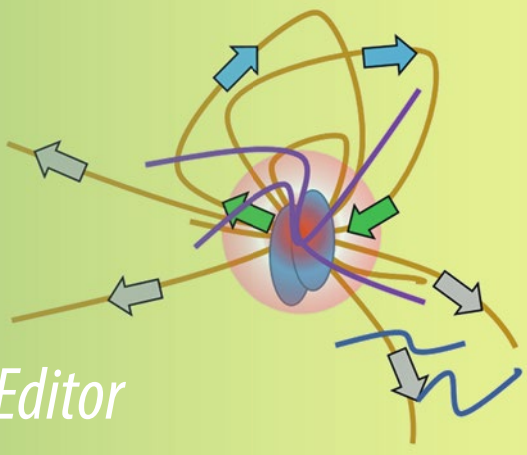
Methods in Molecular Biology 1468

Springer Protocols



Ulf Andersson Ørom *Editor*

Enhancer RNAs

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

Enhancer RNAs

Methods and Protocols

Edited by

Ulf Andersson Ørom

Max Planck Institute for Molecular Genetics, Berlin, Germany

Humana Press

Editor
Ulf Andersson Ørom
Max Planck Institute for Molecular Genetics
Berlin, Germany

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-4033-2 ISBN 978-1-4939-4035-6 (eBook)
DOI 10.1007/978-1-4939-4035-6

Library of Congress Control Number: 2016949091

© Springer Science+Business Media New York 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC New York

Preface

Long noncoding RNAs (long ncRNAs) are being continuously studied as more and more functions are described and they are found to be involved in several processes. Since the discovery that enhancer RNA (eRNA) and enhancer-like long ncRNAs are transcribed from enhancers and contribute to the activity of enhancers [11, 15] and are widely expressed from active enhancers [1, 3, 8, 13, 21], a substantial number of papers have elaborated further on their importance and mechanisms [11, 16]. With this series issue on enhancer-associated RNA, the techniques for both individual transcript studies and genome-wide and transcriptome-wide analyses are provided. The ever-increasing demand for high-throughput data creates a need to understand and apply both wet-lab and dry-lab techniques to explore the full potential of how long ncRNAs can provide functionality in enhancer function.

RNA localization is important for the function of the transcript. While many mRNAs are predominantly localized to the cytoplasm, long ncRNAs are often enriched in the nucleus [4]. This seems to reflect that they are often involved in transcription, but also poses a challenge for the functional studies. Isolating the cellular fractions of interest and the ability to endogenously detect ncRNA at enhancers and target them specifically for decay are important tools to address their functions.

Enhancers and promoters are looped in the chromosome conformation to come into proximity in the nucleus despite a longer distance along the linear chromosome [2]. Long ncRNAs have been shown to affect this process in some instances [12] and while this interaction is important for enhancer function it is still being studied what the impact of enhancer-associated RNA is for chromatin looping [5, 6, 12, 14, 19].

The predominant mechanistic model derived from current studies is that long ncRNAs work through interactions with proteins, often transcription factors, by recruiting them or evicting them from target genes and their promoters to modulate transcription [7, 9, 10, 12, 17, 18, 20, 22].

It is therefore important to measure the impact on transcription genome-wide to properly assess the consequences of enhancer-associated RNA manipulation. Such manipulation can be done either by targeting the transcripts with siRNA or antisense oligos, by activating their transcription targeted by CRISPR-activation, or directly by manipulating the genomic locus by inserting or deleting DNA regions using CRISPR/Cas9.

Finally, this issue provides means of annotating long ncRNAs and exploring transcription by assessing where transcription starts and generally how it occurs.

It is my hope that this issue can contribute to more researchers addressing the importance and mechanisms of long ncRNAs transcribed from enhancers to fully understand this important and developing field.

