian genome. Science 309(5740):1559–1563. doi:10.1126/science.1112014
3. Kapranov P, Cheng J, Dike S (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316(5830):1484–1488. doi:10.1126/science.1138341
4. Mehler MF, Mattick JS (2006) Non-coding RNAs in the nervous system. J Physiol 575 (Pt 2):333–341. doi:10.1113/jphysiol.2006.113191
5. St Laurent G 3rd, Wahlestedt C (2007) Noncoding RNAs: Couplers of analog and digital information in nervous system function? Trends Neurosci 30(12):612–621. doi:10.1016/j.tins.2007.10.002
6. Mercer TR, Dinger ME, Sunkin SM (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 105(2):716–721. doi:10.1073/pnas.0706729105
7. Faghihi MA, Modarresi F, Khalil AM (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14(7):723–730. doi:10.1038/nm1784
8. Modarresi F, Faghihi MA, Lopez-Toledano MA (2012) Inhibition of natural antisense

transcripts in vivo results in gene-specific transcriptional upregulation. Nat Biotechnol 30(5):453–459. doi:10.1038/nbt.2158

9. Yap KL, Li S, Munoz-Cabello AM, Raguz S (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell 38(5):662–674. doi:10.1016/j.molcel.2010.03.021
10. Steijger T, Abril JF, Engstrom PG (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 3(10):1177–1184
11. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10):671–682
12. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25):14863–14868
13. Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20(9):1453–1454. doi:10.1093/bioinformatics/bth078

Chapter 11

Targeting Promoter-Associated RNAs by siRNAs

Sara Napoli

Abstract

Promoter-associated RNAs (paRNA) are long noncoding RNAs transcribed in sense or antisense direction in correspondence of promoters of other genes. Their expression influences promoter activity by means of specific regulatory function. PaRNA can act as a scaffold for proteins involved in processes regulating gene expression, as chromatin remodeling or transcription. RNA-induced transcriptional silencing (RITS) is the mechanism of transcriptional interference achieved by siRNA directed against chromatin-associated RNAs, as paRNAs. In this chapter, it is described how to detect interaction between siRNA and paRNA in cell nuclei and investigate siRNA capability to destabilize paRNA-protein complex.

Key words noncoding RNA, directional RT-PCR, biotin pull down, RNA immunoprecipitation

1 Introduction

Several deep sequencing experiments showed that genome is pervasively transcribed, giving rise to a heterogeneous class of molecules, long noncoding RNAs (lncRNAs), which can originate both from intergenic regions and in association with transcribed genes [1, 2]. That last category can be transcribed in the antisense direction with respect to the coding gene (gene-antisense ncRNAs) [3, 4] or can be transcribed in sense or antisense direction in correspondence of promoting regions; in this case they are named promoter-associated RNA (paRNA) [5–7]. Several examples of paRNAs playing a role in the modulation of adjacent gene transcription are present in literature [8, 9]. PaRNAs represent an additional layer of complexity of the transcriptional regulation in adaptation to particular cell context. A diffuse mechanism is the sequence-specific recruitment of remodeling chromatin proteins, inducing long lasting silencing of neighbor genes. It is the case of many tumor suppressor genes that inhibit normal cellular growth and which are often epigenetically silenced in cancer. Many oncosuppressor genes have nearby antisense RNAs. For instance, p15, a cyclin-dependent kinase, presents an antisense partner, p15AS,

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_11,

which induces its stable and irreversible silencing, both in *cis* and in *trans*, through alteration of H3 methylation pattern [10]. Many natural antisense transcripts (NAT) are reported also for p21 in the EST database. The degradation of one of them by siRNA reverts the suppressive H3K27me3 mark at the p21 promoter resulting in gene reactivation [11].

When paRNAs are transcribed in promoting regions of protein coding genes, they can directly interact with proteins involved in RNA transcription, as general transcription factors or RNA polymerase.

Here, they sustain transcriptional machinery activity and their expression is frequently correlated to the level of expression of the adjacent coding gene.

It is the case of the oncogene c-MYC, a transcription factor often overexpressed in tumors. A ncRNA transcribed in correspondence of Myc promoter overlaps the transcription starting site (TSS). It represents a trail for RNA pol II and general factors, in particular TFIIB, to form the preinitiation complex (PIC) and start Myc transcription correctly. The important function exerted by that paRNA is clearly demonstrated by the complete loss of transcriptional efficiency of c-MYC promoter when a siRNA is directed against its paRNA. The impairment in transcription is dependent on the binding between the siRNA and the paRNA but not on the degradation of the last one. The interference with paRNA-RNA pol II binding destabilizes the constitution of PIC and impairs MYC transcription initiation [12]. The ncRNA associated with Myc promoter is an example of a general class of molecules. PaRNA expression is tightly controlled to assist the right transcription of protein coding genes, influencing them both positively or negatively through the interaction with multiple proteins.

For this reason, paRNAs can be targeted by RNA-induced transcriptional silencing (RITS). SiRNA-paRNA binding can alter paRNA functionality and in consequence neighbor gene expression. General rules [13] useful to design siRNA for post transcriptional gene silencing must be followed also when siRNA targets a paRNA, but particular attention must be paid to strand specificity.

Once siRNAs is loaded by RISC complex, the double-strand RNA must be unwound and the passenger strand is degraded in the 3'–5' direction. The remaining strand is the guide strand that binds RNA target. The thermodynamic stability of siRNA ends determines the choice of guide strand, which leads target recognition. Indeed, unwinding of dsRNA starts from the less thermostable siRNA end. During siRNA design, the thermostability of its extremities must be considered to efficiently target the right RNA molecule [14]. This is especially true to target noncoding RNAs that often overlap transcripts coming from the opposite strand. In this case, if a siRNA is not fully strand specific, it is not simply

ineffective but can interfere with another transcript, complicating experimental readout.

In this chapter, it is described how to prove the correct interaction between siRNA and paRNA of interest. SiRNA with biotinylated sense or antisense strand allows pulling down the bound paRNA, which can be further analyzed. Since paRNAs often overlap another transcript on the opposite direction, either mRNAs or other ncRNAs, directional RT-PCR must be performed to discriminate the transcript directionality. If a signal is amplified after gene-specific retrotranscription with forward primer but not with reverse primer, antisense transcription occurs in the analyzed region; otherwise, it is sense transcription if signal is amplified after retrotranscription with reverse primer but not with forward primer. A signal can be obtained also in both reactions. In this case bidirectional transcription occurs.

PaRNA function is frequently related to its binding with specific proteins. SiRNA-paRNA recognition can impair that interaction. Destabilization of paRNA-protein interaction by RITS can be assessed by RNA ChIP immunoprecipitation, a technique that allows enriching nuclear RNA bound to a given protein.

2 Materials

Prepare all solutions using ultrapure water (prepared by purifying deionized water to attain a sensitivity of 18 MΩ cm at 25 °C) and analytical grade reagents. When RNAse free conditions are necessary, all buffers must be prepared with RNAse-free reagents and DEPC-treated water. Prepare and store all reagents at 4 °C temperature (unless indicated otherwise).

2.1 siRNA Transfection

1. Mammalian cells.
2. Cell culture medium.
3. Annealed siRNA, 100 μM.
4. Sense and antisense single strand biotin-siRNA, 100 μM (*see* **Note 1**).
5. Sense and antisense single strand unlabeled siRNA, 100 μM.
6. Annealing buffer 5×: 50 mM Tris–HCl, 100 mM NaCl, pH 7.5.
7. Lipofectamine (Invitrogen) (*see* **Note 2**).
8. Reduced serum medium.

2.2 Biotin-siRNA Pull-Down

1. Formaldehyde 37 % (v/v).
2. Glycin 10× 1.25 M.
3. DNA extraction kit.
4. PBS.

5. Modified Lysis buffer: NP-40 (0.5 % v/v), 300 mM NaCl, 20 mM Hepes, 2 mM $MgCl_2$.
6. Streptavidin agarose beads.
7. Washing buffer 2×: 10 mM Tris–HCl, 1 mM EDTA, 0.5 M NaCl, pH 8.0.
8. Elution buffer: 10 mM Tris–HCl, 1 mM EDTA, 2.0 M NaCl, pH 6.0.
9. DNAse I (RNAse-free).
10. DNAse I buffer 2×: 50 mM $MgCl_2$, 10 mM $CaCl_2$.

2.3 Directional RT-PCR

1. TRIzol® (Life technologies) (*see* **Note 3**).
2. Chloroform.
3. Isopropanol.
4. Ethanol 75 %.
5. One-step RT PCR kit.
6. Primers.
7. Tris-borate-EDTA (TBE) buffer: 89 mM Tris-base, 89 mM boric acid, 2 mM EDTA, pH 8.0.
8. Agarose gel: 2 % in TBE.
9. RNAse-free tips and eppendorfs.

2.4 RNAChIP Immunoprecipitation

1. Nuclei extraction buffer: 100 mM Tris–HCl, 100 mM NaCl, 2.5 mM $MgCl_2$, 40 μg/ml digitonin, (pH 7.4) (*see* **Note 4**).
2. RIP lysis buffer: 50 mM Hepes, 1 mM EDTA, Triton-X 1 % (v/v), (pH 8.0).
3. RNAse inhibitor (*see* **Note 5**).
4. Protease Inhibitor Cocktail.
5. DNAse I (RNAse-free).
6. DNAse I buffer 2×: 50 mM $MgCl_2$, 10 mM $CaCl_2$.
7. Protein G Magnetic Beads.
8. ChIP grade (or IP grade) antibody.
9. Binding buffer: 50 mM Hepes, Triton-X 0.5 % (v/v), 25 mM $MgCl_2$, 5 mM $CaCl_2$, 20 mM EDTA (pH 8.0), (pH 8.0).
10. FA500 buffer: 50 mM Hepes, Triton-X 1 % (v/v), deoxycholic acid 0.1 % (w/v), 500 mM NaCl, 1 mM EDTA, (pH 8.0).
11. LiCl buffer: 0.25 M LiCl, Triton-X 1 % (v/v), deoxycholic acid 0.5 % (w/v), 10 mM Tris–HCl, (pH 8.1).
12. TES buffer: 10 mM Tris–HCl, 10 mM NaCl, 1 mM EDTA, (pH 8.1).
13. RIP elution buffer: 100 mM Tris–HCl, 10 mM EDTA, SDS 1 % (w/v), (pH 7.8).

14. 5 M NaCl.
15. Proteinase k.
16. TRIzol® (Life technologies), Chloroform, Isopropanol, Ethanol 75 % for RNA extraction.
17. Primers and One-step RT PCR kit for directional RT-PCR.

2.5 Equipment

1. Agarose Electrophoresis Equipment.
2. Dark Reader transilluminator.
3. Sonicator.
4. Magnetic stand.
5. PCR thermal cycle system.
6. Spectrophotometer with high accuracy and reproducibility.
7. Centrifuge with capability to reach 12,000 × *g* at 4 °C.

3 Methods

3.1 siRNA Transfection

1. Seed cells at 70 % confluence in 15 cm dishes (*see* **Note 6**).
2. Anneal sense and antisense strands, coupling one biotin labeled strand and one not, to obtain duplex siRNA with biotinylated-sense or antisense strand. Mix equal amount of sense and antisense strands at concentration 100 μM each and annealing buffer 5×, to have final buffer concentration 1× and duplex siRNA 40 μM.
3. Heat the siRNA at 70 °C for 10 min and then let it to cool at room temperature (*see* **Note 7**).
4. Remove medium from the plate and replace with 6 ml reduced serum medium.
5. Prepare the transfection mix, in two separate sterile RNAse-free tubes:

   30 μl (1.2 nmol) of duplex siRNA in 3 ml of reduced serum medium.

   60 μl of Lipofectamine in 3 ml of reduced serum medium.
6. After 5 min, mix siRNA and Lipofectamine and let stand for 15 min at room temperature, than add it to the cells.
7. After 4–6 h replace transfection mix with fresh complete medium and incubate cells for 24–48 h at 37 °C 5 % CO_2.

3.2 Biotin siRNA Pull-Down

1. Collect the cells, resuspend in complete medium and count.
2. Crosslink cells with 1 % formaldehyde for 10′ at room temperature (*see* **Notes 8, 9**).
3. Stop the crosslinking with glycine 1× for 5′ at room temperature.

4. Wash twice with PBS.
5. Extract DNA according to kit manufacturer's instructions (*see* **Note 10**).
6. Elute DNA in 200 μl of kit elution buffer and save 30 μl as input.
7. Mix the left sample with equal amount of modified lysis buffer.
8. Prewash streptavidin-agarose beads with 300 μl of modified lysis buffer.
9. Mix the beads with sample and incubate on rotation for 2 h at 4 °C (*see* **Note 11**).
10. Wash three times with washing buffer.
11. Add 100 μl of elution buffer.
12. Adjust input volume to 100 μl with elution buffer.
13. Heat inputs and pull down samples at 65 °C for 1 h (*see* **Note 12**).
14. Adjust $MgCl_2$ to 25 mM and $CaCl_2$ to 5 mM.
15. Add DNAse 10 U and incubate at 37 °C for 10 min.
16. Add 2 mM EDTA and heat at 65 °C for 10 min (*see* **Note 13**).

3.3 Directional RT-PCR

1. Add 1 ml of Trizol to the sample.
2. Proceed with chloroform extraction immediately; otherwise keep the lysate at −80 °C (*see* **Note 3**).
3. Check RNA concentration and quality (*see* **Note 14**). Prepare RNA dilutions at 40 ng/μl concentration in RNAse-free water.
4. Prepare RT-PCR reaction mix using One-step RT PCR kit according to supplier's instructions. Assemble RT-PCR reaction on ice, keeping every reagent cold and in RNAse-free conditions; use only RNAse-free tips and tubes. Prepare master mix on ice, adapting what is indicated in Table 1, for the necessary number of reactions. Aliquot the master mix in the 0.2 μl tubes and then add the RNA.

Table 1
Master mix for directional RT-PCR

	Starting concentration	Volume (μl)	Final concentration
Mix	2×	12.5	1×
Primer for retrotranscription	10 μM	0.5–1	0.2–0.4 μM
Enzyme mix		0.5	
Water		to 19–21.5	
RNA	40 ng/μl	2.5–5	100–200 ng

5. Retrotranscribe RNA in a primer-specific manner at 50 °C for 30′ (*see* **Note 15**).
6. Denature cDNA and inactivate retrotranscriptase activity at 94 °C for 8′.
7. Cool the samples to 4 °C. Spin down the tubes quickly and then add 0.5–1 μl of the remaining primer (10 μM) to each tube, one by one. Vortex briefly and spin down the tubes.
8. Perform PCR reaction, according to the following program: Hot Start Taq activation 95 °C/2′, Denaturing step 95 °C/45″ (*see* **Note 16**), Annealing step 55–60 °C/30", Elongating step 72 °C/15″-60″. Repeat from denaturing to elongating step for 30–35 cycles. Final elongation step 72 °C/10′. Annealing temperature is 5 degrees lower than the primers' melting temperature while elongation time is 60″/kb amplicon length (*see* **Note 17**).
9. To exclude contamination by residual genomic DNA amplification, run a control reaction in parallel, assembling it as described at Subheading 3.3, **step 3**, but avoid adding any primers during retrotranscription step. Both primers must be added to the reaction after retrotranscriptase activity inactivation. Positive control where genomic DNA (100 ng) is amplified must be included (*see* **Note 18**).
10. Check the presence of the amplicon on a 2 % agarose minigel, running at 90 V for 30′.

3.4 RNA Chromatin Immunoprecipitation

1. Collect cells transfected with annealed unlabeled siRNA or negative control siGL3 and count them (*see* **Note 19**).
2. Crosslink five million of cells as already described at Subheading 3.2 at **step 2**. Stop crosslinking with glycine at final concentration 3× (*see* **Note 20**).
3. Wash cells in PBS twice.
4. Resuspend two million cells in 1 ml of nuclei extraction buffer.
5. Incubate 30′ on ice.
6. Centrifuge 2000 × *g* 8′ (*see* **Note 21**).
7. Resuspend nuclei in 500 μl of RIP lysis buffer containing protease inhibitor cocktail at final concentration 1× and 50 U of RNAse inhibitor.
8. In order to physically share chromatin, which paRNAs are bound to, sonicate the lysate for five to eight times for 10″ each, at low energy, paused by 30″ between one pulse and the other. Keep the lysate on ice during all the sonication procedure, to limit heat production (*see* **Note 22**).
9. Centrifuge the lysate at 14,000 × *g* for 10′ at 4 °C. Recover the supernatant and transfer it in a clean tube. It is supposed to

contain all nuclear RNA, also bound to chromatin, since sharing made it soluble. The pellet will contain insoluble debris and can be discarded (*see* **Note 23**).

10. Dilute the sample with equal amount of DNAse I buffer 2×, with 50 U RNAsin and 1× protease inhibitor cocktail added, to have following final buffer composition, 0.5 % TritonX, 25 mM $MgCl_2$, 5 mM $CaCl_2$.
11. Add 30 U DNAse I RNAse-free and incubate for 15′ at 37 °C.
12. Stop the digestion, adding EDTA to a final concentration of 20 mM.
13. Add 20 μl of magnetic beads and the antibody of interest (8–10 μg), or the same amount of negative control IgG, to lysate and incubate for 2 h at 4 °C on a rotating platform (*see* **Notes 24–26**).
14. Put the samples in the magnetic stand and remove carefully the supernatant.
15. Wash the beads twice with binding buffer, twice with FA500 buffer, twice with LiCl buffer, and twice with TES buffer. Each wash cycle consists of 5′ on rotation with 1 ml of cold buffer; put the samples in the magnetic stand and remove carefully all supernatants between one wash and the following.
16. Add 75 μl of RIP elution buffer to recover the complex protein-RNA that was captured by the antibody (*see* **Note 27**).
17. Adjust NaCl to 200 mM and treat with 20 μg of proteinase k for 1 h at 42 °C and 1 h at 65 °C.
18. Then extract RNA with Trizol/chlorophorm (*see* **Note 28**).
19. Perform directional RT-PCR to evaluate if paRNA of interest interacts to the immunoprecipitated protein directly (*see* **Note 29**). Use primers and PCR conditions indicated in Subheading 3.3. Retrotranscribe the same volume of RNA, regardless the RNA concentration, to compare paRNA enrichment in sample where the protein of interest was immunoprecipitated with respect to IgG, and then compare the RNA enrichment in cells treated or not with siRNA.

4 Notes

1. Biotin labeled siRNA can be bought from several companies, asking for specific biotin labeling. The biotin moiety must be placed at 5′ of the strand of interest, in order to not interfere with the opening of the dsRNA and the selection of the guide strand by RIST complex.
2. Even if many transfection agents can be found on the market, Lipofectamine® is the best choice to deliver siRNA into the nucleus.

3. TRIzol® reagent is extensively referred to give high-quality, intact RNA from many kinds of biological materials. This reagent is very useful because it can be directly added to cells or to cellular lysates, with high yield of intact RNA extracted. It efficiently preserves RNA from degradation. RNA lysates diluted in TRIzol® can also be saved at −80 °C for even 1 year before performing chloroform extraction, without any loss of material.
4. Digitonin is a steroidal saponin useful to mild permeabilize cellular membrane. It tends to precipitate in solution, so it should be solubilized in water, freshly, at the concentration of 0.5 % (w/v), and saved at 4 °C not more than one week.
5. Due to low stability of RNA and low amount of paRNAs, RNAse Inhibitor must be added to any solution that will be in contact to the sample.
6. Right cell confluence must be established by the operator, since it depends on cell model. The efficiency/toxicity ratio is influenced by cell confluence: the lower is ratio between number of cells and siRNA the higher is the transfection efficiency, the higher is the ratio the lower is the cellular toxicity of transfection.
7. Faster cooling of duplex siRNA may give wrong or incomplete annealing.
8. Crosslinked cells are highly sticky. The cross linking of attached cells in the dish can reduce the recovery of cells, for this reason can be preferred cross linking cells already detached and pulled in a tube.
9. Crosslinking conditions can be adjusted depending on cell model: if recovery is low, it can be improved increasing crosslinking time, even if exceeding crosslinking can result in hard cell lysis or a specific results.
10. This step is meant to recover genomic DNA to which paRNA are bound, getting rid, in the end, of proteins that take part to chromatin structure.
11. Increasing the incubation time can improve RNA recovery but on the other side increase the risk of RNA degradation.
12. Elution buffer contains high salt concentration, and this allows reverting crosslinking during 65 °C incubation. Inefficient reversal of crosslinking results in hard amplification during following RT-PCR. In case, increasing 65 °C incubation time can improve the result. Exceeding 2 h incubation may result in RNA degradation.
13. 65 °C incubation in the presence of EDTA 2 mM inactivates DNAse, preventing RNA degradation.
14. If RNA is contaminated by alcohol, phenols, etc. (Ads 260/230 < 1.6) proceed to reprecipitate your RNA.

15. If the noncoding RNA to amplify is a sense transcript, retrotranscribe with reverse primer, if it is an antisense transcript with the forward primer. The directionality of ncRNA must be previously determined by two RT-PCR performed in parallel, retrotranscribing with either forward or reverse primer.
16. The denaturation step is longer than standard PCR protocol, since often paRNAs are rich in GC, and stronger denaturation is required to get good amplification.
17. The amount of RNA, the primers concentration, and the number of cycles to perform must be experimentally established, because they depend on the amount of ncRNA and on the efficiency of primers amplification.
18. If contamination of genomic DNA is evident, go back to the mother RNA, digest with DNAse for 15′ at rt, then reprecipitate RNA and repeat the directional RT-PCR reaction.
19. The best timing for RNA-ChIP experiment after siRNA transfection depends on the paRNA under investigation. Generally, the interference with ncRNA localized onto the chromatin requires longer time, then common post transcriptional gene silencing. However, the best timing must be experimentally validated.
20. The amount of cells to use for a single RNA chromatin immunoprecipitation depends on the abundance of the protein we want to immunoprecipitate and on the efficiency of the antibody. For this reason, it must be empirically determined in the particular cell model used. In general, start from five million cells/IP.
21. The efficiency of nuclei extraction depends on the cell model, so check efficiency of cytoplasmic membrane permeabilization and integrity of nuclei by Trypan blue staining.
22. The first time the experiment is performed in a certain cellular model, the optimal sonication must be experimentally determined. Lysates (100 μl) must be sonicated with increasing number of pulses. Digest DNA, following the protocol as described at Subheading 3.4, **steps 9–12**. Add 200 mM NaCl and proteinase k and incubate 2 h 65 °C. RNA (20 μl) must be loaded onto a 2 % RNAse-free agarose gel, in denaturing loading buffer, to see the size of the shared RNA. A smear between 500 and 200 nucleotides is acceptable.
23. This is a safe break point. Put the samples at −80 °C, but proceed with processing the day after, immediately.
24. The correct amount of antibody must be empirically evaluated by a previous titration. However, usually 8–10 μg of ChIP grade antibody should work properly.

25. Magnetic beads tend to precipitate, so to distribute them properly in each sample, mix well the vial content, cut off the edge of the tip, and pipet slowly.
26. The operator must include a negative control, as IgG, an antibody that does not recognize any protein and in consequence of that, should not coprecipitate RNA. A positive control may also be useful, even if sometimes it is hard to identify a protein surely bound to RNA molecule under inspection. A proper control is often the RNA polymerase, if it is known transcribing that particular class of RNA molecules.
27. Add to the input 25 μl of 3× RIP elution buffer, to adjust the buffer component concentration.
28. In order to have absolutely pure RNA, after phenol/chloroform extraction, an additional step of DNA digestion can be performed and then RNA can be reprecipitated.
29. Since the signal of the noncoding RNA detected bound to an immunoprecipitated protein can be very weak, a step of nested PCR can help to obtain a more robust result.

References

1. Morris KV, Mattick JS (2014) The rise of regulatory RNA. Nat Rev Genet 15(6):423–437
2. St Laurent G, Wahlestedt C, Kapranov P (2015) The landscape of long noncoding RNA classification. Trends Genet 31(5):239–251
3. Lapidot M, Pilpel Y (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep 7(12):1216–1222
4. Werner A (2013) Biological functions of natural antisense transcripts. BMC Biol 11:31
5. Lepoivre C, Belhocine M, Bergon A et al (2013) Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics 14:914
6. Goodrich JA, Kugel JF (2006) Non-coding-RNA regulators of RNA polymerase II transcription. Nat Rev Mol Cell Biol 7(8):612–616
7. Ntini E, Järvelin AI, Bornholdt J et al (2013) Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol 20(8):923–928
8. Wang X, Arai S, Song X et al (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature 454(7200):126–130
9. Martianov I, Ramadass A, Serra Barros A et al (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445(7128):666–670
10. Yu W, Gius D, Onyango P et al (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature 451(7175):202–206
11. Morris KV, Santoso S, Turner AM et al (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. PLoS Genet 4(11):e1000258
12. Napoli S, Pastori C, Magistri M et al (2009) Promoter-specific transcriptional interference and c-myc gene silencing by siRNAs in human cells. EMBO J 28(12):1708–1719
13. Birmingham A, Anderson E, Sullivan K et al (2007) A protocol for designing siRNAs with high functionality and specificity. Nat Protoc 2(9):2068–2078
14. Chen PY, Weinmann L, Gaidatzis D et al (2008) Strand-specific 5′-O-methylation of siRNA duplexes controls guide strand selection and targeting specificity. RNA 14(2):263–274

Chapter 12

RNA-FISH to Study Regulatory RNA at the Site of Transcription

Marta Soler, Raquel Boque-Sastre, and Sonia Guil

Abstract

The increasing role of all types of regulatory RNAs in the orchestration of cellular programs has enhanced the development of a variety of techniques that allow its precise detection, quantification, and functional scrutiny. Recent advances in imaging and fluoresecent in situ hybridization (FISH) methods have enabled the utilization of user-friendly protocols that provide highly sensitive and accurate detection of ribonucleic acid molecules at both the single cell and subcellular levels. We herein describe the approach originally developed by Stellaris®, in which the target RNA molecule is fluoresecently labeled with multiple tiled complementary probes each carrying a fluorophore, thus improving sensitivity and reducing the chance of false positives. We have applied this method to the detection of nascent RNAs that partake of special regulatory structures called R loops. Their growing role in active gene expression regulation (Aguilera and Garcia-Muse, Mol Cell 46:115–124, 2012; Ginno et al., Mol Cell 45:814–825, 2012; Sun et al., Science 340:619–621, 2013; Bhatia et al., Nature 511:362–365, 2014) imposes the use of a combination of in vivo and in vitro techniques for the detailed analysis of the transcripts involved. Therefore, their study is a good example to illustrate how RNA FISH, combined with transcriptional arrest and/or cell synchronization, permits localization and temporal characterization of potentially regulatory RNA sequences.

Key words RNA-FISH, R loop, Nascent transcript, Intron, Vimentin

1 Introduction

RNA fluorescent in situ hybridization (FISH) represents a powerful tool for dissection of epigenetic mechanisms, and helps our understanding of the creation of specific transcription patterns [1]. Nuclear RNA FISH permits the detection of noncoding RNAs and primary transcripts at gene loci to assay for the transcriptional status of the particular gene of interest [2]. Here, we describe the RNA FISH method based on Stellaris® protocol that allows simultaneous detection, localization, and quantification of individual RNA molecules at the subcellular level using wide field fluorescence microscopy (www.biosearchtech.com/stellarisprotocols), and apply it to characterize the R loop-forming antisense transcript *VIM-AS1* [3–7] (Fig. 1). *VIM* and *VIM-AS1* transcripts represent

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_12,

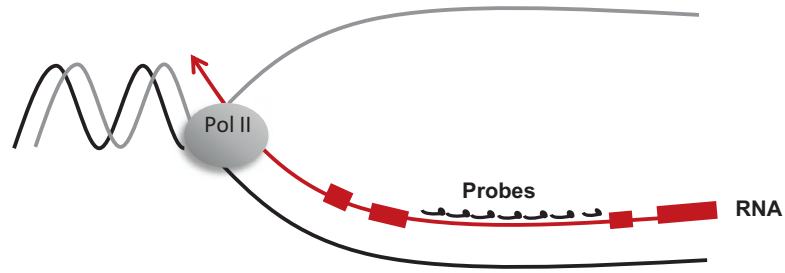


Fig. 1 Stellaris® RNA FISH method comprises multiple (25–40) oligonucleotides coupled to a fluorescent label (e.g., TAMRA, FAM, Quasar®570, etc.). The oligos may anneal to processed RNAs or intronicregions (to detect the nascent transcriptor specially stabilized intronicregions). If necessary, FISH detection can be carried out under conditions of transcriptional arrest and/or synchronized cells

a head-to-head sense/antisense pair, where the first intron of the antisense transcript positively regulates the expression of *VIM* through the formation of an R loop structure. Two pools of 48 probes tiling either the first intron of *VIM-AS1* RNA (starting just downstream of the R loop region to avoid signal interference) or the first intron of *VIM* pre-mRNA are designed following Stellaris RNA FISH Probe. *VIM-AS1* probes are coupled to TAMRA and *VIM* probes to FITC reporter dyes. Transcript abundance can vary depending on cell cycle and thus we include a synchronization protocol that allows analysis at different cell cycle phases. Also, when characterizing non-exonic sequences transcriptional arrest might be necessary to distinguish between nascent transcripts that are rapidly processed and stable, regulatory sequences within pre-mRNA species. Thus, in this protocol, we describe the use of RNA-FISH in synchronized MCF10A cells under Actinomycin D treatment to study regulatory, noncoding transcripts.

2 Materials

Prepare all solutions using Braun sterile water for injections or ultrapure water (prepared by purifying deionized water to obtain a sensitivity of 18 MΩ cm at 25 °C) and analytical grade reagents.

2.1 For Cell Culture, Synchronization, and Actynomicin D Treatment

MCF10A breast cells culture media: Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (DMEM/Ham's F-12 medium) supplemented with 20 ng/mL EGF, 500 ng/mL hydrocortisone, 10 mg/mL insulin, and 100 ng/mL cholera toxin. Store at 4 °C (*see* **Note 1**).

Cell culture dishes 100 × 200 mm style.

12-well cell culture plates.

12 mm diameter × 0.17 thickness coverslip.

Thymidine stock solution: weigh 264 mg of Thymidine (T1895, Sigma) and resuspend with 1 mL of 1M NaOH, to obtain a final concentration of 1 M. Filter the solution through a 0.2 μm filter under laminar flow hood, and keep at room temperature protected from light (*see* **Note 2**).

Actinomycin D stock solution: dissolve 2 mg of Actinomycin D (AD) in 2 mL of Dimethyl Sulfoxide (DMSO), to obtain a stock solution of 1 mg/mL. Store at −20 °C (*see* **Note 3**).

2.2 For RNA FISH

When performing Stellaris RNA FISH, it is imperative to limit RNA degradation. Please ensure that all consumables and reagents are Nuclease-free.

1. Stellaris RNA FISH Probes (*see* **Note 4**).
2. TE buffer: 10 mM Tris–HCl, 1 mM EDTA pH 8.0.
3. Fixation buffer: 3.7 % (v/v) formaldehyde in PBS 1×. Store at room temperature (RT).
4. Permeabilization buffer: 70 % ethanol in nuclease-free H_2O.
5. Dextran sulfate solution (50 % w/v).
6. 20× Saline-Sodium Citrate (SSC): 3 M NaCl, 300 mM sodium citrate.
7. Deionized formamide 100 %.
8. Hybridization buffer: dextran sulfate 5 % (w/v) (*see* **Note 5**), 2× Saline-Sodium Citrate (SSC), deionized formamide 10 % (v/v), in Nuclease-free H_2O. Store at −4 °C (*see* **Note 6**) (*see* **Note 7**).
9. Wash buffer: deionized formamide 10 % (v/v), 2× SSC, in Nuclease-free H_2O. Store at room temperature.
10. Nuclear stain buffer: 4′,6-diamino-2-fenilindol (DAPI) dissolved in wash buffer at 50 ng/mL.
11. Mounting medium (*see* **Note 8**).

2.3 Imaging and Analysis Components

1. Microscope: AxioObserver Z1 (ZEISS).
2. Filters: DAPI (EX365, BS395, EM445/50), eGFP (EX470/40, BS495, EM525/50), and Cy3 (EX550/25, BS570, EM605/70) (or those appropriate for the fluorophores used).
3. Objective: Pan-APOCHROMA 63× NA 1:4 oil.
4. Fluorescence lamp HXP 120 C.
5. Acquisition: AxioCam MRm camera with cooled Monochrome CCD sensor and ZEN 2012 from Carl Zeiss software.
6. Fiji Is Just Image 1.5b software.

3 Methods

3.1 Cell Culture

Carry out all procedures under laminar flow hood. Cells are grown at 37 °C, in a humidified air - 5 % CO_2 incubator.

1. For each condition, place 5–8 sterile 12-mm round coverslips on 100-mm plastic culture dishes (*see* **Note 9**).
2. Plate 1 × 10^6 exponentially growing MCF10A breast cells in 10 mL culture media supplemented with DMEM/F-12. Allow to attach for about 8 h (*see* **Note 10**).

3.2 Synchronization

1. Synchronization of MCF10A cells is performed by double thymidine block [8] (*see* **Note 11**). When cells are completely attached, treat the cells at a final concentration of 2 mM thymidine, adding 20 μL of tymidine stock solution to 10 mL of culture medium, and let them for 14 h at 37 °C in the incubator (*see* **Note 12**). Remember the asynchronous condition without thymidine (*see* **Note 13**).
2. Remove the thymidine-containing medium and rinse the dishes twice with 1× PBS (*see* **Note 14**). Add 10 mL fresh medium and incubate dishes for 10 h at 37 °C.
3. Add 20 μL of thymidine stock solution to each plate, again at a final concentration of 2 mM, and incubate 14 h more at 37 °C.
4. Remove the thymidine-containing medium and rinse the dishes twice with 1× PBS.
5. Add 10 mL of fresh medium. This point will be considered as time 0 and cells are harvested at 0, 5, 6 h.
6. Monitor the progression of cell cycle to G2/M phase by flow cytometry: FACS. The cell cycle (distribution of cells in G0/G1, S, and G2/M phases) is assessed with Propidium Iodide (PI)-stained cells by flow cytometry. Approximately 1–2 × 10^6 cells are fixed in ice-cold 70 % ethanol overnight at −20 °C. Cells are subsequently washed and resuspended in 1× PBS, 1 % FBS. Cells are stained with 0.025 mg/mL PI, and treated with 0.05 mg/mL RNase for 30 min at 37 °C in the dark. The PI fluorescence is measured for individual cells using a FACS flow cytometer (FACS Calibur, Becton Dickinson, USA).

3.3 Treatment with Actinomycin D (AD)

1. Directly after 5 and 6 h of thymidine release, add 50 μL of actinomycin D stock solution or DMSO to each plate, for treatment or control condition, respectively. This gives a final concentration of 5 μg/mL of AD and 0.5 % DMSO (*see* **Note 15**) [9].
2. Incubate the cells for 30 min at 37 °C. From now on, carry out all procedures at room temperature (RT) and out of laminar flow hood unless otherwise specified.

3.4 Cell Fixation and Permeabilization

1. Transfer the coverslips containing adherent cells, synchronized and treated (AD or DMSO), from the 100-mm dishes to a 12-well plate, with the help of sterile forceps and a needle (*see* **Note 16**).
2. Wash each well twice with 1 mL of 1× PBS (*see* **Note 17**).
3. Add 1 mL of fixation buffer and incubate at room temperature for 10 min.
5. Remove fixation buffer by aspiration and wash twice with 1 mL of 1× PBS.
6. To permeabilize, immerse cells in 1 mL of 70 % (v/v) ethanol for at least 1 h at 4 °C (*see* **Note 18**). These are standard permeabilization conditions for most adherent cell cultures.

3.5 Hybridization

1. Reconstitute the dried probe stock: dissolve the dried oligonucleotide probe blend in 200 μL of TE buffer to create a probe stock at a total oligo concentration of 25 μM. Mix well by vortexing and spin down.
2. Warm the reconstituted probe solution to room temperature. Mix well by vortexing, then centrifuge briefly.
3. To prepare the hybridization buffer containing probes for more than one transcript, add 2.5 μL of each probe stock solution (e.g., 2.5 μL of VIM-FITC plus 2.5 μL VIM-AS1-TAMRA) to a final volume of 50 μL of hybridization buffer, and then vortex and centrifuge (this is enough for one coverslip). This creates a working solution with both probes at a concentration of 1.25 M. This solution will be used in **steps 6** and 7 (*see* **Note 19**).
4. Aspirate the 70 % ethanol off the coverslip.
5. Add 1 mL of wash buffer, and incubate at room temperature for 2–5 min.
6. To incubate the coverslips in a humid environment and prevent dehydration, use an empty plastic P1000 pipette tip box. Place a parafilm on the top grid and 2–3 cm of water at the bottom to create a humid chamber (Fig. 2).
7. Within this humidified chamber, dispense 50 μL of the hybridization buffer containing probe onto the Parafilm.
8. Gently transfer the coverslips, cells side down, onto the 50 μL drop of hybridization buffer containing probe (*see* **Note 20**). Incubate in the dark at 37 °C overnight (*see* **Note 21**).
9. Gently transfer the coverslips, cells side up, to a fresh 12-well plate.
10. Add 1 mL of wash buffer and incubate in the dark at 37 °C for 30 min.
11. Aspirate the wash buffer and add 1 mL of DAPI nuclear staining buffer to counterstain the nuclei.

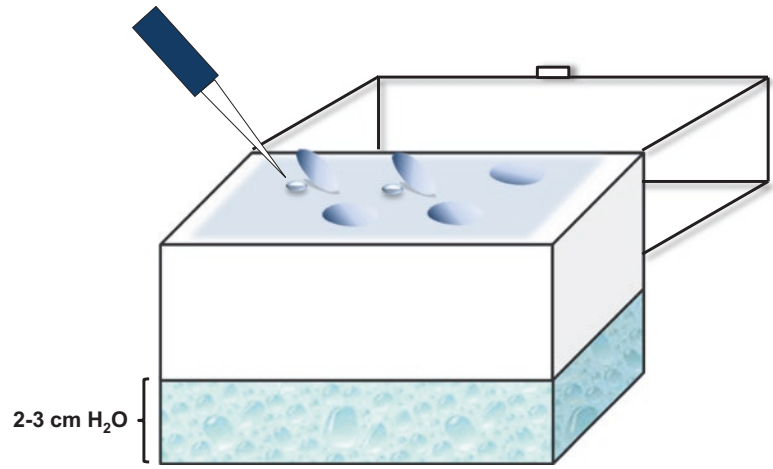


Fig. 2 Hybridization of fixed, permeabilized cells with FISH probes. Coverslips with cells attached are placed upside down onto a drop of hybridization buffer plus probe, in a tip box containing 2–3 cm of H_2O to maintain humidity

12. Incubate in the dark at 37 °C for 15–20 min.
13. Aspirate the DAPI staining buffer, and then wash twice with 1 mL of 2× SSC buffer. Incubate at room temperature for 2–5 min.
14. Add a small drop (approximately 6 μL) of Vectashield hardset mounting medium onto a microscope slide, and mount coverslip onto the slide, cells side down. Try to avoid bubbles between the slide and the coverslip (*see* **Note 22**).
15. Gently wick away excess antifade from the perimeter of the coverslip and allow to dry for 30 min to 1 h, and if necessary wipe away any dried salt off the coverslip with water.

3.6 Imaging and Analysis

1. The RNA FISH samples are imaged by using a wide-field fluorescent microscope AxioObserver Z1 (ZEISS), with a 63×/1.4 oil immersion objective (*see* **Note 23**).
2. Images are acquired with Axio Cam camera at a resolution of 1388 × 1040 (pixel size: 6.45 × 6.45 μm).
3. Images are analyzed by Fiji Is Just ImageJ software (Fig. 3).

4 Notes

1. Aliquote and store all components at −20 °C.
2. Prepare fresh for each experiment.
3. To avoid contact with the AD powder, inject the DMSO through the rubber stopper and into the AD vial with the help of a syringe. It is advisable to prepare aliquotes of approxi-

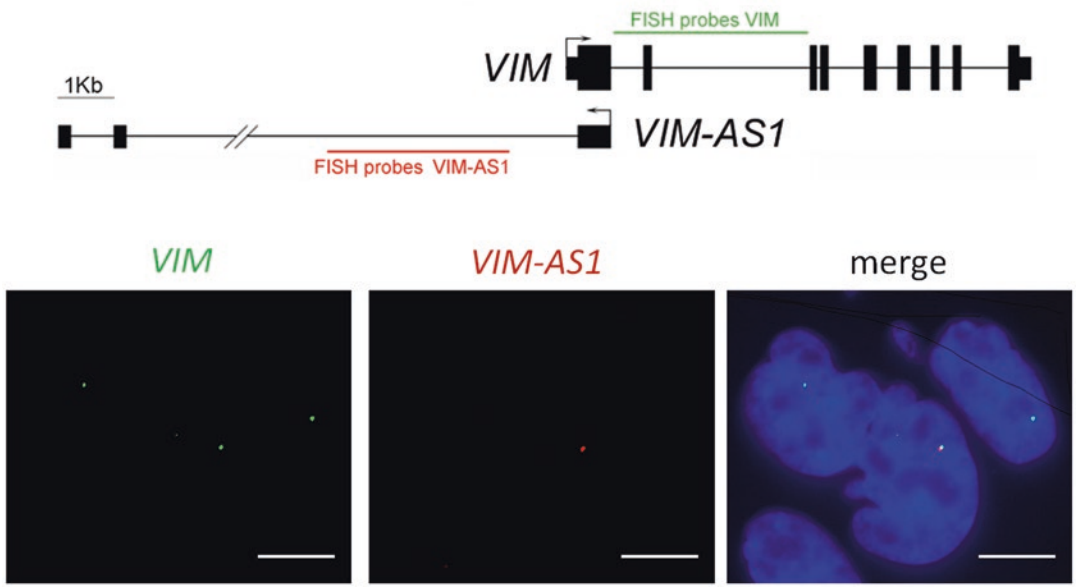


Fig. 3 RNA FISH detection shows presence of sense and antisense transcripts at the VIM locus. *Above*, RNA FISH probe design. Red probes tile the first intron of VIM-AS1, whereas green probes tile the first intron of VIM. *Below*, MCF10A cells were synchronized and released, fixed after 6 h and stained for RNA FISH. Blue represents the DAPI staining to reveal the nucleus. The images show only one allele per cell is transcriptionally active, and the antisense transcript is present in ~30–40 % of cells. Scale bar, 10 μm

mately 500 μL in dark eppendorf tubes to avoid too many freeze/thaw cycles.