Berlin, Germany *Ulf Andersson Ørom*

References

1. 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
2. Bulger M, Groudine M (2011) Functional and mechanistic diversity of distal transcription enhancers. Cell 144:327–339
3. De Santa F, Barozzi I, Mietton F et al (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8:e1000384
4. Derrien T, Johnson R, Bussotti G et al (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res 22:1775–1789
5. Fanucchi S, Shibayama Y, Burd S et al (2013) Chromosomal contact permits transcription between coregulated genes. Cell 155:606–620
6. Gomez JA, Wapinski OL, Yang YW et al (2013) The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. Cell 152:743–754
7. Guttman M, Donaghey J, Carey BW et al (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295–300
8. Hah N, Murakami S, Nagari A et al (2013) Enhancer transcripts mark active estrogen receptor binding sites. Genome Res 23:1210–1223
9. Huarte M, Guttman M, Feldser D et al (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142:409–419
10. Khalil AM, Guttman M, Huarte M et al (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106:11667–11672
11. Kim TK, Hemberg M, Gray JM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
12. Lai F, Orom UA, Cesaroni M et al (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature 494:497–501
13. Lam MT, Cho H, Lesch HP et al (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature 498:511–515
14. Melo CA, Drost J, Wijchers PJ et al (2013) eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell 49:524–535
15. Orom UA, Derrien T, Beringer M et al (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58
16. Orom UA, Shiekhattar R (2013) Long noncoding RNAs usher in a new era in the biology of enhancers. Cell 154:1190–1193
17. Rinn JL, Kertesz M, Wang JK et al (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129:1311–1323
18. Schaukowitch K, Joo JY, Liu X et al (2014) Enhancer RNA Facilitates NELF Release from Immediate Early Genes. Mol Cell 56:29–42
19. Trimarchi T, Bilal E, Ntziachristos P et al (2014) Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. Cell 158: 593–606
20. Vance KW, Sansom SN, Lee S et al (2014) The long non-coding RNA Paupar regulates the expression of both local and distal genes. Embo J 33:296–311
21. Vucicevic D, Corradin O, Ntini E et al (2015) Long ncRNA expression associates with tissue-specific enhancers. Cell Cycle 14: 253–260
22. Wang KC, Yang YW, Liu B et al (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472:120–124

Contents

Contributors

Guillaume Andrey • *Development and Disease Group, Max Planck Institute for Molecular Genetics, Berlin, Germany*

Ana Carolina Ayupe • *Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Sao Paulo, SP, Brazil*

E.L. Blanchard • *Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA*

Steven Blinka • *Department of Cell Biology, Neurobiology, and Anatomy, Medical College of Wisconsin, Milwaukee, WI, USA; BloodCenter of Wisconsin, Blood Research Institute, Milwaukee, WI, USA*

Gerd A. Blobel • *Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; The Perelman School of Medicine, The Perelman School of Medicine, Philadelphia, PA, USA*

Alexander Brown • *Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

Jocelyn Choy • *Cancer Science Institute of Singapore, Centre for Translational Medicine (MD6), National University of Singapore, Singapore, Singapore*

Thomas Conrad • *Max Planck Institute for Molecular Genetics, Berlin, Germany*

Evan Corre • *ABiMS Platform, CNRS-UPMC, Station Biologique de Roscoff, Roscoff, France*

Wulan Deng • *Transcription Imaging Consortium, Howard Hughes Medical Institute, Ashburn, VA, USA*

Thomas Derrien • *CNRS UMR6290, Dog Genetic Team, Rennes, France*

Sarah Djebali • *INRA GenPhySE, Castanet-Tolosan, France*

Stephanie Fanucchi • *Gene Expression and Biophysics Group, Synthetic Biology ERA, CSIR, Pretoria, South Africa; Division of Chemical Systems and Synthetic Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa*

Sylvain Foissac • *INRA GenPhySE, Castanet-Tolosan, France*

Melissa J. Fullwood • *Cancer Science Institute of Singapore, Centre for Translational Medicine (MD6), National University of Singapore, Singapore, Singapore; School of Biological Sciences, Nanyang Technological University, Singapore, Singapore; Agency for Science, Technology and Research (A*STAR), Institute of Molecular and Cell Biology, Singapore, Singapore; Yale-NUS Liberal Arts College, Singapore, Singapore*

Alessandro Gardini • *The Wistar Institute, Philadelphia, PA, USA*

James A. Heward • *Department of Pharmacy and Pharmacology, University of Bath, Bath, UK; Centre for Haemato-Oncology Barts Cancer Institute, London, UK*

Christophe Hitte • *CNRS UMR6290, Dog Genetic Team, Rennes, France*

Jae-Yeol Joo • *Department of Neuroscience, The University of Texas Southwestern Medical Center, Dallas, TX, USA*

J. Jung • *Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA*

Tae-Kyung Kim • *Department of Neuroscience, The University of Texas Southwestern Medical Center, Dallas, TX, USA*

W. Lee Kraus • *The Laboratory of Signaling and Gene Expression, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA; The Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA; Program in Genetics, Development and Disease, Graduate School of Biomedical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA*

Mark A. Lindsay • *Department of Pharmacy and Pharmacology, University of Bath, Bath, UK*

Venkat S. Malladi • *The Laboratory of Signaling and Gene Expression, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA; The Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA*

Annalisa Marsico • *Otto-Warburg-Laboratory, RNA Bioinformatics, Max Planck Institute for Molecular Genetics, Berlin, Germany; Department of Mathematics and Informatics, Free University of Berlin, Berlin, Germany*

Musa M. Mhlanga • *Gene Expression and Biophysics Group, Synthetic Biology ERA, CSIR, Pretoria, South Africa; Division of Chemical Systems and Synthetic Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa; Unidade de Biofisica e Expressao Genetica, Faculdade de Medicina, Instituto de Medicina Molecular, Universidade de Lisboa, Lisbon, Portugal*

Shino Murakami • *The Laboratory of Signaling and Gene Expression, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA; The Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA; Program in Genetics, Development and Disease, Graduate School of Biomedical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA*

Anusha Nagari • *The Laboratory of Signaling and Gene Expression, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA; The Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA*

Ulf A. Ørom • *Max Planck Institute for Molecular Genetics, Berlin, Germany*

Pablo Perez-Pinera • *Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

Luca Pinello • *Department of Biostatistics and Computational Biology, Harvard TH Chan School of Public Health, Dana-Farber Cancer Institute, Boston, MA, USA*

Kirthi Pulakanti • *BloodCenter of Wisconsin, Blood Research Institute, Milwaukee, WI, USA*

Sridhar Rao • *Department of Cell Biology, Neurobiology, and Anatomy, Medical College of Wisconsin, Milwaukee, WI, USA; BloodCenter of Wisconsin, Blood Research Institute, Milwaukee, WI, USA; Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, USA*

Michael H. Reimer Jr. • *Department of Cell Biology, Neurobiology, and Anatomy, Medical College of Wisconsin, Milwaukee, WI, USA; BloodCenter of Wisconsin, Blood Research Institute, Milwaukee, WI, USA*

Eduardo M. Reis • *Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Sao Paulo, SP, Brazil*
Benoit T. Roux • *Department of Pharmacy and Pharmacology, University of Bath, Bath, UK*
P.J. Santangelo • *Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA*
Katie Schaukowitch • *Department of Neuroscience, The University of Texas Southwestern Medical Center, Dallas, TX, USA*
Youtaro Shibayama • *Gene Expression and Biophysics Group, Synthetic Biology ERA, CSIR, Pretoria, South Africa; Division of Chemical Systems and Synthetic Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa*
Malte Spielmann • *Development and Disease Group, Max Planck Institute for Molecular Genetics, Berlin, Germany*
Keith W. Vance • *Department of Biology and Biochemistry, University of Bath, Bath, UK*
Wendy S. Woods • *Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
Valentin Wucher • *CNRS UMR6290, Dog Genetic Team, Rennes, France*
Xinyi Yang • *Otto-Warburg-Laboratory, Epigenomics, Max Planck Institute for Molecular Genetics, Berlin, Germany*
Guo-Cheng Yuan • *Department of Biostatistics and Computational Biology, Harvard TH Chan School of Public Health, Dana-Farber Cancer Institute, Boston, MA, USA*
C. Zurla • *Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA*