4. Stellaris FISH probes are sold under license from Rutgers, The State University of New Jersey, and may be used under its patent rights for Research Use Only. Stellaris RNA FISH probes are shipped dry and can be stored frozen or at 4 °C in this state. Dissolved probe mix should be subjected to a minimum number of freeze-thaw cycles. For daily and short-term use of dissolved probe mix, storage at 4 °C in the dark for up to a month is recommended. For storage longer than a month, we recommend aliquoting and freezing probes in the dark at −20 °C.
5. Dextran sulfate concentration may be increased up to 10 % (w/v) to reduce nonspecific signal.
6. When using formamide, first warm the solution to room temperature before opening to avoid oxidation; also, care must be taken when using formamide (i.e., use in the hood, wear protection, etc.) because it is a suspected carcinogen and teratogen and is readily absorbed through the skin.
7. For long-term storage, hybridization buffer can be aliquoted in 1 mL tubes and stored at −20 °C.
8. Vectashield Hardset (H1400; Vector Laboratories) or Prolong Gold (Invitrogen) work well for most Stellaris probes.

9. It is important to plan the experiment taking into account synchronization time points (for instance, 0, 5, and 6 h), and an asynchronous condition to establish the settings in the cell cycle analysis. Also, it must be taken in consideration many time points will be combined with the Actinomycin D or control treatment.
10. Plate exponentially growing cells at 30–50 % confluence. MCF10A cell line attaches very quickly, but other cell types might need to be seeded at 450,000–500,000 cells/plate and left to attach for about 24 h.
11. Double-thymidine block as described here works well for most cell types with doubling time of between 24 and 30 h. For other specific cell types, optimization of the time of incubation with thymidine might be necessary. See reference [8] for further assistance in setting up synchronization protocols.
12. When thymidine is added to the medium, slightly move the plate to mix the thymidine and obtain a homogeneous concentration. Remember that the culture medium may become dark due to basification produced by NaOH.
13. Asynchronous culture serves to establish the cell settings.
14. Thoroughly rinse all surfaces of the dish to prevent carryover of thymidine.
15. In the particular case of *vimentin*, we are interested in just the G2-phase of the cell [10]. 5 and 6 h after the release of thymidine is when the MCF10A cell line reaches mostly G2/M.
16. This protocol is set for a 12-well plate system. To adapt this protocol to your preferred system, volumes should be adjusted accordingly.
17. Be careful not to take off cells when vacuuming the medium.
18. Cells can be stored at 4 °C in 70 % ethanol up to one week before hybridization. Seal the plate to prevent ethanol evaporation.
19. Prepare the required volume for the number of cover slips over 12 % of excess.
20. Be very careful not to confuse on which side the cells are. You can use sterile forceps and needle to move the coverslips.
21. A good option is to cover the box with aluminum foil.
22. Gradually dropping the covers on the Vectashield hardset mounting medium drop.
23. For best results, samples mounted with Vectashield Mounting Medium should be imaged on the same day.

Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad (MINECO, grant number SAF2014-56894-R), the Fundació La Marató de TV3 (grant number 20131610), and the Asociación Española contra el Cáncer-Junta de Barcelona. We are grateful to Dr. Manel Esteller for his advice and support during the preparation of this manuscript.

References

1. Namekawa SH, Lee JT (2011) Detection of nascent RNA, single-copy DNA and protein localization by immunoFISH in mouse germ cells and preimplantation embryos. Nat Protoc 6:270–284
2. Lawrence JB, Singer RH (1985) Quantitative analysis of in situ hybridization methods for the detection of actin gene expression. Nucleic Acids Res 13:1777–1799
3. Sun Q, Csorba T, Skourti-Stathaki K et al (2013) R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. Science 340:619–621
4. Bhatia V et al (2014) BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. Nature 511:362–365
5. Aguilera A, Garcia-Muse R (2012) R loops: from transcription byproducts to threats to genome stability. Mol Cell 46:115–124
6. Ginno PA, Lott PL, Christensen HC et al (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45:814–825
7. Boque-Sastre R, Soler M, Oliveira-Mateos C et al (2015) Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. Proc Natl Acad Sci U S A 112:5785–5790
8. Jackman J, O'Connor PM (2001) Methods for synchronizing cells at specific stages of the cell cycle. In: Wiley J and Sons (ed). Curr Protoc Cell Biol. Chapter 8: Unit 8.3. doi:10.1002/0471143030.cb0803s00
9. Guy AL, Taylor JH (1978) Actinomycin D inhibits initiation of DNA replication in mammalian cells. Proc Natl Acad Sci U S A 75:6088–6092
10. Rittling SR, Baserga R (1987) Functional analysis and growth factor regulation of the human vimentin promoter. Mol Cell Biol 7:3908–3915

Chapter 13

Detection and Characterization of R Loop Structures

Raquel Boque-Sastre, Marta Soler, and Sonia Guil

Abstract

R loops are special three stranded nucleic acid structures that comprise a nascent RNA hybridized with the DNA template strand, leaving a non-template DNA single-stranded. More specifically, R loops form in vivo as G-rich RNA transcripts invade the DNA duplex and anneal to the template strand to generate an RNA:DNA hybrid, leaving the non-template, G-rich DNA strand in a largely single-stranded conformation (Aguilera and Garcia-Muse, Mol Cell 46:115–124, 2012).

DNA-RNA hybrids are a natural occurrence within eukaryotic cells, with levels of these hybrids increasing at sites with high transcriptional activity, such as during transcription initiation, repression, and elongation. RNA-DNA hybrids influence genomic instability, and growing evidence points to an important role for R loops in active gene expression regulation (Ginno et al., Mol Cell 45, 814–825, 2012; Sun et al., Science 340: 619–621, 2013; Bhatia et al., Nature 511, 362–365, 2014). Analysis of the occurrence of such structures is therefore of increasing relevance and herein we describe methods for the in vivo and in vitro identification and characterization of R loops in mammalian systems.

Key words R loop, CpG island, RNA-FISH, Vimentin, DRIP, Bisulfite sequencing, In vitro transcription

1 Introduction

R loops (DNA:RNA hybrids and the associated single-stranded DNA) have been traditionally associated with threats to genome integrity, making some regions of the genome more prone to DNA-damaging and mutagenic agents. Initially considered to be rare byproducts of transcription, over the last decade accumulating evidence has pointed to a new view in which R loops form more frequently than previously thought. The R loop field has become an increasingly expanded area of research, placing these structures as a major threat to genome stability but also as potential regulators of gene expression. Special interest has arisen as they have also been linked to a variety of diseases, including neurological disorders and cancer, positioning them as potential therapeutic targets [1–5].

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_13,

The interest in R loop research expanded as the widespread presence of R loops over the 5′ region of genes was revealed, and a correlation with CpG-rich, nonmethylated promoters in the human genome was apparent [6]. These genomic regions are GC-rich and have a strong positive GC skew (template strand having an excess of C vs. G residues), and a link between R loop formation and protection from methylation was suggested [7]. Also, different studies suggest that transcription termination in RNA Pol II-driven genes may represent an additional and relevant case in which R loops could form naturally [8, 9], with R loops directly implicated in transcriptional termination of some human genes [10]. R loops have also been shown to play a role in the regulation of ncRNAs [3], and a link with the transcription machinery has been established [11].

The growing interest to understand the biology of R loop formation and their roles as regulators of gene expression have enhanced the development of in vivo and in vitro techniques that help in their study. R loop formation depends on a number of features [12, 13], among them high G/C content, negative supercoiling and DNA nicks, and only experimental proof can finally assess the actual ability of a given genomic region to form stable R loops when transcribed. We herein present three types of approaches to identify and characterize long (~1 Kb) stretches of RNA:DNA hybrids with the potential to regulate the expression of nearby genes. We first present an in vitro protocol to establish the ability of a RNA transcript to form an R loop structure as it is transcribed from its plasmidic DNA template. Moving to the cellular context, we then describe two protocols to identify, on the one hand, the bisulfite-sensitive single-stranded DNA associated with R loops, and, on the other hand, the RNA:DNA hybrid as it is recognized by the specific S9.6 antibody. Altogether these techniques help in the precise localization and characterization of stable R loops that might form under different cellular contexts.

1.1 In Vitro R Loop Formation

In vitro R loop formation assays is a relatively fast and easy way to establish the potential ability of a particular DNA region to form R loops in the event of a polymerase transcribing through it. As such, it should be envisioned as a primary test that gives support to other in vivo evidences. It takes advantage of the shift in mobility of the plasmid DNA that will occur when most of the DNA molecules remain bound to the RNA transcript. We use plasmids containing the fragment of interest (~1 Kb in length) downstream of a phage promoter to in vitro synthesize the encoded RNA. The ability of this RNA to remain attached to the DNA template and form stable R loops will be shown by the aberrant mobility of the plasmid in agarose electrophoresis together with the sensibility to RNaseH digestion (which specifically resolves R loops by degrading the

RNA strand of DNA:RNA hybrids). In addition, presence of RNA in the shifted band can be confirmed by radiolabeling the transcription reaction and analyzing an autoradiography of the gel. R loop-forming transcripts are typically rich in Guanosine nucleotides, and thus the antisense transcript (that can be easily tested in parallel if the sequence is cloned in between two different phage promoters) will not display the same mobility shift and will lack radioactive signal.

1.2 Detection of R loops In Vivo Through Immunoprecipitation of DNA:RNA Hybrids (DRIP)

R loops that form in living cells can be effectively immunoprecipitated by the S9.6 antibody, which recognizes DNA:RNA hybrids with high specificity. Chromatin from a variety of cell types under particular growing conditions can be isolated in native conditions, fractionated with restriction enzymes, and pulled down with S9.6 antibody [14]. The retrieved material can then be analyzed by qPCR to assess the enrichment of individual genomic fragments. A negative control in which the sample is processed in parallel but where RNaseH is added is necessary to assess specificity of the pulldown. This experimental approach, together with transcriptomic data and GC skew determination in the studied DNA region, can help confirm the presence of a stable, local R loop, suggestive of a potential regulatory structure.

1.3 Detection of R Loops In Vivo by Native Bisulfite Sequencing

An alternative approach for in vivo detection of R loops takes advantage of the fact that, under native conditions, the single-stranded DNA within the R loop is susceptible to sodium bisulfite-induced conversion of cytosines to uracil (read as thymidine by DNA Polymerases), whereas a double-stranded DNA is commonly resistant to such conversion. In this section, we describe some modifications of the method initially designed by Yu and colleagues [15]. The native bisulfite treatment of genomic DNA is followed by PCR with primers specific to one half of the predicted R loop-forming region, the PCR products are then ligated and the resulting clones sequenced. By using this method, and in contrast to a conventional denaturing treatment of DNA, we are able to distinguish the initial template strand in the first round of PCR based on the sequence of the final PCR product. That is, we can distinguish between the G-rich or the C-rich strand. Changes C to T reveal a single-strandedness in the C-rich strand, whereas changes G to A reveal single-strandedness in the G-rich strand. For reliable results, it is necessary to sequence a high number of clones (at least 20 clones). Only sequences that feature long stretches (>100 bp) of uninterrupted G-to-A conversions represent potential R loop-forming regions. This technique allows us to have a qualitative indication of the existence of a local R loop. The following protocol refers to the detection of an R loop near the transcription start site of *vimentin* gene in SW480 cells [11].

2 Materials

2.1 In Vitro R Loop Formation

1. DNA plasmid template with potential R loop forming fragment downstream of a bacteriophage promoter (typically, T7, T3, or SP6) (*see* **Note 1**).
2. 10× RNA Polymerase buffer (0.4 M Tris–HCl, pH 8.0, 60 mM $MgCl_2$, 100 mM Dithiothreitol, 20 mM spermidine).
3. 100 mM $MgCl_2$ in H_2O.
4. Mix of rNTPs (10 mM each ATP, CTP, GTP, 1 mM UTP) in H_2O.
5. [α-32P]-UTP (10 μCi/μl).
6. Bacteriophage RNA Polymerases (e.g., T7, T3, SP6, at 10 U/μl, Roche) to drive transcription of the downstream sequence.
7. 10 mg/ml RnaseA in H_2O.
8. RNaseH (5 U/μl).
9. Phenol:Chloroform:Isoamyl Alcohol 25:24:1, saturated with 10 mM Tris, pH 8.0, 1 mM EDTA.
10. 10× DNA Loading buffer (0.2 % w/v Orange G, 50 % glycerol in H_2O).
11. 10× TBE (1 M Tris base, 1 M Boric Acid, 0.02 M EDTA).
12. High-grade agarose for nucleic acid electrophoresis.
13. 1 Kb DNA Ladder.
14. Horizontal electrophoresis tank and power supply.
15. DNA gel stain (Ethidium bromide or SYBR®Safe, ThermoFisher).
16. 3 MM Whatman filter paper, gel dryer, autoradiography films, and cassettes.

2.2 Detection of R Loops In Vivo Through Immunoprecipitation of DNA:RNA Hybrids (DRIP)

2.2.1 For DNA Isolation and Restriction Enzyme Digestion:

1. 10×10^6 cells growing in culture under the particular condition of study.
2. Proteinase K (10 mg/ml).
3. DNA lysis buffer: 10 mM Tris pH 8.0, 1 mM EDTA, SDS 0.5 % (w/v).
4. 5 M NaCl.
5. Isopropanol, 70 % Ethanol, absolute Ethanol.
6. TE buffer: 10 mM Tris, 1 mM EDTA.
7. HindIII, EcoRI, XbaI, BamHI, and 10× restriction buffer (*see* **Note 2**).
8. RNaseH (5 U/μl).
9. 1 % agarose gel.
10. Phenol:Chloroform:Isoamyl Alcohol 25:24:1, saturated with 10 mM Tris, pH 8.0, 1 mM EDTA.

2.2.2 For Immunoprecipitation:

1. IP buffer: Triton X-100 0.05 % (v/v) in phosphate-buffered saline (PBS).
2. Anti-DNA-RNA Hybrid [S9.6] Antibody (Kerafast).
3. 1 mg/ml Normal Mouse IgG.
4. Magnetic beads anti-Rabbit IgG.

2.2.3 For qPCR:

1. Specific oligonucleotides (forward and reverse sequences to obtain 80–200 nts amplicon from within predicted R loop region. Use also control primers outside R loop-forming region).
2. SYBR® Green PCR Master Mix.
3. Applied Biosystems 7900HT Fast Real Time PCR (or equivalent equipment).

2.3 Detection of R Loops In Vivo by Native Bisulfite Sequencing

1. SW480 human colon adenocarcinoma cell line.
2. DMEM containing stable glutamine.
3. Heat-inactivated Fetal Bovine serum.
4. Lysis buffer: 10 mM Tris pH 8.0, 1 mM EDTA, SDS 0.5 % (w/v).
5. 5 M NaCl.
6. Isopropanol, 70 % EtOH.
7. Braun sterile water for injections.
8. RNase inhibitor.
9. Bisulfite conversion kit to study DNA methylation. We recommend EZ DNA Methylation-Gold™ Kit (Zymo Research), which in our hands gives the best conversion efficiency.
10. PCR reagents.
11. 2 % Agarose gel.
12. Gel and PCR Clean-up kit.
13. pGEM®-T Vector System (Promega).
14. DH5-α *E. coli* competent cells.
15. LB Broth: 10 g Tryptone, 10 g NaCl, 5 g yeast extract and bring volume up to 1 l with deionized water. Autoclave.
16. Agar ampicillin plates: prepare LB medium as above, but agar at 15 g/L before autoclaving. After autoclaving, cool, and add ampicillin (50 μg/mL), and pour into petri dishes. Store at 4 °C.
17. 20 mg/ml X-Gal in DMSO.
18. 0.1 M IPTG in H_2O.
19. DNA miniprep kit.
20. T7 primer (5′-TAATACGACTCACTATAGGG-3′).
21. X-terminator: *BigDye® XTerminator*™ (Applied Biosystems).

3 Methods

3.1 In Vitro R Loop Formation

3.1.1 In Vitro Transcription Reactions

For each transcript to be assayed (usually two per plasmid, corresponding to sense and antisense directions) carry out the in vitro transcription in duplicate (one reaction will contain +RNaseH and the other one –RNaseH). Also, include a negative control for each plasmid (- polymerase, this will indicate basal migration rate of the pure DNA plasmid) (*see* Fig. 1).

1. Prepare transcription reactions at room temperature, adding last the RNA Polymerase (*see* **Note 3**). Mix 2 μl of 10× buffer with 0.5 μl of 100 mM $MgCl_2$, 500 ng of plasmid DNA, 0.2 μl of the rNTP mix, 0.3 μl of [α-32P]-UTP, 1 μl of the RNA Polymerase, and H_2O up to 20 μl.
2. Incubate for 45 min at 37°C, followed by inactivation 15 min at 70 °C.

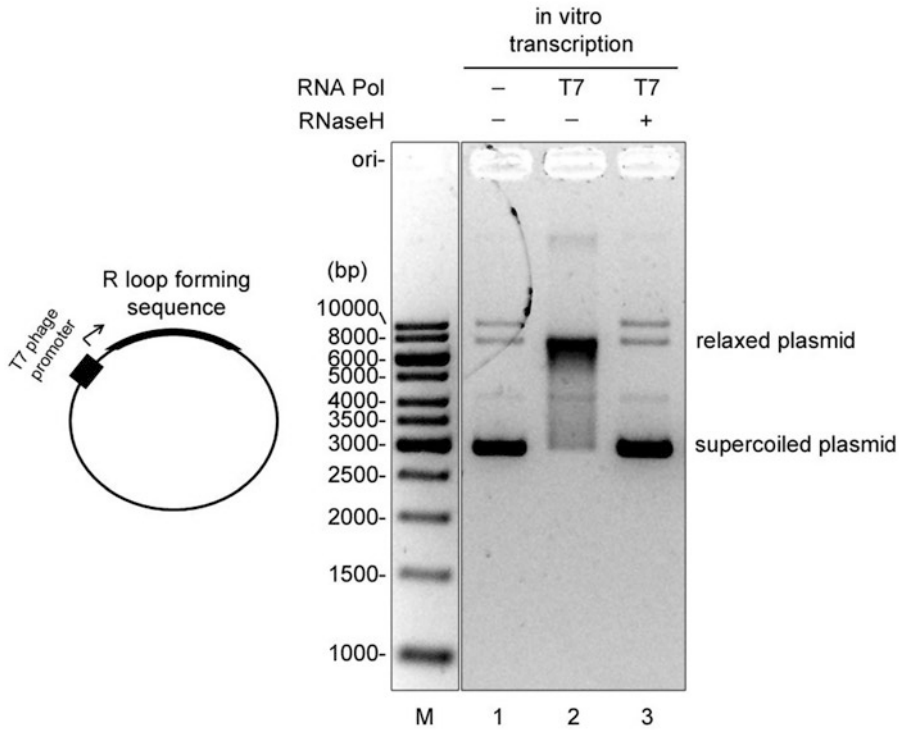


Fig. 1 In vitro R loop formation assay. The potential R loop forming region (~1 Kb long) is cloned downstream of a bacteriophage promoter, and the in vitro transcription reaction is carried out. After resolving in a 1 % agarose gel, DNA bands are stained with ethidium bromide. Shifted bands are indicative of a stable interaction between the template DNA plasmid and the transcribed RNA (*lane* 2). To confirm the presence of RNA:DNA hybrids in the shifted band, the reactions are incubated in the presence of the bacterial recombinant RNAseH (*lane* 3). As a control, a mock reaction without polymerase is also analyzed (*lane* 1). *M*, 1 Kb DNA ladder

3.1.2 RNAse Digestion and Gel Electrophoresis

1. Add to each tube 2.5 μl of RNase A diluted 1/5000 (*see* **Note 4**).
2. Add 0.5 μl of RNaseH to the "+RNaseH" tubes.
3. Incubate for 30 min at 37 °C.
4. Add H_2O to 50 μl, extract with one volume of Phenol:Chloroform:Isoamyl Alcohol and add 5 μl of 10× DNA Loading buffer.
5. Load a 1 % agarose gel in 1× TBE, together with DNA ladder (*see* **Note 5**).
6. Run at 120 V constant until dye reaches bottom of gel (~12–15 cm).
7. Remove gel from tank, stain gel with DNA dye by soaking it in 100 ml of 1× TBE containing the dye in a separate tray for at least 1 h. Visualize on a UV-illuminator.
8. To reveal the RNA content of the shifted bands, put the gel on a Whatman paper and cover with cling film, vacuum-dry the gel for 1 h at 80 °C, and expose to an autoradiography film overnight.