Chapter 1

Cellular Fractionation and Isolation of Chromatin-Associated RNA

Thomas Conrad and Ulf Andersson Ørom

Abstract

In eukaryotic cells, the synthesis, processing, and functions of RNA molecules are confined to distinct subcellular compartments. Biochemical fractionation of cells prior to RNA isolation thus enables the analysis of distinct steps in the lifetime of individual RNA molecules that would be masked in bulk RNA preparations from whole cells. Here, we describe a simple two-step differential centrifugation protocol for the isolation of cytoplasmic, nucleoplasmic, and chromatin-associated RNA that can be used in downstream applications such as qPCR or deep sequencing. We discuss various aspects of this fractionation protocol, which can be readily applied to many mammalian cell types. For the study of long noncoding RNAs and enhancer RNAs in regulation of transcription especially the preparation of chromatin-associated RNA can contribute significantly to further developments.

Key words Cellular fractionation, Chromatin–RNA, Nascent RNA, Long noncoding RNA, Enhancer RNA, Primary transcripts, RNA processing, RNA splicing, RNA nuclear export, RNA subcellular localization

1 Introduction

Cellular fractionation techniques have been widely used to study synthesis, processing, and trafficking of biomolecules and organelles. In recent years, this approach has been combined with novel deep sequencing technologies, yielding unprecedented insights into the biogenesis and fate of primary RNA transcripts [1–3]. We use a differential centrifugation protocol to obtain chromatin-associated, nucleoplasmic and cytoplasmic RNA for the study of nascent transcription, RNA processing, and nuclear export. While absolute separation of organelles and compartments is rarely possible even with sophisticated isolation procedures, the simple method described in this chapter robustly enriches compartment-specific RNA species.

As in common nuclear isolation protocols, cells are first homogenized by lysis with a mild detergent to release organelles and other

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_1, © Springer Science+Business Media New York 2017

cellular constituents into suspension, while leaving the nuclear envelope intact. Centrifugation of the lysate through a 24% sucrose cushion yields a pellet with purified nuclei, while the supernatant, after further clearing by high speed centrifugation, represents the cytoplasmic extract. The second fractionation step was developed by Wuarin and Schibler in 1994, who wanted to separate chromatin in complex with nascent RNA from the nucleoplasm, to test if splicing occurs co-transcriptionally [4]. To achieve this, sedimented nuclei are gently resuspended in a buffer containing 50% glycerol and then rapidly lysed by addition of 1% Igepal and 1 M Urea. This method was based on the earlier observation that ternary complexes of initiated Pol II, DNA, and RNA are resistant to high salt concentrations and detergents [5], although it is now clear that transcripts can remain stably attached to the chromatin template under these conditions even after RNA Pol II has dissociated [1, 2]. The key innovation by Wuarin and Schibler was the use of Urea instead of ionic detergents to disrupt the nuclear envelope, since this maintains the association of histone proteins, including histone H1, with the genomic DNA. As a consequence, the preserved compact chromatin structure enables precipitation of the chromatin–RNA complex by centrifugation in a tabletop centrifuge. RNA can eventually be recovered from all isolated fractions by phenol–chloroform extraction or with Trizol. Efficient isolation of RNA molecules from the different subcompartments can be verified by gel electrophoresis and by quantitative RT-PCR against compartment-specific transcripts like the cytoplasmic 7SL RNA and the chromatin-associated 45S rRNA precursor.

The protocol outlined above has been used with minor modifications by various laboratories to study the dynamics of nascent transcription, exon usage, splicing kinetics, RNA processing, transport, and decay [1, 3, 4, 6]. We have used this approach to obtain a snapshot on the short-lived chromatin-associated primary transcripts that harbor microRNA precursors [2]. We routinely use the method in our laboratory, since it is relatively easy to perform and works reproducibly with various cell types.

2 Materials

2.1 Cellular Fractionation and RNA Isolation

1. Dulbecco's Modified Eagle Medium (DMEM).
2. Tris buffered saline (PBS): 150 mM NaCl, 10 mM Tris–HCl, pH 7.4.
3. 0.25% Trypsin-EDTA, Phenol Red.
4. Refrigerated benchtop centrifuge.
5. Refrigerated microcentrifuge.
6. 1.5 ml Protein LoBind tubes.
7. 1.5 ml DNA LoBind tubes.

8. 1 M NaCl solution: weigh 58.44 g of sodium chloride in a 1 l graduated cylinder or a glass beaker. Make up to 1 l with RNAse-free water. Sterile filter through a Steritop-GP 0.22 μ filter unit.
9. 10% Igepal CA-630 solution: add 20 ml Igepal CA-630 to 150 ml of RNAse-free water in a 200 ml graduated cylinder or a glass beaker and dissolve using a magnetic stirrer. Make up to 200 ml with RNAse-free water and sterile filter through a Steritop-GP 0.22 μ filter unit.
10. Cell lysis buffer: 10 mM Tris pH 7.4, 150 mM NaCl, 0.15% Igepal CA-630. Fill about 100 ml of RNAse-free water into a 200 ml graduated cylinder. Add 30 ml of 1 M NaCl solution, 2 ml 1 M Tris pH 7.4, and 3 ml 10% Igepal CA-630 solution. Fill up to 200 ml with RNAse-free water and sterile filter through a Steritop-GP 0.22 μ filter unit.
11. Sucrose buffer: 10 mM Tris pH 7.4, 150 mM NaCl, 24% sucrose. Weigh 48 g of sucrose in a glass beaker. Add 30 ml of 1 M NaCl solution and 2 ml of 1 M Tris pH 7.4. Add RNAse-free water to a volume of about 180 ml and dissolve the sucrose on a magnetic stirrer. Make up to 200 ml with RNAse-free water in a graduated cylinder. Sterile filter through a Steritop-GP 0.22 μ filter unit.
12. PBS-EDTA: Fill 100 ml 10× PBS pH 7.4 in a 1 l graded cylinder. Add 1 ml 500 mM EDTA pH 8.0 and fill up to 1 l with RNAse-free water.
13. Glycerol buffer: 20 mM Tris pH 7.4, 75 mM NaCl, 0.5 mM EDTA, 50% Glycerol. Measure 50 ml Glycerol in a graded cylinder, add 1 ml of 1 M Tris pH 7.4, 7.5 ml of 1 M NaCl solution, 100 μl 500 mM EDTA pH 8.0. Make up to 100 ml with RNAse-free water and mix by pipetting up and down with a 50 ml pipette. Sterile filter through a Steritop-GP 0.22 μ filter unit.
14. Nuclear lysis buffer: 10 mM Tris pH 7.4, 1 M Urea, 0.3 M NaCl, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 1% Igepal CA-630. Weigh 6 g of Urea into a 100 ml graduated cylinder or a glass beaker. Add 1 ml of 1 M Tris pH 7.4, 30 ml NaCl solution, 300 μl 2.5 M $MgCl_2$ solution, 40 μl 500 mM EDTA pH 8.0, and 10 ml of 10% Igepal CA-630 solution. Make up to 100 ml with RNAse-free water and sterile filter through a Steritop-GP 0.22 μ filter unit.
15. SUPERase-in (Ambion, Life Technologies).
16. GlycoBlue (Ambion, Life Technologies).
17. Isopropanol.
18. RNAse-free water.
19. TRIzol (Life Technologies).