3.2 Detection of R Loops In Vivo Through Immunoprecipitation of DNA:RNA Hybrids (DRIP)

3.2.1 DNA Isolation and Restriction Enzyme Digestion

1. Scrape or trypsinize cells and pellet in an eppendorf tube. Add 500 μl of DNA lysis buffer, and 5.5 μl Proteinase K. Incubate at 37 °C, a minimum of 4 h or overnight.
2. Add 250 μl of 5 M NaCl to get a final concentration of 1.5 M. Mix and spin at maximum speed for 15 min.
3. Keep supernatant (if the phases do not separate well, spin again 15 min top speed), and add 0.8 Volumes of isopropanol (600 μl).
4. Mix gently by inversion until the DNA precipitate (the "jellyfish") appears. If it is very abundant, fish it out and proceed with the 70 % Ethanol washes.
5. Spin for 10 min at top speed. Wash pellet with 1 ml 70 % Ethanol, then dry it and resuspend in TE buffer (in enough volume to avoid too much viscosity). Measure DNA concentration at nanodrop.
6. Digest 40–50 μg of genomic DNA with the mix of restriction enzymes in 1× restriction buffer in 100 μl (*see* **Note 6**). Incubate for 2 h at 37 °C. Sample should become much less viscous.
7. Split the DNA sample in two and add to one tube 5 μl of RNaseH. Digest for a further 2 h at 37 °C.
8. Load 2 μl of DNA in 1 % agarose gel to check size (*see* **Note 7**).
9. Do a phenol/chloroform extraction, precipitate with 2.5 volumes of Ethanol, and resuspend in 40–50 μl of IP buffer. Measure amount at nanodrop.

3.2.2 Immunoprecipitation (IP) with S9.6 Antibody

1. Use 10 μg of DNA per IP, and keep a small amount (500 ng–1 μg) of this same DNA to be used later as input material. Incubate DNA with 10 μl of S9.6 monoclonal antibody or control IgG in a total volume of 500 μl for 2 h at 4 °C. Standards experiments include, per each DNA sample: S9.6 –RNaseH, S9.6 +RNaseH, IgG –RNaseH, IgG+RNaseH.
2. Wash magnetic beads three times in IP buffer and add 30 μl of beads to each sample. Rock for 2 h at 4 °C (see **Note 8**).
3. Wash three times with 700 μl of IP buffer, rocking for 5 min at 4 °C for each wash.
4. Treat beads with Proteinase K: add 2 μl of Proteinase K and 48 μl of PBS to the beads, and incubate for 3 h at 50 °C. Recover DNA by phenol/chloroform extraction and ethanol precipitation. Resuspend in 100 μl of TE buffer.
5. Analyze recovered fragments by qPCR: use 25 ng of input DNA and 5 μl of the immunoprecipitated DNA (straight or diluted if needed) and 0.5 μM each primer in 10–12 μl of SYBR® Green reaction mix. Analyze on a real-time PCR machine, using the ΔCt method comparing the signal from the immunoprecipitated material to the signal from the input DNA (correcting by dilutions used).

3.3 Detection of R Loops In Vivo by Native Bisulfite Sequencing

3.3.1 Cell Culture and DNA Extraction

1. SW480 human colon adenocarcinoma cell lines are cultured in DMEM containing stable glutamine, and supplemented with 10 % (v/v) heat-inactivated FBS (). Cells are grown at 37 °C in a humidified atmosphere of 5 % (v/v) CO_2 and 95 % (v/v) air.
2. Pellet cells (from 10 cm dishes), when cells are at 80 % of confluence (*see* **Note 9**). Proceed with DNA extraction as described above ("Detection of R loops in vivo through immunoprecipitation of DNA:RNA hybrids (DRIP)") and measure at nanodrop. It is very important not to freeze the sample and continue with the protocol (*see* **Note 10**).

3.3.2 Bisulfite Treatment

1. Transfer 5 μg of the extracted DNA in 20 μl of water into 0.5 μl tubes.
2. Add 1 μl of RNaseOUT to each sample.
3. Add 130 μl of CT Conversion Reagent (from *EZ DNA Methylation-Gold™ Kit*). This reagent will convert the unprotected cytosines (unmethylated in single-stranded DNA) to uracil.
4. Incubate in a thermocycler overnight at 37 °C (12–16 h).
5. Purify samples following the kit protocol (*EZ DNA Methylation-Gold™ Kit*).
6. Elute with 25 μl of hot water.

3.3.3 Primer Design

We design primers for the first half of the R loop forming region, corresponding to a ~500 bp fragment. Forward primer was designed upstream of the predicted R loop forming region, and the reverse primer was designed in the middle of the structure. We introduce in the reverse primer changes G to T, which finally we will read as G to A changes on the plus strand of the sequence if the minus strand is originally in a single strand conformation (*see* Fig. 2).

3.3.4 Sequencing PCR

1. PCR with the specific primers for the R loop region: use 5–8.4 μl of DNA, 0.2 μM primers, 1.5 mM $MgCl_2$, and 1.5 U Taq Polymerase in 1× buffer in a total volume of 15 μl. A typical PCR program would be (might need some optimization depending on primer sequence and polymerase used): initial denaturing step 7 min at 95 °C, 32 cycles (30 s at 95 °C, 30 s at 58 °C, 30 s 72 °C), final elongation step 7 min at 72 °C.
2. Run the samples in a 2 % agarose gel, cut out bands, and purify the samples with Gel and PCR Clean-up kit, eluting with 15 μl of hot water.

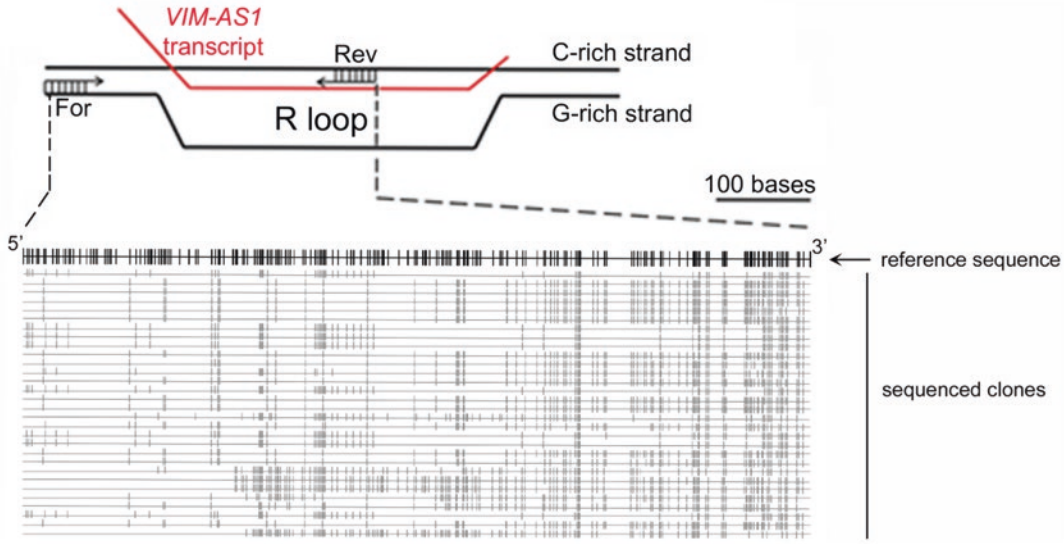


Fig. 2 In vivo detection of R loop following native bisulfite treatment of DNA. *Above*, the diagram depicts the RNA:DNA hybrid (in this example, formed between with *VIM-AS1* transcript near the transcription start site of *VIM*) and the displaced, single-stranded, minus DNA strand. The PCR is performed with a forward, native (For) oligonucleotide and a reverse, converted (Rev) oligonucleotide, which takes into account the C-to-U changes (C-to-T after PCR) that appear only on the minus DNA strand following native bisulfite conversion. *Below*, PCR amplification and sequencing of 20 clones corresponding to the first half of the C skew-containing region. The reference top line depicts every G position (vertical lines), and the clones sequenced are represented below as light gray, with every G-to-A change on the plus strand (indicative of an unprotected minus strand) indicated by a vertical line

3.3.5 Ligation

1. Use the pGEM-T ligation system, mixing 5 μl of buffer, 1 μl pGEM vector, 1 μl T4 Ligase, and 3 μl of purified DNA. Incubate for 2–3 h at room temperature or overnight at 4 °C.

3.3.6 Transformation

1. Add 90 μl of competent bacteria to the ligation reaction and incubate for 30 min on ice.
2. Heat-shock the bacteria for 1 min at 42 °C, leave 3 min on ice. Add 900 μl of LB medium and incubate for 1 h at 37 °C in agitation.
3. Centrifuge tubes for 4 min at 800 *g*, discard two-thirds of the supernatant, and resuspend bacteria in the remaining medium.
4. Plate some or all of the transformation onto 10 cm agar plate containing the appropriate antibiotic and 20 μl 20 mg/ml X-Gal, 20 μl 0.1 M IPTG. Incubate overnight at 37 °C.

3.3.7 DNA Miniprep and Sequencing of DNA Insert

1. Pick white colonies (*see* **Note 11**) and put in 1 ml of LB medium with Ampicillin at 50 μg/mL. We pick approximately 30–40 colonies per DNA band and use 96-well plates. Incubate overnight at 37 °C with agitation.
2. Purify plasmid DNA with the miniprep kit, following the provider's instructions.
3. Set up sequencing reactions following standard protocols by using T7 primer, clean up with X-terminator kit *BigDye® XTerminator™* (Applied Biosystems), and ABI 377 sequencer (Applied Biosystems).
4. Align and compare sequences to reference genome with BioEdit Sequence Alignment Editor.

4 Notes

1. For example, we use the pGEM®-T family of vectors (Promega), which provides a quick method for TA cloning. The presence of both T7 and SP6 promoters at both ends of the insert allows the comparison between sense and antisense transcription and thus is useful to conduct a negative control in parallel. High-quality plasmids are needed (e.g., column-purified).
2. This restriction enzyme combination might vary depending on the specific locus of study. It is important to make sure the region of interest and control regions do not contain any of the restriction sites used.
3. Ensure the 10× buffer is at room temperature before mixing with the plasmid DNA, to avoid precipitation due to the presence of spermidine in the buffer.
4. This step is necessary to remove most of the free RNA. Optimization of the amount used might be needed depending on the RNase batch.

5. It is important to use material completely free of DNA dye traces before and during the electrophoresis. DNA intercalating agents such as Ethidium bromide will highly interfere with R loop formation and might alter DNA mobility during electrophoresis.
6. We use each enzyme at 0.5 U/μl final concentration. Choose a reaction buffer in which your combination of enzymes works at their optimal efficiency.
7. DNA should appear as a smear ranging in size from ~800 bp to a few thousands Kbases.
8. After this incubation, the supernatant can be kept as unbound fraction to estimate the efficiency of the pulldown.
9. These cells can be overexpressed with RNASE H (which specifically degrades the RNA in RNA/DNA hybrid) to see differences when R loop is forming or when it is depleted. We also performed shRNAs and *Antisense LNA™ GapmeRs (Exiqon)* to depleted de R loop forming gene, in our case the antisense transcript of *VIM* gene (*VIM-AS1*).
10. It is very important not to freeze the sample to avoid RNA degradation.
11. White colonies indicate gene fragment insertion in the plasmid. Sometimes, very short fragments of DNA do not effectively disrupt β-galactosidase activity and may be seen as light blue colonies.

Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad (MINECO, grant number SAF2014-56894-R), the Fundació La Marató de TV3 (grant number 20131610), and the Asociación Española contra el Cáncer-Junta de Barcelona. We are grateful to Dr. Manel Esteller for his advice and support during the preparation of this manuscript.

References

1. Aguilera A, Garcia-Muse R (2012) R loops: from transcription byproducts to threats to genome stability. Mol Cell 46:115–124
2. Ginno PA, Lott PL, Christensen HC et al (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45:814–825
3. Sun Q, Csorba T, Skourti-Stathaki K et al (2013) R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. Science 340:619–621
4. Bhatia V et al (2014) BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. Nature 511:362–365
5. Santos-Pereira JM, Aguilera A (2015) R loops: new modulators of genome dynamics and function. Nat Rev Genet 16:583–597
6. Ginno PA et al (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45:814–825
7. Ginno PA et al (2013) GC skew at the 5′ and 3′ ends of human genes links R-loop formation

to epigenetic regulation and transcription termination. Genome Res 23:1590–1600

8. Skourti-Stathaki K, Proudfoot NJ, Gromak N (2011) Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol Cell 42:794–805
9. Hatchi E, Skourti-Stathaki K, Ventz S et al (2015) BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. Mol Cell 57:636–647
10. Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ (2014) R-loops induce repressive chromatin marks over mammalian gene terminators. Nature 516:436–439
11. Boque-Sastre R, Soler M, Oliveira-Mateos C et al (2015) Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. Proc Natl Acad Sci U S A 112:5785–5790
12. Roy D, Lieber MR (2009) G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Mol Cell Biol 29:3124–3133
13. Roy D et al (2010) Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: a nick can serve as a strong R-loop initiation site. Mol Cell Biol 30:146–159
14. Yu K, Roy D, Huang FT et al (2006) Detection and structural analysis of R-loops. Methods Enzymol 409:316–329
15. Yu K, Chedin F, Hsieh CL et al (2003) R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat Immunol 4:442–451

Part IV

Therapeutic Potential of Promoter-Associated RNAs Targeting

Chapter 14

Induction of Transcriptional Gene Silencing by Expression of shRNA Directed to *c-Myc* P2 Promoter in Hepatocellular Carcinoma by Tissue-Specific Virosomal Delivery

Mohammad Khalid Zakaria, Debi P. Sarkar, and Parthaprasad Chattopadhyay

Abstract

Double-stranded RNA-mediated transcriptional gene silencing (TGS) has shown promising results over posttranscriptional gene silencing (PTGS) due to its long term and heritable nature. Various research groups have shed light on different mechanisms by which TGS operate. Some of these include histone modification, DNA methylation, or restriction of RNA polymerase binding onto the target gene's promoter. This serves as an added advantage since permanent *c-Myc* inactivation is critical for suppressing hepatocellular carcinoma (HCC). Inability to target cancer cells specifically, without affecting the normal cells, has been one of the biggest drawbacks of an effective cancer therapy. Therefore, we aimed to overcome this barrier by first generating tumor-specific transcriptional units expressing TGS inducing shRNAs against *c-Myc*'s P2 promoter only in neoplastic liver cells. Secondly, we coupled this TGS inducing system with Sendai fusion virosomes for liver-specific delivery to minimize nonspecific side effects in vitro.

Key words Transcriptional gene silencing, Virosomal delivery, shRNAs

1 Introduction

Majority of hepatocellular carcinoma (HCC) related deaths are due to its asymptomatic nature as the observable symptoms are only manifested in the later stage of disease progression [1]. Major limitations in HCC targeted therapy are efficiency and selectivity of gene transfer. Targeting of neoplastic cells has been limited due to the lack of ways which could specifically and effectively deliver the therapeutic modalities to cancer cells [2, 3]. Such discrimination between normal and transformed cells would prevent systemic cytotoxicity that is often observed in patients being treated with chemotherapy. Also, HCC-targeted gene/chemo therapies are often limiting since the patients have impaired hepatic functions [4]. For HCC, α-fetoprotein (AFP) promoter could help in

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_14,

achieving specific gene therapy [20–22]. Several vehicular trafficking systems have been utilized for delivery of either siRNA or chemotherapeutic drugs. Some of these include modified liposomes [5], nanoparticles [6], receptor-ligand based systems [7], etc.

Sendai virus is a para-influenza virus with negative sense RNA as its genome. The ease of its culture, inability to infect humans and high efficiency of liver-specific delivery has proved its importance in the targeted therapy [8]. Sendai fusion (F) virosomes are the derivatives of Sendai virus consisting of the surface F protein and reconstituted lipid bilayer. The surface F protein interacts specifically with the asialoglycoprotein receptors (ASGPRs) present on the hepatocyte's surface [9]. The lack of surface hemagglutinin neuraminidase (HN) protein prevents its nonspecific interaction [9]. We have shown earlier that the Sendai virosome could deliver the entrapped cargo specifically and efficiently in liver cells both in in vitro and in vivo systems [3, 10]. The cargo is delivered directly into the cytoplasm and hence escapes degradation by the lysosomal pathway. This enhances the half-life as well as the expression of the cargo within the target cells [11].

Targeting regulatory region of gene, by dsRNA, has been shown to induce transcriptional gene silencing (TGS) [12]. For induction of TGS, shRNA against specific sites on the regulatory region of a gene, such as polymerase or transcription factor binding region, has shown good outcomes [13–15]. TGS has shown promising results for longer durations by involving heritable epigenetic changes leading to prominent transcriptional repression [13, 15, 16]. This could serve as an advantage since permanent *c-Myc* inactivation is critical for suppressing neoplastic liver cells [17]. ME1a1 binding site, upstream of *c-Myc* P2 promoter, has shown to be critical for the maintenance of heterochromatin architecture [18]. *c-Myc* P2 promoter seems to be the dominant one as it is responsible for transcribing majority of the *c-Myc* transcripts [19]. Since up-regulated *c-Myc* has a role in HCC development [20], we designed shRNA encompassing ME1a1 site to repress *c-Myc* P2 promoter.

First, we cloned AFP promoter from −230 to +2 bp (*see* **Notes 1** and **2**) relative to the transcription start site (TSS) and fused nuclear factor kappa β (NF-κB) enhancer upstream to it since NF-κB is critically involved in HCC [21]. We hypothesized that NF-κB linked AFP promoter could strongly drive the downstream transgene expression without losing tumor-specific expression (*see* **Note 3**). *c-Myc* shRNA, targeting the ME1a1 binding site, was cloned downstream to the NF-κB–AFP promoter fusion construct for the induction of TGS. Furthermore, we created another similar fusion construct having AFP enhancer (−4 kb to −3.3 kb) upstream to the AFP promoter to enhance expression of *c-Myc* shRNA. Next, we utilized liver-specific Sendai F virosomes for packaging and delivery of these TGS-inducing shRNA constructs to induce

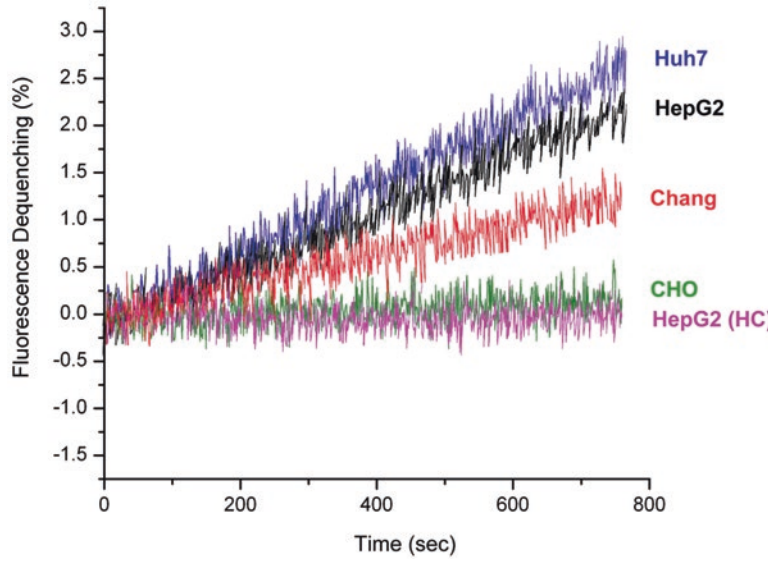


Fig. 1 Model of Sendai F-Virosome mediated transcriptional gene silencing of *c-Myc* promoter in neoplastic liver cells

specific cell death of neoplastic liver cells. The Sendai virus was lysed and reconstituted containing only F-proteins on its surface and having the AFP enhancer/promoter *c-Myc* shRNA plasmid entrapped within its aqueous core (Fig. 1).

2 Materials

1. BamHI and EcoRI restriction endonucleases.
2. dTNPs (10 mM), primers 20 pmoles, Taqpolymerase, Taqpolymerase buffer (50 mM KCl, 10 mM Tris–Cl pH 9. Triton-X-100 0, 0.1 % (v/v)).
3. $MgCl_2$ (1.5 mM).
4. *c-Myc* shRNA sense and antisense oligonucleotides (Table 1).
5. Annealing Buffer (10 mM Tris-HCL (pH 8.0), 1 mM EDTA, 100 mM NaCl).
6. NF-κB oligonucleotides (Table 1).
7. Inactivated Hemaglutinating virus of Japan (Sendai virus-Z strain).
8. 10–11 days old embryonated chicken eggs.
9. PBS (10 mM) pH 7.4.
10. Octadecyl rhodamine beta chloride (R18).
11. 3 mM DTT.
12. Dialysis bag (cutoff MW 12,000–14,000).

Table 1
List of sequences used in the study

Name	Sequence (5′–3′)
c-Myc shRNA sense	**GATCCGAACGGAGGGAGGGATCGCGCTTTTTCAAGAGAAGCGCGATCCCTCCCTCCGTTCTTA**
c-Myc shRNA antisense	**AGCTTAAGAACGGAGGGAGGGATCGCGCTTCTCTTGAAAAAGCGCGATCCCTCCCTCCGTTCG**
c-Myc Scrambled sense	**GATCCAGCGGTCGAGACGTGGCGGAGATTTTCAAGAGATCTCCGCCACGTCTCGACCGCTTTA**
c-Myc Scrambled antisense	**AGCTTAAGCGGTCGAGACGTGGCGGAGATCTCTTGAAAATCTCCGCCACGTCTCGACCGCTG**
NF-KB Enhancer	GGGAATTTCCGGGAATTTCCGGGAATTTCCGGGAATTTCC
c-Myc Bisulfite PCR Outer Reverse	TTCCAATACAAAATACCC
c-Myc Bisulfite PCR Outer Forward	TTTGAGAGGGAGTAAAAG
c-Myc Bisulfite PCR Inner Forward	CGCCAGGGTTTTCCCAGTCACGACGTGGGAAAAGAAAAAAG
c-Myc Bisulfite PCR Inner Reverse	AGCGGATAACAATTTCACACAGGAAACCAAAAACRAAACCCC
c-Myc ChIP Forward	CCCACCGGCCCTTTATAATGC
c-Myc ChIP Reverse	GCTCGGGTGTTGTAAGTTCC
Chromosome 16 ChIP Forward	GTCTCTTTCTTGTTTTTAAGCTGGG
Chromosome 16 ChIP Reverse	TGAGCTCATTGAGACATTTGG

13. Triton X-100 10 % (v/v), maintaining the final percentage of triton between 2 and 5 %.
14. Cell lines—HepG2, Huh7, and Chang Liver.
15. Ca^{2+} (1.5 mM).
16. 0.02 M Tris-buffered saline (pH 8.3), 10 mM Tris-buffered saline, pH 7.4.
17. Dithiothreitol (DTT) is dissolved in 0.02 M Tris-buffered saline (pH 8.3) at 30 mM. Final concentration should be 3 mM.
18. EDTA 2 mM.
19. SM2-Biobeads.
20. 26-G needle.
21. Cell culture: Dulbecco modified eagles's medium (DMEM) and Opti MEM (Invitrogen), 6-well culture plates and T175 flasks.
22. EZ-ChIP™ (Millipore).
23. EpiTect® Bisulfite kit (Qiagen).

3 Methods

Carry out the process at room temperature unless specified

3.1 Generation of AFP Promoter/ Enhancer Driven TGS Inducing c-Myc shRNA System

1. Extract genomic DNA from HepG2 cells by conventional phenol chloroform method.
2. Amplify AFP promoter (−230 to +2 bp), AFP Enhancer (700 bp; −4 to −3.3 kb) by PCR. Use promoters depending on the type of required tissue specificity (*see* **Note 4**). We amplified the region from previously cloned AFP promoter constructs using specific primers with 5′EcoRI and 3′BamHI restriction sites respectively [3].
3. Carry out amplification in a 25 μl reaction volume with approximately 100 ng of plasmid DNA as template, 10 mM of each dATP, dCTP, dTTP and dGTP, 20 pmole of each of the primers, 1× Taq polymerase buffer, 1.5 mM $MgCl_2$, and 2.5 units of Taq DNA polymerase. The PCR conditions used were 94 °C for 1 min; 60 °C for 1 min and 72 °C for 1 min for 30 cycles followed by a final extension at 72 °C for 10 min.
4. Cut the Agarose gel and elute AFP promoter/enhancer from 1 % agarose gel by gel extraction kit following manufacturer's protocol.
5. Mix 100 pmoles of forward and reverse *c-Myc* shRNA oligos (Table 1) in microfuge tube containing 100 μl of 1× Annealing Buffer.

6. Boil Water in a beaker and half dip the microfuge tubes with oligos.
7. Maintain temperature at 95–100°C for 5 min and allow the beaker to cool overnight to room temperature (RT).
8. Perform agarose Gel (2.5 %) electrophoresis to analyze and excise annealed shRNA.
9. Ligate NF-κB enhancer (Table 1) upstream to AFP Promoter (−230 to +2 bp) by following standard cloning protocol or as described previously [3].

3.2 Sendai Virus Culture

1. Propagate Sendai virus (Z-strain) within the allantoic sac of 10–11 days old embryonated chicken eggs.
2. Infect the eggs with viral stock diluted in 10 mM PBS containing 2 to 4 Hemagglutinating units (HAU). Postinfection, incubate eggs at 37 °C for 48 h and then at 4 °C for 24 h.
3. Harvest and purify the virus as per the standard methods described by Peretz *et al* [22] and has also been described previously by one of us [23]. Collect the allantoic fluid and centrifuge at 1000 × *g* for 10 min and collect the clear supernatant. Pellet the virus by centrifuging the supernatant for 60 min at 100,000 × *g*, 4 °C.
4. Suspend the virus in 10 mM PBS and homogenize. Estimate the virus yield in terms of protein by Bradford's method. Also, check the activity by hemagglutination and hemolysis assays and the purity by running 10 % SDS-PAGE. These processes are also described earlier [9]. Store the virus at −70 °C till further use.

3.3 Generation of Sendai Fusion (F) Virosomes and Labeling with Octadecyl Rhodamine B Chloride (R18) to Check Fusion Efficiency with Liver Cell Lines

Before studying the effects of TGS, post shRNA delivery, label the Sendai fusion (F) virosomes with R18 dye and check their fusion efficacy with the liver cell lines.

1. The Sendai viral envelopes, containing the F-protein, could be reconstituted and prepared by following the methods described earlier [9]. Briefly, carry out reduction of 50 mg of Sendai virus with 3 mM DTT at 37 °C. Dialyze to remove unreacted free DTT.
2. Resuspend this suspension in nonionic detergent Triton X-100 for 1 h; this would result in the removal of viral genetic material containing the RNA genome and other proteins including HN (*see* **Note 5**).
3. Label the F-virosomes (1 mg/ml) suspension with R-18 by injecting 10 μl (1 mg/ml) ethanolic solution (of R-18) with vortex mixing. Keep this mixture in the dark at RT for 30 min.
4. Remove the unwanted unbound R18 by ultracentrifugation at 100,000*g* for 1 h at 4 °C followed by resuspending the pellet in 10 mM PBS.

3.4 Studying Fusion Kinetics of R18 Labeled Sendai F-virosomes

1. Incubate HepG2, Huh7, and Chang Liver cells (10^6 cells) with 20–50 μg of R-18 labeled F-virosomes for 1 h at 4 °C.
2. After incubation, centrifuge the cells at 500 × g for 5 min to remove unbound virosomes.
3. Resuspend the pellet in 100 μl of cold 10 mM PBS. For measuring fusion kinetics, in a cuvette, take 50 μl of the labeled F-virosome-cell complex suspension having PBS (3 ml) and 1.5 mM Ca^{2+} which was pre-warmed to 37 °C.
4. Record the kinetics of fusion online by a spectrofluorimeter. To normalize the data, the percent fluorescence dequenched (% FDQ) at a time point was estimated as per the equation: % FDQ = $[(F - F_0)/F_t - F_0)] \times 100$, where F_0 indicates fluorescence intensity at time zero and F at a given time point. F_t is the fluorescence intensity recorded in 0.1 % Triton X-100 treated sample that results in 100 % dilution of the probe (Fig. 2).

3.5 Entrapment of AFP Enhancer/Promoter Driven c-Myc shRNA Constructs in Sendai F-virosomes

The general entrapment protocol has also been described previously [9, 23].

1. Take around 100–150 mg of Sendai virus and centrifuge it at 100,000 × *g* for 1 h at 4 °C.
2. Suspend the pellet in 0.02 M Tris-buffered saline (pH 8.3) and DTT solution (keep the final concentration of DTT around 3 mM).
3. Incubate the above mix at 37 °C for 4 h (mild shaking). Boil dialysis bag (range 12,000–14,000) in deionized autoclaved water for 10 min and then dip it in cold TBS. Add viral suspension within this bag and dialyze overnight with 4–6 changes at 2 h.
4. Centrifuge at 100,000 × *g* for 1 h at 4 °C. Use 10 mM Tris-buffered saline pH 7.4 to homogenize the pellet and then add 10 % Triton X-100, keeping the final percentage of triton to 2–5 %.
5. Spin the sample for around 1 h at RT followed by centrifugation at 100,000 × *g* for 1 h at 4 °C.
6. From detergent extract, take the supernatant having only the viral F proteins and lipids.
7. Add 15–20 μg of *c-Myc* shRNA plasmid containing 2 mM EDTA and mix to the supernatant.
8. Remove the detergent from the final solution by using SM-2 Biobeads (usually eight times the amount of detergent) followed by rotating the sample for 2 h at 4 °C. Repeat the above step at RT.
9. Use 26-G needle to collect viral suspension and centrifuge at 100,000*g* for 1 h at 4 °C. Wash the pellet in TBS (pH 7.4) at 100,000 × *g* for 1 h at 4 °C. Take care during this step (*see* **Note 6**).

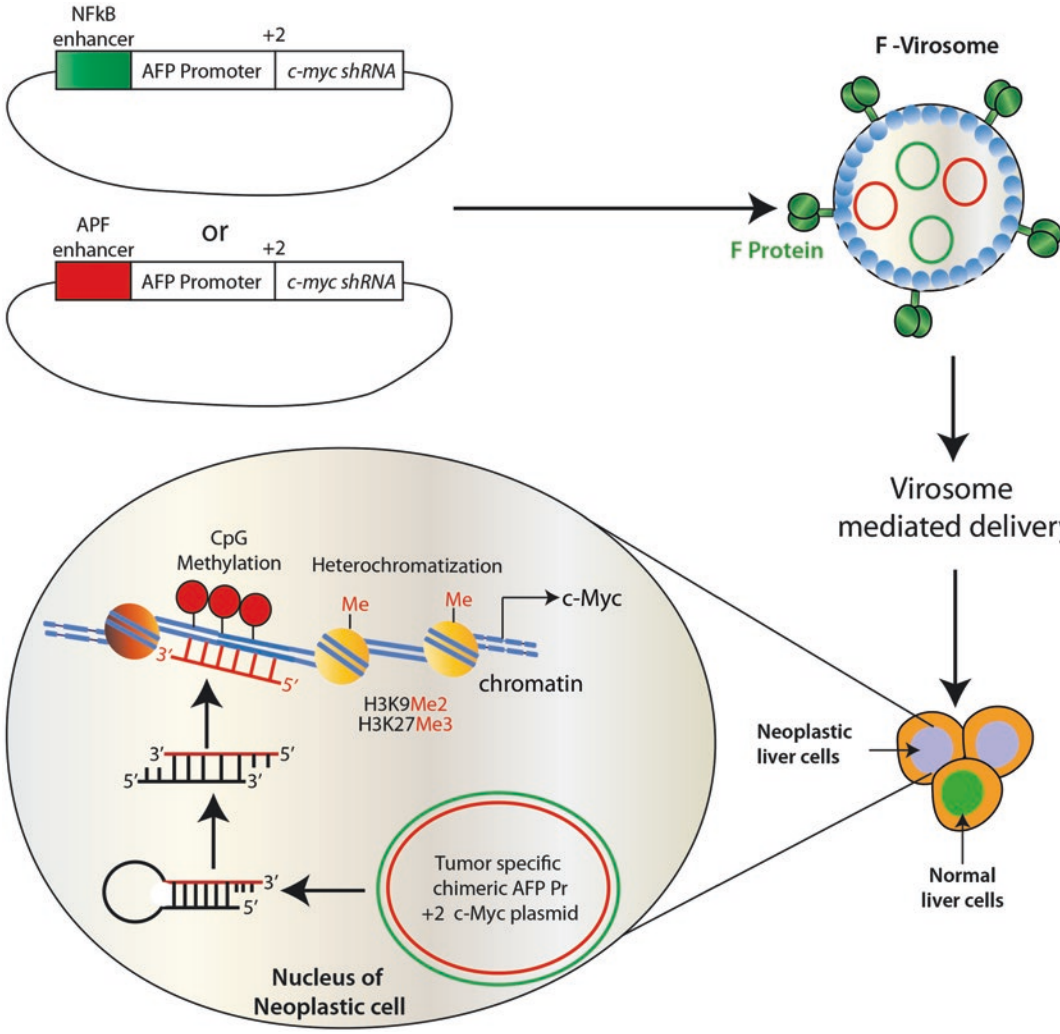


Fig. 2 Fusion Kinetics of R-18 labeled Sendai F virosomes with liver cells (*Source*: Zakaria et al. BMC Cancer. DOI: 10.1186/1471-2407-14-582). The fusion efficacy of Sendai F-virosomes is shown with Huh7, HepG2, and Chang Liver cells. Huh7 and HepG2 show comparable fusion with the virosomes, whereas it was slightly lesser with Chang Liver. The fusion is based on the number of ASGPRs being expressed by individual cell type. Heat-treated Sendai F-virosomes (HC) demonstrate no fusion with the HepG2 cells and is used as a control. Also, HeLa cells, as a nonliver cell line, could be used as a control since it lacks ASGPRs

10. Finally, suspend the pellet in 10 mM phosphate buffered saline pH 7.4. Virosome samples can be stored at 4 °C.
11. Finally, evaluate the structure and function of shRNA loaded virosomes by following published procedure [9].

3.6 Fusion Mediated Delivery of AFP Promoter/Enhancer Driven c-Myc shRNA to Liver Cells In Vitro

1. Prepare monolayer of HepG2, Huh7, or Chang Liver cells (3 x10^5 in 6-well culture plates or 10^6 in T175 culture flasks). Next day, wash twice with 1 ml of OPTI MEM medium to remove DMEM containing serum.

2. Incubate cells with Sendai F-virosomes containing NF-κB–AFP promoter and AFP Enhancer-AFP promoter *c-Myc* shRNA (2–4 mg per 3 × 10^5 cells) at 37 °C in 1 ml of serum-free medium for 2 h at 37 °C in an incubator. Use heat-treated F-virosomes (incubated at 56 °C for 30 min) or cells alone as controls (*see* **Note 7**).
3. After 2 h, remove the medium and supplement with DMEM containing 5–10 % FBS followed by incubation for 5 days (We evaluated the effects after 5 days of treatment).
4. After 5 days, wash cells with PBS and proceed for different assays as required. The cells are now ready for analysis of transgene expression, knockdown studies, proliferation assays, etc. (*see* **Note 8**). We studied these parameters before doing the epigenetic analysis of the *c-Myc* promoter [3].

3.7 Epigenetic Studies Post Sendai F-virosomal Delivery of TGS Inducing Constructs

3.7.1 Heterochromatization Study of the c-Myc Target Locus

We utilized the commercially available Millipore kit for the ChIP Assay to assess the status of the target *c-Myc* P2 promoter post Sendai F-virosomal delivery.

1. ChIP assays were performed using EZ-ChIP™ Chromatin Immunoprecipitation Kit according to the manufacturer's protocol. Sonicate the samples to approximately 200–1000 bp. Adjust the power of your sonicator or vary the pulses during sonication of the samples. Always perform sonication on ice (*see* **Note 9**).
2. Briefly, incubate 2.0 μg of antibodies—Histone three lysine nine dimethylated (H3K9Me2), histone three lysine twenty seven trimethylated (H3K27Me3), histone three acetylated (H3Ac) and control mouse IgG antibody diluted in the antibody buffer (100 μl) for around 90–100 min (*see* **Note 10**). Use the strip wells provided in the kit for incubation.
3. At the same time, use formaldehyde (1 %) to treat 10^6 HepG2 cells for approximately 10 min at 25 °C. This would crosslink the DNA-protein complexes.
4. Use 125 mM glycine to stop crosslinking and isolate nuclei from cells. Utilize the shearing buffer, provided in the kit, to suspend the nuclear pellet and sonicate. We sonicated with three 10 s pulses, with two minutes on ice in between pulses (*see* **Note 11**).
5. Centrifuge at 20,000 × *g* for 10 min at 4°C to separate sheared chromatin from debris. The supernatant could be used now to incubate with antibodies prepared in the strip wells.
6. For reverse crosslinking of histone-DNA, add 5 M NaCl eluent (500 μL) followed by heating at 65 °C for 3–4 h. You can store the samples at −20 °C at this stage.
7. Add 0.5 M EDTA, 1 M Tris-HCl, pH 6.5, and 10 mg/mL Proteinase K to the eluent and incubate at 45 °C for approximately one hour for degrading proteins.

8. Recover the DNA by conventional phenol/chloroform extraction followed by ethanol precipitation (*see* **Note 12**). Wash pellets with 70 % ethanol and air dry.
9. Resuspend pellet in an appropriate buffer for PCR.
10. Use immunoprecipitated DNA as template for Real Time PCR (Primers Table 1). As an input control, use sheared DNA without immunoprecipitation for PCR.
11. Primers against chromosome 16 centrosome (Table 1) could be used as positive control (Table 1; *see* **Note 13**).

3.7.2 CpG Methylation Study of the c-Myc Promoter

To evaluate shRNA mediated methylation of cyosine residues, we used EpiTect® Bisulfite kit (Qiagen).

1. Treat 500 ng of genomic DNA (from the cells treated with shRNA loaded F-virosomes) with sodium bisulfite according to the manufacturer's instructions.
2. Amplify with separate pairs of primers that are specific for either the top or bottom DNA strands.
3. Following PCR amplification, the uracils are amplified as thymines, whereas 5-methyl cytosine (MeC) residues are amplified as cytosines. We use outer and inner primers to amplify, where the inner primers were tagged for sequencing of the PCR products (Table 1; *see* **Note 14**).

4 Notes

1. The aim is to use the minimal promoter region that would provide tissue specificity and minimize nonspecific expression. We used AFP promoter to target HCC; hence, the minimal region of around 230 bp was sufficient to drive transgene expression specifically in HCC cell lines.
2. We used AFP promoter upto +2 bp only. This was aimed at minimizing the shRNA sense strand that could compromise its efficient processing into mature siRNA.
3. We utilized NF-κB as well as AFP enhancer region to augment *c-Myc* shRNA expression specifically in HCC cells. NF-κB is one of the pathways activated in HCC; hence, we utilized the NF-κB enhancer in tandem with AFP promoter for stronger expression of shRNA without losing HCC specificity. Various enhancers could be tried and checked in combination with promoters depending upon the type of study and tissue specificity.
4. Various tumor-specific promoters could be used to target specific cancer cells, like prostate-specific antigen (PSA) promoter to target prostate cancer or carcinoembryonic antigen promoter to target colorectal carcinoma [24–26].

5. It is very important to get rid of the HN proteins during reconstitution process. HN interacts with terminal sialic acid moieties of various cells (like RBCs) and could lead to nonspecific targeting or agglutination.
6. During collection of the viral suspension by 26-G needle, take extreme care not to collect biobeads.
7. Heat treatment inactivates virosome's F protein. Hence, it could be used as a control if the virosome is showing fusion with the target cells. Heat inactivated virosomes should thereby show no fusion with the same cell type. It also proves that the F-proteins are necessary and helps in the fusion process by interacting with the ASGPRs of the target liver cell since the non livers cells (lacking ASGPRs) show no fusion.
8. We also checked the levels of c-Myc post shRNA knockdown before going for the epigenetic analysis of the promoter. The decrease in the c-Myc levels was established by Real time pcr and western blotting, as described previously [3].
9. Perform sonication on ice. The samples might get denatured due to the generation of extreme heat during the sonication process. Also, perform agarose gel electrophoresis post sonication to verify the proper shearing before proceeding to the next step.
10. H3K9Me2 and H3K27Me3 were used for detecting the heterochromatization of the *c-Myc* locus, whereas H3Ac was used to detect the acetylation status post shRNA delivery.
11. Sonication efficiency depends upon the sample type. So, the method for ideal sonication has to be standardized separately.
12. Addition of yeast tRNA or glycogen helps visualize the DNA pellet.
13. Primers against chromosome 16 centromere were utilized as a positive control for heterochromatin marker as histones here are 100 % methylated [26].
14. Bisulfite primers were designed from http://bisearch.enzim.hu/. Nested PCR was done with *c-Myc* outer and inner primers (Table 1) encompassing the shRNA target region. The inner primers were tagged with M13 complementary sequences for sequencing purposes.
15. Instead of a single treatment with shRNA loaded virosomes, two treatments could also be tried (with half of the virosome concentration) in a gap of 2–3 days.
16. Chang Liver cell line was first checked for liver-specific markers and then used in the study. The expression levels of various liver-specific markers were evaluated by real-time pcr and have been shown by us previously [3].

17. Huh7 cells have lower basal level of c-Myc as compared to HepG2 cells. Hence, the TGS effects were more profound in HepG2 cells [3]. Chang Liver cells served as untransformed control cells that showed no induction of TGS since the AFP promoter/enhancer system was inactive in it.

Acknowledgment

This work was supported by a grant from Department of Biotechnology, Government of India (Grant No BT/PR13733/AGR/36/667/2010) and BMC Cancer (Biomedcentral) for permitting the reuse of already published data by MKZ **(DOI:**10.1186/1471-2407-14-582). Department of Science and Technology's J.C. Bose fellowship and R&D grant from Delhi University to Professor Debi P. Sarkar is also acknowledged. Mohammad Khalid Zakaria was supported by a research fellowship from Indian Council of Medical Research (I.C.M.R).