2.2 Agarose Gel Electrophoresis

1. 10× Tris-Borate-EDTA buffer (TBE).
2. LE Agarose.
3. 6× Glycerol loading buffer: mix 3 ml Glycerol and 7 ml of RNAse-free water in a 15 ml falcon tube. Add a trace of Orange G as loading dye (~0.05% w/v).
4. SYBR safe DNA stain.
5. Mini Cell electrophoresis chamber.

3 Methods

All centrifugation steps are carried out at 4 °C and all buffers are ice cold. Tubes should be kept on ice throughout the process.

3.1 Nuclei Isolation by Differential Centrifugation

Here we use a common detergent-based homogenization method that works well with many mammalian cell types.

1. Grow HeLa cells in a 10 cm tissue culture dish to 90% confluency (*see* **Note 1**).
2. Prepare all buffers and label all necessary tubes that you need during the fractionation process (*see* **Note 2**). Calculate the required amounts of all buffers, depending on the total number of samples in your experiment. Fractionation of cells from one 10 cm dish requires 400 µl Igepal lysis buffer, 1 ml sucrose buffer, 250 µl glycerol buffer, and 250 µl Urea buffer. Prepare about 10% more buffer than required, complement with 20 U/ml SUPERase-In, and store on ice.
3. Briefly rinse cultured cells with PBS and aspirate the liquid. Trypsinize cells by adding 1 ml of 0.25% Trypsin solution, incubate at 37 °C for 5 min in a CO_2 incubator, and stop the trypsinization reaction by adding 10 ml cold DMEM (*see* **Note 3**).
4. Transfer cell suspension into a 15 ml falcon tube, spin for 5 min at 200 × *g* in a tabletop centrifuge (*see* **Note 4**), and aspirate the supernatant (*see* **Note 5**).
5. Resuspend the cell pellet in 10 ml PBS and spin at 200 × *g* for 5 min. Remove the supernatant (*see* **Note 5**).
6. Resuspend the cell pellet in 1 ml PBS and transfer to a 1.5 ml Eppendorf tube. Spin at 200 × *g* in a microcentrifuge for 2 min. Carefully remove the supernatant.
7. Add 400 µl Igepal lysis buffer to the pellet and gently pipette up and down 3–5 times to resuspend the cells (*see* **Note 6**). Incubate on ice for 5 min.
8. In the meantime, prepare a protein LoBind 1.5 ml tube with 1 ml (2.5 volumes) of cold sucrose buffer (*see* **Note 7**).

9. Gently overlay the cell lysate on top of the sucrose buffer by slowly pipetting to the wall of the tube. The cell lysate should form a visible upper phase due to the higher density of the sucrose cushion (*see* **Note 8**).
10. Centrifuge at 3500 × *g* for **10** min (*see* **Note 9**). The resulting pellet contains cell nuclei; the supernatant contains the cytoplasmic fraction (*see* **Note 10**).
11. Clear the cytoplasmic fraction again by centrifugation at 14,000 × *g* for 1 min in a new 1.5 ml microcentrifuge tube and collect the supernatant. The cytoplasmic extract can be snap frozen in liquid nitrogen or directly used for RNA isolation with TRIzol (*see* **Note 11**). Use 1 ml of TRIzol reagent per 200 µl of the cytoplasmic extract and follow the manufacturer's instructions (*see* **Note 12**). Quantify by nanodrop. The amount of isolated RNA depends on the cell type, but ranges around 20–30 µg for 200 µl Hela cytoplasmic extract.
12. Briefly rinse isolated nuclei from **step 10** with 1 ml ice cold PBS-EDTA (*see* **Note 13**). If the pellet gets disturbed, perform a 5 s spin at 3500 × *g* before gently removing the PBS-EDTA from the nuclear pellet.
13. For isolation of total nuclear RNA, add 1 ml of Trizol reagent directly to the nuclear pellet. Resuspend the pellet by pipetting up and down with a 1 ml pipette, followed by passaging through a 21 gauge needle with a 2 ml syringe (*see* **Note 14**). Follow the manufacturer's instructions for RNA isolation (*see* **Note 12**).