References

1. Ryder SD (2003) Guidelines for the diagnosis and treatment of hepatocellular carcinoma (HCC) in adults. Gut 52(suppl 3):iii1–iii8
2. Khan I, Zakaria MK, Kumar M, Mani P, Chattopadhyay P, Sarkar DP et al (2015) A novel placental like alkaline phosphate promoter driven transcriptional silencing combined with single chain variable fragment antibody based virosomal delivery for neoplastic cell targeting. J Transl Med 13:254
3. Zakaria MK, Khan I, Mani P, Chattopadhyay P, Sarkar DP, Sinha S (2014) Combination of hepatocyte specific delivery and transformation dependent expression of shRNA inducing transcriptional gene silencing of c-Myc promoter in hepatocellular carcinoma cells. BMC Cancer 14(1):582
4. Lin C-P, Liu J-D, Chow J-M, Liu C-R, Liu HE (2007) Small-molecule c-Myc inhibitor, 10058-F4, inhibits proliferation, downregulates human telomerase reverse transcriptase and enhances chemosensitivity in human hepatocellular carcinoma cells. Anticancer Drugs 18(2):161–170
5. Buyens K, De Smedt SC, Braeckmans K, Demeester J, Peeters L, van Grunsven LA et al (2012) Liposome based systems for systemic siRNA delivery: stability in blood sets the requirements for optimal carrier design. J Control Release 158(3):362–370
6. Dahlman JE, Barnes C, Khan OF, Thiriot A, Jhunjunwala S, Shaw TE et al (2014) In vivo endothelial siRNA delivery using polymeric nanoparticles with low molecular weight. Nat Nanotechnol 9(8):648–655
7. Vyas SP, Singh A, Sihorkar V (2001) Ligand-receptor-mediated drug delivery: an emerging paradigm in cellular drug targeting. Crit Rev Ther Drug Carrier Syst 18(1):1–76
8. Wang X, DP S, Mani P, CJ S, Chen Y, Guha C et al (2009) Long-term reduction of jaundice in Gunn rats by nonviral liver-targeted delivery of Sleeping Beauty transposon. Hepatol Baltim Md 50(3):815–824
9. Bagai S, Puri A, Blumenthal R, Sarkar DP (1993) Hemagglutinin-neuraminidase enhances F protein-mediated membrane fusion of reconstituted Sendai virus envelopes with cells. J Virol 67(6):3312–3318
10. Ray U, Roy CL, Kumar A, Mani P, Joseph AP, Sudha G et al (2013) Inhibition of the interaction between NS3 protease and HCV IRES with a small peptide: a novel therapeutic strategy. Mol Ther 21(1):57–67
11. Kumar M, Mani P, Pratheesh P, Chandra S, Jeyakkodi M, Chattopadhyay P et al (2015) Membrane fusion mediated targeted cytosolic drug delivery through scFv engineered sendai viral envelopes. Curr Mol Med 15(4):386–400
12. MS W, LM V, Ehsani A, Amarzguioui M, Aagaard L, Z-X C et al (2006) The antisense strand of small interfering RNAs directs histone methylation and transcriptional gene silencing in human cells. RNA NY 12(2):256–262

13. Napoli S, Pastori C, Magistri M, Carbone GM, Catapano CV (2009) Promoter-specific transcriptional interference and c-myc gene silencing by siRNAs in human cells. EMBO J 28(12):1708–1719
14. Palanichamy JK, Mehndiratta M, Bhagat M, Ramalingam P, Das B, Das P et al (2010) Silencing of integrated human papillomavirus-16 oncogenes by small interfering RNA-mediated heterochromatization. Mol Cancer Ther 9(7):2114–2122
15. Morris KV (2008) RNA-mediated transcriptional gene silencing in human cells. Curr Top Microbiol Immunol 320:211–224
16. Civenni G, Malek A, Albino D, Garcia-Escudero R, Napoli S, Di Marco S et al (2013) RNAi-mediated silencing of Myc transcription inhibits stem-like cell maintenance and tumorigenicity in prostate cancer. Cancer Res 73(22): 6816–6827
17. Shachaf CM, Kopelman AM, Arvanitis C, Karlsson A, Beer S, Mandl S et al (2004) MYC inactivation uncovers pluripotent differentiation and tumour dormancy in hepatocellular cancer. Nature 431(7012):1112–1117
18. Albert T, Wells J, Funk JO, Pullner A, Raschke EE, Stelzer G et al (2001) The chromatin structure of the dual c-myc promoter P1/P2 is regulated by separate elements. J Biol Chem 276(23):20482–20490
19. Wierstra I, Alves J (2008) The c-myc promoter: still MysterY and challenge. Adv Cancer Res 99:113–333
20. Lin C-P, Liu C-R, Lee C-N, Chan T-S, Liu HE (2010) Targeting c-Myc as a novel approach for hepatocellular carcinoma. World J Hepatol 2(1):16–20
21. Luedde T, Schwabe RF (2011) NF-κB in the liver—linking injury, fibrosis and hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 8(2):108–118
22. Peretz H (1974) Fusion of intact human erythrocytes and erythrocyte ghosts. J Cell Biol 63(1):1–11
23. Wang X, Mani P, Sarkar DP, Roy-Chowdhury N, Roy-Chowdhury J (2009) Ex vivo gene transfer into hepatocytes. In: Dhawan A, Hughes RD (eds)*Hepatocyte transplantation* [Internet]. Totowa, NJ: Humana Press. pp 117–139[cited 2016 Apr 2]. Available from http://link.springer.com/10.1007/978-1-59745-201-4_11
24. Latham JP, Searle PF, Mautner V, James ND (2000) Prostate-specific antigen promoter/enhancer driven gene therapy for prostate cancer: construction and testing of a tissue-specific adenovirus vector. Cancer Res 60(2):334–341
25. Li Y, Chen Y, Dilley J, Arroyo T, Ko D, Working P et al (2003) Carcinoembryonic antigen-producing cell-specific oncolytic adenovirus, OV798, for colorectal cancer therapy. Mol Cancer Ther. 2(10):1003–1009
26. Haring M, Offermann S, Danker T, Horst I, Peterhansel C, Stam M (2007) Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. Plant Methods 3(1):11

Chapter 15

Targeting Promoter-Associated Noncoding RNA In Vivo

Gianluca Civenni

Abstract

There are many classes of noncoding RNAs (ncRNAs), with wide-ranging functionalities (e.g., RNA editing, mediation of mRNA splicing, ribosomal function). MicroRNAs (miRNAs) and long ncRNAs (lncRNAs) are implicated in a wide variety of cellular processes, including the regulation of gene expression. Incorrect expression or mutation of lncRNAs has been reported to be associated with several disease conditions, such a malignant transformation in humans. Importantly, pivotal players in tumorigenesis and cancer progression, such as c-Myc, may be regulated by lncRNA at promoter level. The function of lncRNA can be reduced with antisense oligonucleotides that sequester or degrade mature lncRNAs. In alternative, lncRNA transcription can be blocked by small interference RNA (RNAi), which had acquired, recently, broad interested in clinical applications. In vivo-jetPEI™ is a linear polyethylenimine mediating nucleic acid (DNA, shRNA, siRNA, oligonucelotides) delivery with high efficiency. Different in vivo delivery routes have been validated: intravenous (IV), intraperitoneal (IP), intratumoral, subcutaneous, topical, and intrathecal. High levels of nucleic acid delivery are achieved into a broad range of tissues, such as lung, salivary glands, heart, spleen, liver, and prostate upon systemic administration. In addition, in vivo-jetPEI™ is also an efficient carrier for local gene and siRNA delivery such as intratumoral or topical application on the skin. After systemic injection, siRNA can be detected and the levels can be validated in target tissues by qRT-PCR. Targeting promoter-associated lncRNAs with siRNAs (small interfering RNAs) in vivo is becoming an exciting breakthrough for the treatment of human disease.

Key words Noncoding RNAs, Transcriptional gene silencing, RNA interference, Small interfering RNA,, Polyethylenimine

1 Introduction

Exclusively a minor fraction (ca. 1 %) of the whole genome codes for protein and therefore the majority of cellular transcriptional RNA output is noncoding. There are many classes of ncRNAs with a broad range of functions (e.g., RNA editing, mediation of mRNA splicing, ribosomal function). Because ncRNA may modulate gene expression and their miss-expression and mutation are associated with cancer in humans, they are therapeutic targets with a strong potential [1].

Since the discovery of RNA interference (RNAi), there has been an increasing interest in using this technology for clinical

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_15,

applications [2]. Cancer therapies include many small molecule inhibitors and monoclonal antibodies. Unfortunately, many important targets for cancer therapies are hard to be inhibited using these strategies. Most small-molecule inhibitors are not specific leading to undesirable toxicity and in the case of monoclonal antibodies, the target protein may not be accessible. In addition, in case of lncRNAs the absence of a translated protein excludes the possibility of applying antibodies technology.

The use of RNAi in the clinic is attractive as it can circumvent many of these problems and its potential for use as a therapeutic has been supported by a report of systemic small interfering RNA(siRNA) delivery into human tumors [3].

There are several challenges that currently limit the use of siRNA in the clinic. In the bloodstream, naked siRNAs have a short half-life because their rapid degradation by serum RNase A-type nucleases and fast renal clearance [4]. Chemical modification of the sugars backbone or the bases of oligoribonucleotides have been introduced for stabilization [5, 6]. In addition, the hydrophobic cell membranes are a challenge for the intracellular delivery of negatively charged polymers. Moreover, once siRNAs are intracellular, they only transiently cause gene silencing, as the concentration of the siRNAs decreases with each cell division [7].

Regardless of the route of entry, either via receptor mediated endocytosis or by pinocytosis, siRNA will be moved into the early endosomes, which commonly fuse with other endocytic vesicles, late endosomes. Late endosomes finally fuse with the lysosomes where degradation takes place due to acidic environment and degrading enzymes. Thus for a successful gene transfer to take place siRNAs have to escape from the endosomes before they fuse with lysosomes.

One of the most successful approaches to avoid enzymatic degradation is to condense the oligonucelotides into a compact form, so the sites susceptible to cleavages are protected. This condensation is based on the electrostatic interactions between the anionic nucleic acid and the positive charges of the synthetic carrier, which complex and condense the oligonucelotides into nanoparticles. Several polycations such as Polyethylenimine (PEI), Poly-L-Lysine (PLL), cationic lipids, and dendrimers have been used to overcome degradative enzymes.

Polyethylenimine forms stable complexes with oligonucleotides [8, 9]. The resulting positively charged particles are able to interact with anionic proteoglycans at the cell surface and to enter cells by endocytosis [10]. Polyethylenimine possesses the unique property of acting as a “proton sponge” that buffers the endosomal acidity (pH 2.5) and protects oligonucleotides from degradation [9]. The continuous proton influx also induces endosome osmotic swelling and break, which allow oligonucleotides to escape into the cytoplasm.

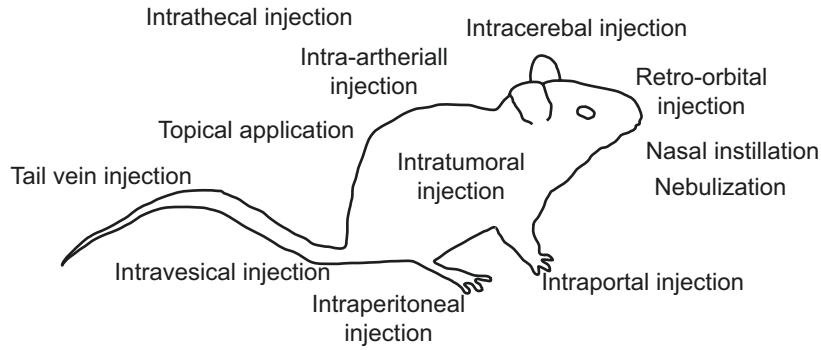


Fig. 1 Published delivery routes in mouse models

Finally, in order that a delivered siRNA to be functional it has to be transported into the nucleus. This entry is governed by the nuclear membrane where the nuclear pore complex mediates the transport of molecules. This complex allows passive passage of small molecules but severely limits passage of larger molecules of more than 50 kD across membrane [11, 12].

In vivo-jetPEI™ is a linear polyethylenimine for effective and reproducible in vivo gene and oligonucleotide delivery with low toxicity [13]. It is more efficient than other cationic lipids and polymers [14], including the branched PEIs [15–17]. In vivo-jetPEI™ mediates efficient nucleic acid delivery to a wide range of tissues using various delivery routes (Fig. 1). For example, it has been used successfully for local [18, 19], intravenous [11, 12, 16, 17, 20–22], intracerebral [23–25], intraperitoneal injections [26–28], and intratracheal instillation [15, 22] (Fig. 1). One of the most efficient administration routes for high levels of gene delivery in the lung consists in systemic administration [11, 12, 22].

Other organs such as the heart, spleen, pancreas, and liver are also transfected following IV injection [11, 12]. Interestingly, in vivo-jetPEI™ is also an effective carrier for local gene delivery and in particular in the brain [29]. For example, in vivo-jetPEI™ efficiently delivers DNA following intraventricular or intrathecal injections [24, 30]. Furthermore, the expression of functional proteins following DNA delivery with in vivo-jetPEI™ in neurons has been observed in genomic studies and therapeutic approaches [30–32]. Other local delivery routes include topical application on the skin [33, 34] and intratumoral injections [35]. More recently, in vivo-jetPEI™ has become a powerful tool for in vivo RNA interference experiments [28, 36–38].

Regarding the transfection mechanism, in vivo-jetPEI™ condenses oligonucelotides into positively charged particles or complexes capable of interacting with anionic proteoglycans at the cell surface and subsequently entering cells by endocytosis [10]. It possesses the unique property of acting as a “proton sponge” within the

endosome and buffers the endosomal pH protecting the oligonucelotides from degradation. Continuous proton influx progressively induces endosome osmotic swelling and rupture thereby providing an escape mechanism for oligonucelotides complexes, which are released in the cytoplasm [7, 38]. Moreover, linear PEI was shown to enhance entry to the nucleus when compared with branched PEI and cationic lipids [14].

In vivo-jetPEI™ had been successfully used to carry gene silencing with siRNA by intraperitoneal injections [28, 39], targeting, for example, the oncogene c-Myc [28], one of the most important players in human neoplasia [40]. In 2009, we showed that transcription of c-Myc is silenced in vitro with high efficiency and selectivity by targeting its promoter with a specific siRNA (siMyc13) [41]. Our approach was based on the presence of a *cis*-acting noncoding promoter-associated RNA (paRNA) overlapping the Myc transcription start site (TSS) and positively regulating transcription initiation [41]. Interfering with this paRNA using a promoter-targeting siRNA inhibited transcription of Myc in a sequence- and gene-specific manner in vitro [41]. Transcriptional silencing by promoter-targeting siRNAs indeed could provide effective means to modulate gene expression in addition to canonical posttranscriptional RNA interference (RNAi) that target mRNAs [42]. As we showed more recently, a similar approach can be used in vivo, delivering siRNA-Myc13 by intraperitoneal injection in mouse carrying human prostate cancer xenografts [28]. In this study, we showed that Myc silencing by promoter-targeting siRNA impairs prostate cancer stem cells maintenance and tumorigenicity and induces senescence in the prostate cancer stem cell subpopulation in vivo [28].

These findings demonstrate also that RNAi-based targeting of regulatory noncoding paRNAs is an effective strategy to modulate transcription of genes involved in critical oncogenic pathways for therapeutic applications.

2 Materials

1. 10 % sterile isotonic glucose solution (w/v).
2. In vivo-jetPEI™ (Polyplus Transfection).
3. Phosphate buffered saline: DULBECCO'S D-PBS without Ca & Mg, 1X Liquid CE.
4. Cellular culture media: RPMI-1640.
5. Penicillin–Streptomycin.
6. Bovine serum.
7. Matrigel: Matrigel® basement membrane matrix (Corning).
8. Ethylenediaminetetraacetic acid disodiumsaltdihydrate (EDTA), pH 8.0.

9. siRNA.
10. Negative control.
11. Counting chamber or an automatic cell counter.
12. Trizol.
13. Sonicator.
14. Micro RNA retrotanscription kit (Applied Biosystem).
15. Custom Small RNA Taqman assay (Applied Biosystem).
16. Taqman universal Master mix (Applied Biosystem).
17. One-step qRT PCR SYBR green kit.
18. Primers for mRNA detection.

3 Methods

3.1 Generating Subcutaneous Xenografts Using Human Prostate Cancer Cell Lines

1. To generate subcutaneous tumor xenografts, human prostate cancer cell lines PC3 (CRL-1435) were purchased from ATCC.
2. Cells have to be cultured and expanded in vitro in adherence conditions using RPMI-1640 supplemented with FCS and penicillin-streptomycin.
3. Detach the cells with PBS supplemented with 2 mM EDTA.
4. Cells in suspension are transferred in 15-ml Falcon tube and centrifuged at 1200 RPM (G force = 153) at 4 °C.
5. Supernatant was decanted and the pellet was resuspended in PBS.
6. Cell numbers can be determined with a counting chamber or an automatic cell counter.
7. Put the required cell amount (ca. 3×10^6) in 100 µl of PBS buffer (*see* **Note 1**).
8. Add equal volume (100 µl) of Matrigel (*see* **Note 2**).
9. Mix gently and keep on ice (*see* **Note 2**).
10. Load the cells mixture using a 1 ml syringe without needle (*see* **Note 3**).
11. Mount the needle (25 gauge) on the syringe and remove bobbles (*see* **Notes 3** and **4**).
12. Injected subcutaneously in athymic nude mice (*see* **Notes 5**–**7**).
13. Wait 2–3 weeks and use themice for siRNA systemic treatment when the tumor xenografts reached the volume of ca. 100 mm^3.

3.2 Preparation of the Complexes with 10 % Glucose Stock Solution

The following protocol is given for intraperitoneal injection using 2 mg/kg of siRNA (siGL3 or siMyc13) and in vivo-jetPEI™ (*see* **Note 8**) at N/P ratio (see **Note 9**) of 5 in a final volume of 600 µl (refer to Table 1 for other DNA amounts and other N/P ratios) (*see* **Notes 8** and **9**).

The preparation of the in vivo-jetPEI™/DNA complexes has to be performed under sterile conditions (Fig. 2):

Table 1
Volumes of in vivo-jetPEI™ according to the amount of DNA and N/P ratios

Amount of oligonucleoties (μg)	Volume (μl) of in vivo-jetPEI™	
	N/P = 5	N/P = 8
15	0.5	0.8
60	2.0	3.2
150	5.0	8.0

1. For one mouse (see **Note 10**) with a weight of 30 g, dilute 60 μg of siRNA (*see* **Note 11**) into 150 μl of 10 % sterile isotonic glucose solution (w/v). Adjust the volume to 300 μl with sterile water (see **Note 12**) to obtain a final concentration of 5 % glucose (*see* **Note 13**). Vortex gently and centrifuge briefly.
2. Dilute 6 μl of in vivo-jetPEI™ in 150 μl of 10 % glucose (*see* **Note 14**). Adjust the volume to 300 μl with sterile water to obtain a final concentration of 5 % glucose (*see* **Note 13**). Vortex gently and spin down briefly.
3. Add 300 μl in vivo-jetPEI™ to 300 μl DNA at once (important: do not mix in reverse order).
4. Briefly vortex the solution immediately and spin down.
5. Incubate for 15 min at room temperature (complex is stable 24 h if stored at 4 °C).
6. Perform injections into animal (*see* **Notes 15** and **16**).

3.3 Treatment with a Promoter-Targeting siRNA (Myc13) and Monitoring In Vivo Delivery, Gene Target, and Tumor Growth at the Appropriate Time Points

To assess the efficiency of the systemic delivery of the promoter-targeting siRNA, Myc13 siRNA, c-Myc mRNA expression levels, and tumor growth were evaluated

1. The delivery of Myc13 siRNA in tumor tissues following intraperitoneal injections can be determined with different doses and time points.
2. Mice can be injected intraperitoneal once with the three different doses (e.g., 0.5, 2, and 5 mg/kg) and sacrificed after 2 days. In a second set, mice can receive a single injection of a dose of 2 mg/kg and sacrificed at different time points (e.g., 1, 3, or 7 days).
3. Prostate xenograft tissues are explanted and transferred in a 3.5 cm^2 dish containing 2 ml of PBS.
4. Using a sterile scalpel, tissue is cut in small pieces (ca. 30 mm^3).
5. Each tissue piece is transferred in a cryotube and snap-frozen.
6. At this point, tissues can be stored at −80 °C.

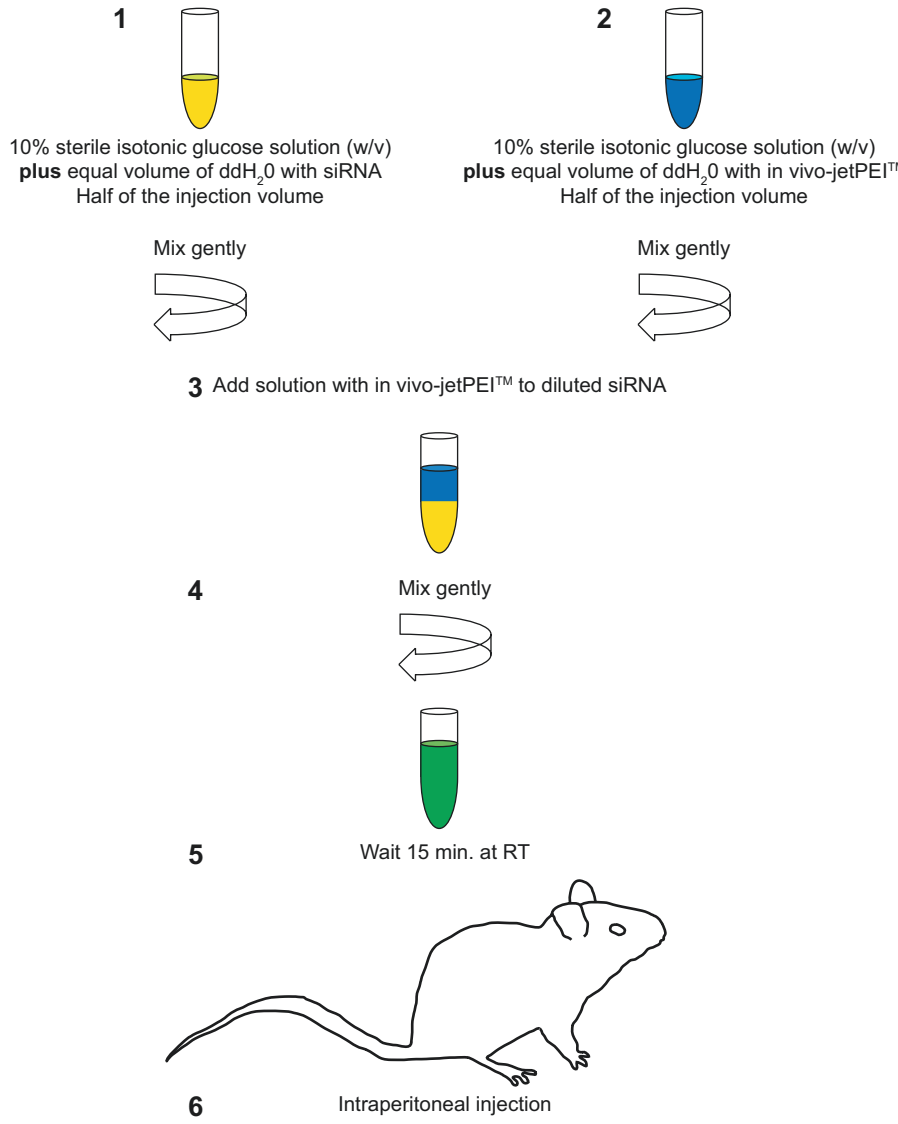


Fig. 2 Steps of the preparation of the siRNA/jetPEI™ complexes with 10 % glucose stock solution

7. For further analysis, frozen tissues are dissolved using TRizol (Invitrogen).
8. To reach complete dissociation, tissues are homogenized using a Bandelin sonicator (30 % power, 9 cycles, 30 s).
9. RNA can be isolated according to TRizol manufacturer's instructions (Invitrogen).
10. To accurately measure the Myc13 siRNA in tissue samples, we use a qRT-PCR method with Myc13-specific primers and U6 small RNA as control [43] (*see* **Note 17**).

Fig. 3 Athymic nude mice carrying a subcutaneous xenograft (*dotted red line*). The dimensions *x* and *y* to calculate the tumor volume are indicated (*solid black lines*)

11. In parallel, the levels of Myc mRNA can be measured in the same samples.
12. Based on the pharmacodynamics and pharmacokinetics results [28], mice carrying human prostate xenografts tumor with a volume ca. 100 cm^3 can be treated twice a week with intraperitoneal injections of control (GL3) and Myc13 siRNA (with freshly prepared complexes each time).
13. Tumors were measured with a calliper (Fig. 3). Two dimensions were measured, x and y, and volume was calculated with the following formula:

$$Tumor\ volume\ (mm3) = \frac{x^2 \times y}{2}$$

where x is bigger than *y* (Fig. 3).

14. For accurate measurement of tumor size, when groups of control and Myc13 siRNA-treated mice were sacrificed the tumor weight was determined using an analytical balance.

4 Notes

1. The solution with cancer cells, PBS, and Matrigel has to be kept in a sterile round-bottom disposable tube. This makes easier to load the syringe without needle prior to the injection.
2. Matrigel solution has always been kept on ice. For this reason, fill the syringe just shortly before injection. On the contrary, Matrigel can solidify in the syringe.
3. Do not fill the syringe with needle; this could break the cells reducing the number of injected alive cells.
4. Since cells tend to precipitate and Matrigel is a dense solution, mix the cells, PBS, and Matrigel mixture very well just before filling the syringe.
5. Athymic nude mice (Balb/c nu/nu, 4–6 weeks old) can be purchased from Harlan.

   Animals have to be maintained under sterile and controlled conditions of temperature (22–24°C), light (12 h light/12 h dark), and humidity (45–65 %), with food and water ad libitum.

6. Mice have to be housed and sacrificed according to the local Law.
7. When more than one mouse is injected subcutaneously with cancer cells, prepare a unique cell, PBS, and Matrigel mixture and keep on ice.
8. JetPEI™ concentration is expressed in nitrogen residues molarity and 1 μg of oligonucleotide contains 3 n moles of anionic phosphate.
9. Efficient cell entry necessitates cationic particles. For this reason, the ionic balance of jetPEI™ cations and siRNA anions has to be in favor of the cations [44]. The N/P ratio is defined as measure of the ionic balance of the complexes. The N/P ratio is the number of nitrogen residues of jetPEI™ per oligonucleotide phosphate. Since nearly one of every three nitrogen atoms of PEI is a cation, electroneutrality of the complexes jetPEI™/siRNA is reached with a N/P ratio equal to 2–3. The optimal transfection conditions for in vivo delivery required a N/P ratio between 5 and 8.

The volume of jetPEI™ solution to be mixed with oligonucleotides to obtain a desired N/P ratio is given in calculated using Formula 1:

Formula 1:

$$\mu l \text{ of jetPEI™ } to\ be \text{ used} = \frac{\left(\mu g \text{ of oligonucleotide} \times 3\right) \times \text{N / P ratio}}{\text{jetPEI™concentration in nitrogen residues}\left(mM\right)}$$

To calculate the volume of jetPEI™ solution to be mixed with oligonucleotides (ON) according to the ON size, use formula 2:

Formula 2:

$$\mu l \text{ of jetPEI™ } to\ be \text{ used} = \frac{\left(ON \text{ base number}\right) \times \left(\text{pmoles of } ON\right) \times \text{N / P ratio} \times 10^{-3}}{\text{jetPEI™concentration in nitrogen residues}\left(mM\right)}$$

10. Experimental groups consisted of at least five animals with equal age (8–10 weeks) and body weight (ca. 30 g).
11. The amount of siRNA delivered is established in part by the animal model, target, organ, and route of injection. As a general principle, it is recommended to use in vivo-jetPEI™ at a N/P ratio between 5 and 8. The amount of siRNA and injection volume should be adjusted to the size of the animal and the route of administration. In order to prevent precipitation, the final concentration of DNA in the final volume should not exceed 0.5 μg/μl.
12. Use siRNA preparation that does not contain salt. It is best to use oligonucleotides prepared in water. For siRNA, the best

results are achieved with high quality endotoxin-free DNA resuspended in ddH_2O. For siRNA, order high quality grade siRNA (PAGE or HPLC purification).

13. The final concentration of glucose in the injection solution is 5 %.
14. It is recommended to use 10 % sterile isotonic glucose solution (w/v), which is needed to generate small and stable nucleic acids/in vivo-jetPEI™ complexes. Since high salt concentrations prevent the formation of small and stable in vivo-jetPEI™/oligonucelotides complexes, the use of ionic solutions such as PBS or cell culture media for complex preparation has to be avoided. The nucleic acid has to be resuspended in low salt buffer since high salt content in the nucleic acid preparation leads to precipitation upon complexes formation. The best results are achieved with high quality siRNA that is prepared in ddH_2O.
15. Prior to injections, ensure that in vivo-jetPEI™ and all the solutions are at room temperature.
16. If mortality occurs the amount of siRNA used in keeping the N/P ratio constant should be decreased. In alternative, the N/P ratio can be reduced, keeping the amount of RNA constant. Mortality may also be due to wrong oligonucleotides preparation. The solution must be endotoxin-free.
17. If low siRNA delivery is detected, it is possible to optimize the amount of siRNA used in the injection volume. An option is to increase the N/P ratio but this could lead to toxicity.

References

1. Roberts TC, Wood MJA (2013) Therapeutic targeting of non-coding RNAs. Biochemical Society Essays Biochem 54:127–145
2. Fire A, Xu S, Montgomery MK, Kostas SA et al (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391(6669):806–811
3. Davis ME, Zuckerman JE, Choi CH et al (2010) Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. Nature 464(7291):1067–1070
4. Bumcrot D, Manoharan M, Koteliansky V, Sah DW (2006) RNAi therapeutics: a potential new class of pharmaceutical drugs. Nat Chem Biol 2(12):711–719
5. Soutschek J, Akinc A, Bramlage B et al (2004) Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. Nature 432(7014):173–178
6. Czauderna F, Fechtner M, Dames S et al (2003) Structural variations and stabilising modifications of synthetic siRNAs in mammalian cells. Nucleic Acids Res 31(11):2705–2716
7. Kim DH, Rossi JJ (2007) Strategies for silencing human disease using RNA interference. Nat Rev Genet 8(3):173–184
8. Boussif O, Lezoualc'h F, Zanta MA et al (1995) A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. Proc Natl Acad Sci USA 92(16):7297–7301
9. Dheur S, Dias N, van Aerschot A et al (1999) Polyethylenimine but not cationic lipid improves antisense activity of 3′-capped phosphodiester oligonucleotides. Antisense Nucleic Acid Drug Dev 9(6):515–525
10. Mislick KA, Baldeschwieler JD (1996) Evidence for the role of proteoglycans in cation-mediated gene transfer. Proc Natl Acad Sci USA 93(22):12349–12354
11. Goula D, Benoist C, Mantero S et al (1998) Polyethylenimine-based intravenous delivery

of transgenes to mouse lung. Gene Ther 5(9): 1291–1295

12. Zou SM, Erbacher P, Remy JS, Behr JP (2000) Systemic linear polyethylenimine (L-PEI)-mediated gene delivery in the mouse. J Gene Med 2(2):128–134
13. Demeneix B, Behr J, Boussif O et al (1998) Gene transfer with lipospermines and polyethylenimines. Adv Drug Deliv Rev 30(1–3):85–95
14. Brunner S, Furtbauer E, Sauer T et al (2002) Overcoming the nuclear barrier: cell cycle independent nonviral gene transfer with linear polyethylenimine or electroporation. Mol Ther 5(1):80–86
15. Ferrari S, Pettenazzo A, Garbati N et al (1999) Polyethylenimine shows properties of interest for cystic fibrosis gene therapy. Biochim Biophys Acta 1447(2–3):219–225
16. Chemin I, Moradpour D, Wieland S et al (1998) Liver-directed gene transfer: a linear polyethlenimine derivative mediates highly efficient DNA delivery to primary hepatocytes in vitro and in vivo. J Viral Hepat 5(6):369–375
17. Bragonzi A, Boletta A, Biffi A et al (1999) Comparison between cationic polymers and lipids in mediating systemic gene delivery to the lungs. Gene Ther 6(12):1995–2004
18. Lisziewicz J, Trocio J, Whitman L et al (2005) DermaVir: a novel topical vaccine for HIV/AIDS. J Invest Dermatol 124(1):160–169
19. Ohana P, Schachter P, Ayesh B et al (2005) Regulatory sequences of H19 and IGF2 genes in DNA-based therapy of colorectal rat liver metastases. J Gene Med 7(3):366–374
20. Goula D, Becker N, Lemkine GF et al (2000) Rapid crossing of the pulmonary endothelial barrier by polyethylenimine/DNA complexes. Gene Ther 7(6):499–504
21. Coll JL, Chollet P, Brambilla E et al (1999) In vivo delivery to tumors of DNA complexed with linear polyethylenimine. Hum Gene Ther 10(10):1659–1666
22. Bragonzi A, Dina G, Villa A et al (2000) Biodistribution and transgene expression with nonviral cationic vector/DNA complexes in the lungs. Gene Ther 7(20):1753–1760
23. Ouatas T, Le Mevel S, Demeneix BA, de Luze A (1998) T3-dependent physiological regulation of transcription in the Xenopus tadpole brain studied by polyethylenimine based in vivo gene transfer. Int J Dev Biol 42(8):1159–1164
24. Goula D, Remy JS, Erbacher P et al (1998) Size, diffusibility and transfection performance of linear PEI/DNA complexes in the mouse central nervous system. Gene Ther 5(5): 712–717
25. Demeneix BA, Ghorbel M, Goula D (2000) Optimizing polyethylenimine-based gene transfer into mammalian brain for analysis of promoter regulation and protein function. Methods Mol Biol 133:21–35
26. Louis MH, Dutoit S, Denoux Y et al (2006) Intraperitoneal linear polyethylenimine (L-PEI)-mediated gene delivery to ovarian carcinoma nodes in mice. Cancer Gene Ther 13(4):367–374
27. Aoki K, Furuhata S, Hatanaka K et al (2001) Polyethylenimine-mediated gene transfer into pancreatic tumor dissemination in the murine peritoneal cavity. Gene Ther 8(7):508–514
28. Civenni G, Malek A, Albino D et al (2013) RNAi-mediated silencing of Myc transcription inhibits stem-like cell maintenance and tumorigenicity in prostate cancer. Cancer Res 73(22):6816–6827
29. Abdallah B, Hassan A, Benoist C et al (1996) A powerful nonviral vector for in vivo gene transfer into the adult mammalian brain: polyethylenimine. Hum Gene Ther 7(16):1947–1954
30. Guissouma H, Dupre SM, Becker N et al (2002) Feedback on hypothalamic TRH transcription is dependent on thyroid hormone receptor N terminus. Mol Endocrinol 16(7):1652–1666
31. Wu K, Meyers CA, Bennett JA et al (2004) Polyethylenimine-mediated NGF gene delivery protects transected septal cholinergic neurons. Brain Res 1008(2):284–287
32. Lemkine GF, Mantero S, Migne C et al (2002) Preferential transfection of adult mouse neural stem cells and their immediate progeny in vivo with polyethylenimine. Mol Cell Neurosci 19(2):165–174
33. Lisziewicz J, Trocio J, Xu J et al (2005) Control of viral rebound through therapeutic immunization with DermaVir. AIDS 19(1):35–43
34. Lisziewicz J, Gabrilovich DI, Varga G et al (2001) Induction of potent human immunodeficiency virus type 1-specific T-cell-restricted immunity by genetically modified dendritic cells. J Virol 75(16):7621–7628
35. Ohana P et al (2004) Regulatory sequences of the H19 gene in DNA based therapy of bladder cancer. Gene Ther Mol Biol 8:181–192
36. Paranjpe S, Bowen WC, Bell AW et al (2007) Cell cycle effects resulting from inhibition of hepatocyte growth factor and its receptor c-Met in regenerating rat livers by RNA interference. *Hepatology* 45(6):1471–1477
37. Liao HW, Yau KW (2007) In vivo gene delivery in the retina using polyethylenimine. *Biotechniques* 42(3):285–286, 288

38. George J, Tsutsumi M (2007) siRNA-mediated knockdown of connective tissue growth factor prevents N-nitrosodimethylamine-induced hepatic fibrosis in rats. Gene Ther 14(10): 790–803
39. Campbell M, Hanrahan F, Gobbo OL et al (2012) Targeted suppression of claudin-5 decreases cerebral oedema and improves cognitive outcome following traumatic brain injury. Nat Commun 3:849
40. Dang CV (2012) MYC on the path to cancer. Cell 149(1):22–35
41. Napoli S, Pastori C, Magistri M et al (2009) Promoter-specific transcriptional interference and c-myc gene silencing by siRNAs in human cells. EMBO J 28(12):1708–1719
42. Pastori C, Magistri M, Napoli S et al (2010) Small RNA-directed transcriptional control: new insights into mechanisms and therapeutic applications. Cell Cycle 9(12): 2353–2362
43. Malek A, Catapano CV, Czubayko F, Aigner A (2010) A sensitive polymerase chain reaction-based method for detection and quantification of metastasis in human xenograft mouse models. Clin Exp Metastasis 27(4):261–271
44. Dheur S, Saison-Behmoaras TE (2000) Polyethyleneimine-mediated transfection to improve antisense activity of 3′-capped phosphodiester oligonucleotides. *Methods Enzymol* 313:56–73

Chapter 16

Manipulation of Promoter-Associated Noncoding RNAs in Mouse Early Embryos for Controlling Sequence-Specific Epigenetic Status

Nobuhiko Hamazaki, Kinichi Nakashima, and Takuya Imamura

Abstract

In mammals, transcription in the zygote begins after fertilization. This transcriptional wave is called zygotic gene activation (ZGA). During ZGA, epigenetic modifications, such as DNA methylation and histone modifications, are dynamically and drastically reconstructed in a sequence-specific manner. However, how such orchestrated gene upregulation is regulated remains unknown. Recently, using microinjection techniques, we have revealed that a class of long noncoding RNAs, named promoter-associated noncoding RNAs (pancRNAs), mediates specific gene upregulation through promoter DNA demethylation during ZGA. Here, we describe the experimental methods available to control the expression levels of pancRNAs and to evaluate epigenetic status after pancRNA manipulation.

Key words Microinjection, Long noncoding RNA, Mouse, DNA methylation

1 Introduction

Genome-wide gene activation in the zygote, called zygotic gene activation (ZGA), begins mainly at the 2-cell stage in mouse [1, 2]. ZGA allows embryos to acquire the potency to form all cell types. During ZGA, epigenetic modifications of DNA and of histone proteins are dynamically and drastically reconstructed [3]. DNA demethylation is a well-known epigenetic event involved in the reconstruction of the zygote's chromatin structure during ZGA. It has long been thought that after fertilization, the bulk of the genomic DNA, including the repeated sequences, such as long interspersed nuclear elements called LINEs, become demethylated as a major part of epigenetic reprogramming [4–6]. In actuality, the DNA methylation of promoter regions behaves quite uniquely during the epigenetic reprogramming [7–9]. The DNA methylation pattern of promoter regions seems to be determined in gene-specific manners. Thus, sequence-specific machineries should

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2_16,

regulate the gene activation pattern. One key issue is how such sequence-specific gene activation is achieved in the process of acquisition of pluripotency in early mouse embryos.

Long noncoding RNAs (lncRNAs) constitute one group of factors that play roles in such local epigenetic alterations. A set of lncRNAs transcribed from bidirectional promoters, named promoter-associated noncoding RNAs (pancRNAs), are polyA+ RNAs involved in the sequence-specific upregulation of their oppositely transcribed partner genes via sequence-specific promoter DNA demethylation [10, 11]. Recently, using microinjection techniques, we have shown that pancRNAs also played essential roles in sequence-specific gene activations during ZGA via promoter DNA demethylation (Fig. 1) [12].

Typically, cell heterogeneity in a sample population makes it difficult to dissect possible DNA methylation changes. Since DNA methylation levels are calculated as the frequency of methylation of alleles in the cell population, low manipulation efficiency compromises the evaluation of the functionality of molecules that are introduced into cells. Accordingly, the microinjection technique is useful because it enables one to obtain a set of homogeneous populations

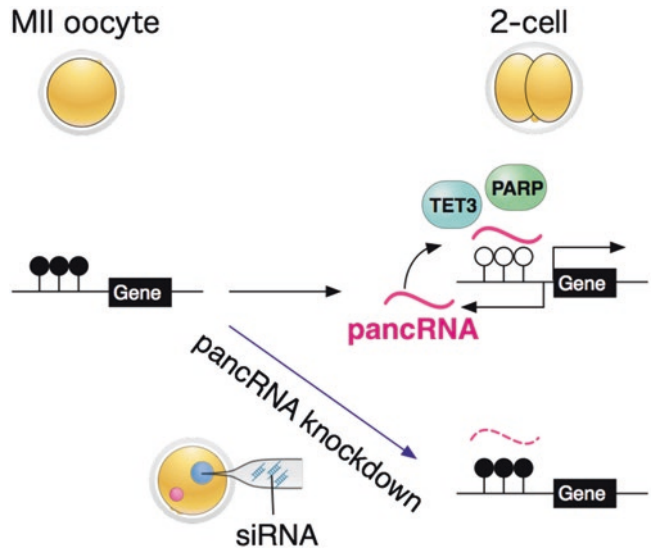


Fig. 1 Schematic model of pancRNA-driven DNA demethylation and gene upregulation during ZGA. Black and white circles indicate methylated and unmethylated cytosines of CpG positions. In this scheme, in unfertilized MII oocytes, DNA methylation levels of the gene promoter region were relatively high. After fertilization, expression of a pancRNA starts at around the 2-cell stage. This pancRNA expression precedes DNA demethylation of the promoter, possibly due to recruitment of DNA demethylation components, such as TET3 and PARP, and a partner-specific gene activation. Knockdown of such pancRNA by siRNA microinjection into the pronucleus caused aberrant high DNA methylation levels and repression of the partner gene

since one can easily select the embryos into which molecules have been properly injected. We have taken advantage of the embryo microinjection system for loss and gain experiments to understand the role of pancRNAs during ZGA [12]. In this system, siRNA and pancRNA overexpression plasmid DNA are injected into the pronuclei for knockdown and overexpression of pancRNA, respectively.

Using this system, we have found that knockdown of pancRNAs inhibits the corresponding DNA demethylation and gene upregulation during ZGA [12]. Conversely, overexpression of pancRNAs can re-induce the DNA demethylation in pancRNA-knocked-down embryos. This clearly indicates that microinjection can be a powerful tool for dissecting thus-far hidden roles of noncoding RNAs in controlling the epigenetic status in preimplantation embryos. Usually, it is difficult to dissect the DNA methylation pattern of microinjected embryos due to the limitation of available samples in a microinjection experiment (only 20–35 embryos per donor female mouse). Standard bisulfite-reaction-based DNA methylation analysis needs more than 10,000 cells on average [13]. To overcome this difficulty, we have optimized the bisulfite conversion protocol for small-scale samples of manipulated embryos. As described below, our method enables us to analyze the DNA methylation profiles from a minimum of 10 embryos.

Here, we describe how to control pancRNAs' expression and to evaluate the epigenetic status in mouse embryos.

2 Materials

2.1 Oocyte or Fertilized Embryo Collection

1. 8-week-old mice of specific strains, such as B6C3F1.
2. Pregnant mare's serum gonadotropin (PMSG).
3. Human chorionic gonadotropin (hCG).
4. M2 medium.
5. M16 medium.
6. CO_2 incubator.
7. 35-mm Petri dishes.
8. Mouth pipet, which is equipped with a mouthpiece and proximal tubing.
9. 1 % hyaluronidase, dissolved in M2 medium.
10. Paraffin oil.
11. Scissors and forceps.
12. 100-mm Petridishes.

2.2 Pipettes for Microinjection

1. Glass capillary (GD-1).
2. Micropipette Puller (Sutter).
3. Micro Forge.
4. Beveller.

2.3 DNA or RNA for Injection

1. siRNAs for pancRNA knockdown (designed according to http://sidirect2.rnai.jp/, for example).
2. Negative Control siRNA.
3. DNA for RNA overexpression.
4. Fluorescent gene expression plasmid (final concentration: less than 10 ng/μl).
5. Microinjector.

2.4 RNA Expression Analysis by RT-qPCR

1. DEPC-treated water.
2. RNA extraction kit.
3. Reverse transcription kit.
4. RT-qPCR primers.
5. Chloroform.
6. Isopropyl alcohol.
7. 75 % ethanol (in DEPC-treated water).
8. DNase I.
9. qPCR kit.
10. qPCR Instrument.

2.5 DNA Methylation Analysis

1. 0.1 % SDS in sterile DNase- and RNase-free water.
2. Block incubator.
3. Thermal cycler.
4. Bisulfite conversion kits.
5. DNA polymerase for bisulfite PCR.
6. Gel extraction kit.
7. Primer for amplification of bisulfite-converted DNA (Designed by using MethPrimer (http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi), [14]).
8. 200-μl PCR tubes.
9. PCR DNA polymerase.
10. 1.5 % agarose gel.
11. Sequencing kit.
12. Lysis solution (0.1 % SDS, 1 mg/mL proteinase K).

3 Methods

3.1 Superovulation of Mice

Daly 1

1. Administer 5 U–7.5 U of PMSG into 8-week-old female mice by intraperitoneal injection.

Day 3

2. 46–48 h after injection of PMSG, inject the mice with 5 U of hCG.

For collection of fertilized embryos, mate the hCG-treated female with male mice.

3.2 Preparation of Mouth Pipettes for Embryo Transfer

1. To make a mouth pipette for transfer and manipulation of embryos, heat a glass capillary in a flame, and when it becomes pliable, pull it immediately.
2. Cut the pulled glass capillary using a diamond pencil or scissors.
3. The diameter of the pipette tip should be around 90–100 μm.

3.3 Preparation of Injection Pipettes

1. Injection pipettes are made using a micropipette puller (Sutter).
2. Optimal conditions for pulling glass capillaries to make injection pipettes should be determined beforehand. Usually, we use the following parameters.

   Heat 850

   Pull 20

   Vel 150

   Delay 100

   Pressure 500
3. Pore the tip of pulled glass capillaries using a beveller. Beveling also enables pipettes to penetrate smoothly into oocytes (*see* **Notes 1** and **2**). For a detailed description about beveling of pipettes, see the chapter of Douglas Kline in [15].

3.4 Preparation of Holding Pipettes

1. Holding pipettes are also made using a puller (Sutter).
2. Optimal conditions for pulling glass capillaries should be determined before making those for use in injection. Usually, we use the following parameters.

   Heat 820.

   Pull 20.

   Vel 160.

   Delay 100.

   Pressure 0.

If you get one "good-quality" holding pipette, you can use it for an entire experiment.

3.5 Preparation of In Vitro Embryo Culture Medium

1. Place 50-μl drops of M16 on 35-mm dishes and cover them with mineral oil or liquid paraffin to prevent evaporation.
2. For pH equilibration of drops, place these dishes in a 37 °C CO_2 incubator overnight.
   Day 4

3.6 Collection of Oocytes or Fertilized Embryos from Superovulated Mice

1. After 16–18 h of hCG treatment, sacrifice superovulated mice and collect mature oocytes from ampullae of oviduct tubes (*see* **Note 3**).
2. Dilute the stock 1 % hyaluronidase 10 fold with M2 medium (final conc.: 0.1 % hyaluronidase).
3. Place the oocytes in the hyaluronidase solution, incubate in a 37 °C incubator for 3–5 min and remove the cumulus cells by pipetting the cumulus-oocyte complex.
4. Wash the embryos by pipetting the M2 drops at least four times.
5. Incubate the collected oocytes in pre-equilibrated M16 medium until microinjection.

3.7 Preparation of an Injection Chamber

1. Place a 200-μl drop of M2 medium in a 100-mm dish (Fig. 2a).
2. In the same dish, make 1–2 μl drops containing the nucleic acids that will be used for microinjection.
3. Cover these drops with mineral oil.

3.8 Preparation of Holding Pipettes

1. Put the tip of a holding pipette into a drop of M2 medium and wait for 1 min (Fig. 3).
2. Attach the holding pipette to a microinjector and fill it with M2 medium by aspirating the M2 drop in the microinjection chamber.

3.9 Preparation of Injection Solution

1. For knockdown of pancRNAs, prepare siRNAs that target pancRNAs and control siRNA as a negative control (final concentration: 2 nM).
2. To visualize whether nucleic acids were truly injected into the embryo, prepare a fluorescent-gene-expressing plasmid (final concentration: 1–5 ng/μl).

3.10 Preparation of Injection Pipettes

1. Using a mouth pipette, put some of the nucleic acid solution into an injection pipette (Fig. 4).
2. Attach the injection pipette to a microinjector and fill the tip of the injection pipette with the desired nucleic acid solution by aspirating the nucleic acid solution on the microinjection chamber (Fig. 2b, *see* **Note 4**).

3.11 Injection of siRNA/DNA

1. Before the injection, confirm that embryos contain visible pronuclei under a microscope. Usually, pronuclei are clearly observed around 24 h after hCG treatment.

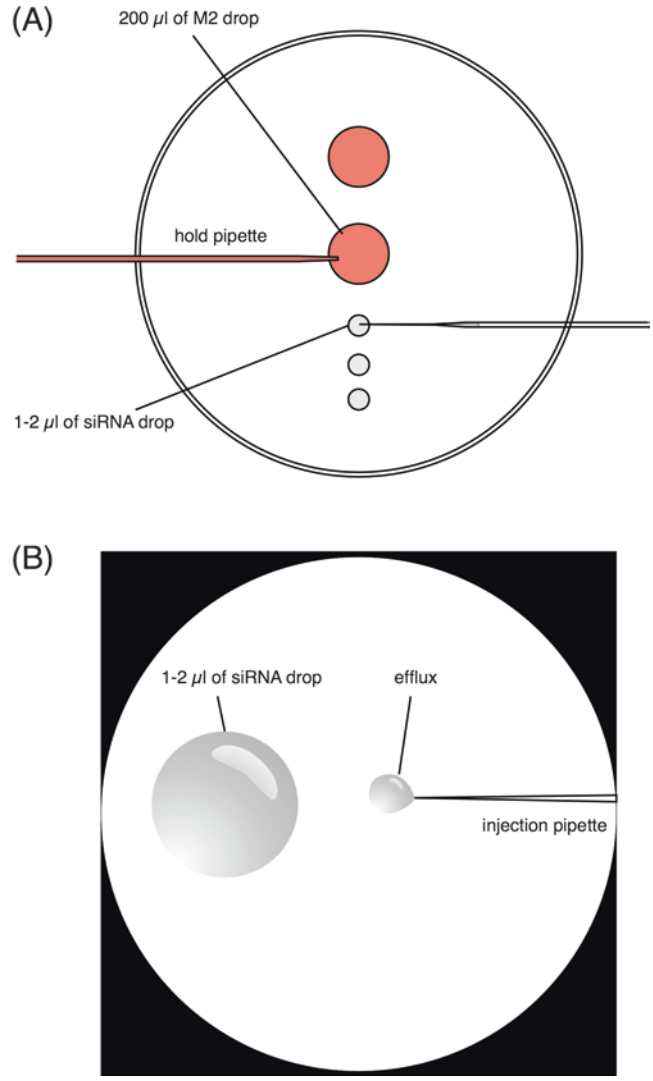


Fig. 2 (**a**) Diagram of injection chamber. A drop containing 200 μl of M2 medium is located at the center of a 100-mm dish for manipulation of embryos. Another drop of M2 should be located near the central M2 drop for use as a replacement of the old M2 medium of the holding pipette by fresh M2 medium. (**b**) Efflux is pushed out from the injection pipette

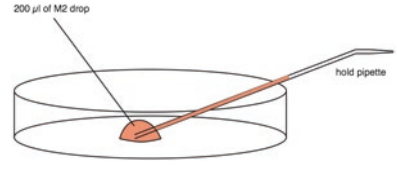


Fig. 3 Method of filling a holding pipette with M2 medium. The tip of a holding pipette is put into a 200-μl drop of M2 medium, drop and is held there for a few minutes

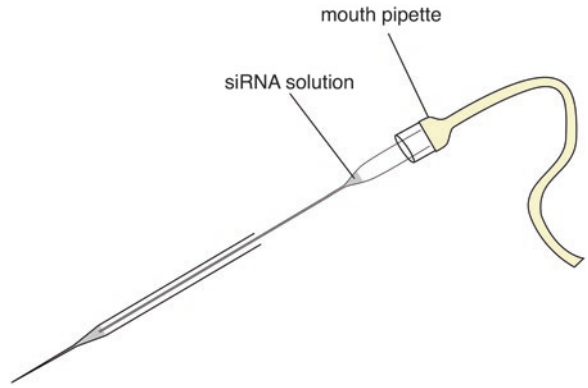


Fig. 4 Filling injection pipette with siRNA solution by using a mouth pipette

2. Pick 10–20 embryos from the embryo pool and move them into an M2 drop in the microinjection chamber.
3. Hold the embryos using the holding pipette and microinject the siRNA-containing solution.
4. If siRNA-containing solutions were properly injected into the paternal pronucleus, you can see the pronucleus swelling immediately.
5. Gently pull out the injection pipette from the embryo (*see* **Note 5**).
6. In this step, approximately 1–5 pl of siRNAs should be injected into the pronucleus.
7. Return the embryos into M16 medium (*see* **Note 6**) as soon as possible and incubate the embryos in a 37 °C chamber.

Day 5

3.12 Analysis of RNA Expression Levels and DNA Methylation Status

1. When the embryos grow up to the 2-cell stage, you can definitively see the EGFP fluorescence derived from the expression plasmid, which means that the siRNA or overexpression construct has also been successfully delivered (Fig. 5, *see* **Notes 7** and **8**).
2. Collect the GFP-positive embryos for RNA expression analysis (Subheading 3.12, **step 1**) or DNA methylation analysis (Subheading 3.12, **step 2**).

3.12.1 RNA Extraction and Reverse Transcription from Microinjected Embryos

1. Transfer 50–100 microinjected embryos into a 1.5-ml tube using a mouth pipette.
2. Extract RNAs using a RNA extraction kit according to the manufacturer's instructions.
3. Dissolve the pellet in 10 μl of DEPC-treated water and vortex briefly.
4. Prepare the following solution.

   sample 10.0 μl.

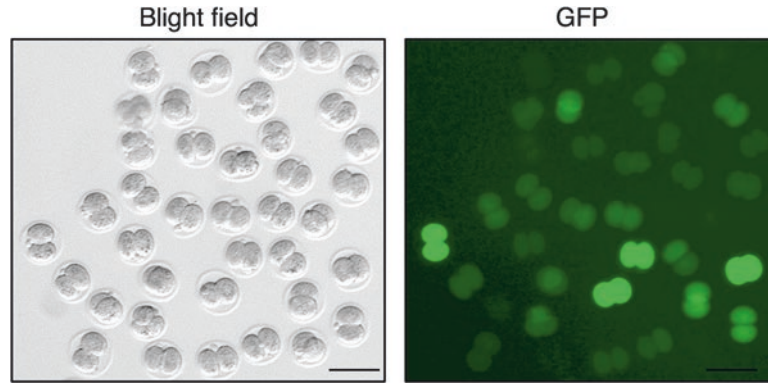


Fig. 5 GFP signals of injected 2-cell embryos. Scale bar indicates 100 μm

1U/μl DNaseI 1.0 μl.

10x buffer 1.5 μl.

DEPC-ddH2O 2.5 μl.

Total volume 15.0 μl.

5. Incubate for 15 min, at 37 °C.
6. Stop the reaction by adding 1.5 μl of 25 mM EDTA, and mix.
7. Add 2 μl of 50 ng/μl random or oligo dT primer, and mix.
8. Incubate for 15 min at 70 °C.
9. Place samples on ice for 2 min.
10. Perform a reverse transcription reaction using reverse transcription kit according to the manufacturer's instructions. For polyA-plus RNAs, use oligo dT primer. For polyA-minus RNAs, use a random primer for reverse transcription.
11. Quantify the expression levels of target RNAs of control and siRNA-injected embryos using qPCR enzyme according to the manufacturer's instruct.

3.12.2 Bisulfite Treatment and DNA Methylation Analysis

1. Put 10–50 embryos into a 200-μl PCR tube.
2. Add 15 μl of lysis solution and incubate for 60 min at 37 °C, and then 15 min at 98 °C, then carry out bisulfite treatment using an DNA Methylation kit (*see* **Note 9**).
3. Perform bisulfite reaction using bisulfite conversion kits according to the manufacturer's instructions.
4. Use 1–4 μl of the bisulfite-converted DNA for PCR amplification, and then sequence the amplified fragment (*see* **Notes 10** and **11**).

An example of PCR reaction mix

Bisulfite-converted DNA 1–4 μl

10× PCR buffer 2 μl

Primer Fw (2 nM) 2 μl

Primer Rv (2 nM) 2 μl
dNTP 2 μl
$MgCl_2$ 2 μl
H_2O to 20 μl
Taq 0.1 μl
total 20 μl

5. Set the sample in the thermal cycler and perform PCR.
   (a) 94 °C 5 min
   (b) (43 cycles)
   (c) 94 °C 10 sec
   (d) 60 °C 30 sec
   (e) 72 °C 30 sec
   (f) 72 °C 10 min
   (g) 4 °C hold
6. Perform agarose gel electrophoresis and purify the amplified DNA using a gel extraction kit (*see* **Note 12**).
7. Perform standard TA-cloning using competent *E. coli*.

Day 6

8. Pick up and transfer each colony into 10 μl of H_2O.
9. Boil the samples (98 °C, 5 min), and then store them at −20 °C or lower temperature until use.
10. In order to check which samples contain DNA-inserted plasmid, perform 43 cycles of PCR.
    Boiled colonies 1 μl
    H_2O to 10 μl
    2× AmpliTaq Gold 360 Master Mix 5 μl
    total volume 10 μl
11. Load 2 μl of each sample onto an agarose gel and perform electrophoresis.
12. Purify DNA samples of positive colonies by ethanol precipitation.
13. Perform sequencing reaction using sequencing kits and capillary sequencing according to the manufacturer's instructions (*see* **Note 13**).
14. For visualization and analysis of the results of bisulfite sequencing, we use web application QUMA (http://quma.cdb.riken.jp/top/index.html) [16] (*see* **Note 14**).

4 Notes

1. We recommend making more than 10 injection pipettes for an experiment, since injection pipettes should be changed when the tip of the pipette becomes sticky or dirty.
2. The condition of injection pipettes rapidly goes bad; therefore, injection pipettes should be made at the time of each experiment.
3. Usually, 20–35 oocytes or embryos were recovered from one B6C3F1 female mouse.
4. In order to test whether the injection pipette can eject solutions properly, move the injection pipette to the paraffin region and try to push the contents out (Fig. 2b). If there is no efflux from the pipette, replace the pipette with another one. The effluent volume can be a barometer of pipette quality. If the volume of effluent is too large (not a picoliter level), change to a different pipette.
5. If an injection pipette becomes sticky, change the pipette immediately. Otherwise, the survival rate of embryos after injection goes down.
6. If injected embryos seem to be enlarged and swollen, those embryos are dying and therefore should be discarded.
7. Usually, 60–90 % of embryos proceed to the 2-cell stage after microinjection.
8. The strength of the GFP fluorescence will vary depending on the volume of injected solution. You can classify the samples according to their brightness of GFP for the subsequent analysis.
9. We confirmed that EZ DNA Methylation-Gold Kit (Zymo research), MethylAmp DNA modification kit (Epigentek), and a MethylCode kit (Thermo Fisher Scientific) can efficiently recover bisulfite-converted DNA from small-scale samples.
10. Since bisulfite-converted DNA is fragile even when stored at −30°C, we recommend using these DNAs within 1–2 months after preparation.
11. Before you try to amplify the bisulfite-treated DNAs from small-scale samples, we strongly recommend that you determine the optimal conditions for the particular PCR primer pairs you will be using, such as the length and degree of annealing and extension, and the DNA polymerase of your choice, using bisulfite-treated DNA from control tissue samples, Especially, it should be noted that annealing temperature is one other critical parameter for the amplification of bisulfite-treated DNA.

12. Nested PCR is one of the options for amplifying "difficult" templates.
13. In order to avoid PCR bias, we subclone more than five PCR bands, and perform bisulfite sequencing of more than 20 of the resulting subclones in total.
14. On the QUMA web site, you can analyze the DNA methylation pattern and perform statistical analysis simply by uploading a zipped sequenced file and a control (not bisulfite-converted) FASTA file of the amplified region.

Acknowledgments

We thank Dr. Elizabeth Nakajima for proofreading the manuscript. We thank Katsuhiko Hayashi for discussion. This work was supported by Grants-in-Aid [No. 15H04603] to T.I. from Japan Society for the Promotion of Science (JSPS). N.H. is supported by a JSPS Research Fellowship.

References

1. Aoki F, Worrad DM, Schultz RM (1997) Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. Dev Biol 181:296–307
2. Latham KE, Garrels JI, Chang C et al (1991) Quantitative analysis of protein synthesis in mouse embryos. I. Extensive reprogramming at the one-and two-cell stages. Development 112:921–932
3. Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet 3:662–673
4. Farthing CR, Ficz G, Ng RK et al (2008) Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. PLoS Genet 4:e1000116
5. Mayer W, Niveleau A, Walter J et al (2000) Demethylation of the zygotic paternal genome. Nature 403:501–502
6. Oswald J, Engemann S, Lane N et al (2000) Active demethylation of the paternal genome in the mouse zygote. Curr Biol 10:475–478
7. Borgel J, Guibert S, Li Y et al (2010) Targets and dynamics of promoter DNA methylation during early mouse development. Nat Genet 42:1093–1100
8. Smallwood SA, Tomizawa S-I, Krueger F et al (2011) Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. Nat Genet 43:811–814
9. Smith ZD, Chan MM, Mikkelsen TS et al (2012) A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature 484:339–344
10. Imamura T, Yamamoto S, Ohgane J et al (2004) Non-coding RNA directed DNA demethylation of Sphk1 CpG island. Biochem Biophys Res Commun 322:593–600
11. Tomikawa J, Shimokawa H, Uesaka M et al (2011) Single-stranded noncoding RNAs mediate local epigenetic alterations at gene promoters in rat cell lines. J Biol Chem 286:34788–34799
12. Hamazaki N, Uesaka M, Nakashima K et al (2015) Gene activation-associated long noncoding RNAs function in mouse preimplantation development. Development 142: 910–920
13. Zhang Y, Rohde C, Tierling S et al (2009) DNA methylation analysis by bisulfite conversion, cloning, and sequencing of individual clones. Methods Mol Biol 507:177–187
14. Li L-C, Dahiya R (2002) MethPrimer: Designing primers for methylation PCRs. Bioinformatics (Oxford, England) 18:1427–1431
15. Kline D (2009) Quantitative microinjection of mouse oocytes and eggs. Methods Mol Biol 518:135–156
16. Kumaki Y, Oda M, Okano M (2008) QUMA: quantification tool for methylation analysis. Nucl Acids Res 36:W170–5

INDEX

Sara Napoli (ed.), *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1543,
DOI 10.1007/978-1-4939-6716-2, © Springer Science+Business Media LLC 2017

L

M

N

O

P

Q

R

S

T

U

V

W

Z