3.2 Nuclear Fractionation

This step separates chromatin in complex with nascent transcripts and other chromatin-associated RNA species from the nucleoplasm.

1. To separate nuclei into nucleoplasm and chromatin, resuspend isolated nuclei in 250 µl glycerol buffer (*see* **Note 15**), then immediately add 250 µl Urea buffer. Mix by vortexing for 4 s and incubate on ice for 2 min.
2. Centrifuge the lysate at 13,000 × *g* for 2 min to precipitate the chromatin–RNA complex. Collect the supernatant with the nucleoplasm in a new tube (*see* **Note 16**). The nucleoplasm can be snap frozen in liquid nitrogen or directly used for RNA isolation with TRIzol. Use 1 ml of TRIzol for 200 µl nucleoplasmic extract (*see* **Note 12**). Quantify by nanodrop. This step yields around 10 µg nucleoplasmic RNA, depending on the cell type used.
3. Briefly rinse the chromatin pellet with PBS-EDTA.
4. To isolate chromatin-associated RNA, resuspend the chromatin pellet in 1 ml of Trizol reagent. Use a 21 gauge needle and syringe to fully solubilize the pellet (*see* **Note 14**). Follow the manufacturer's instructions for RNA isolation (*see* **Note 12**) and quantify by nanodrop. Depending on the cell type, this will yield around 20 µg of chromatin RNA.

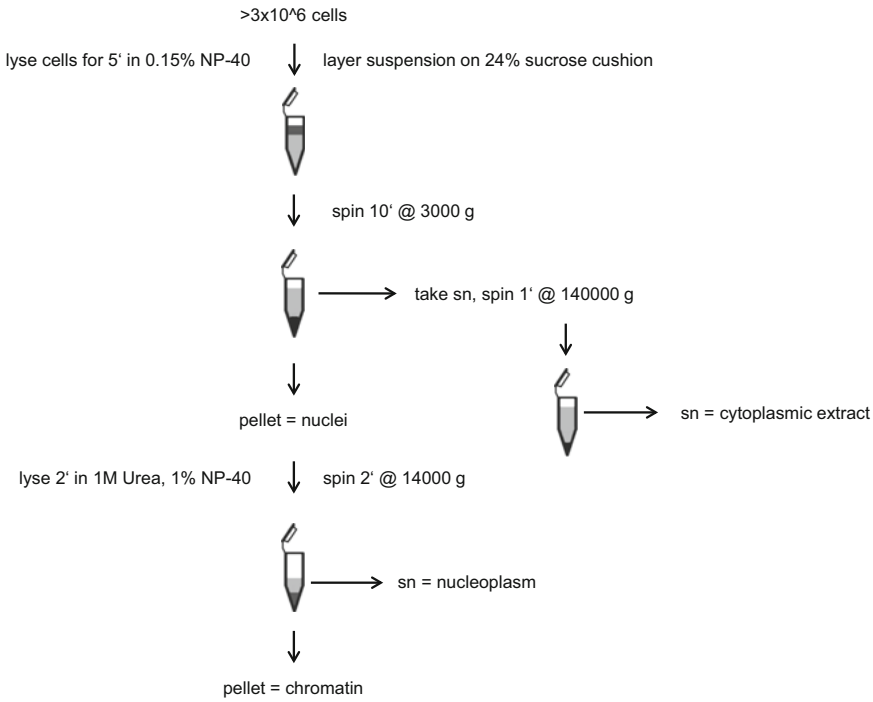


Fig. 1 Cellular fractionation scheme. All incubation times and centrifugation speeds are indicated

3.3 Gel Electrophoresis

This serves as a quick verification of a successful fractionation procedure.

1. If the amounts of isolated RNA are not limited, a quick verification of a successful fractionation can be achieved by visualizing major RNA species by electrophoresis in a 1% agarose gel (*see* **Note 17**). To this end, weigh 1 g of low melting agarose into a glass beaker and dissolve in 100 ml RNAse-free TBE buffer by boiling in a microwave oven (*see* **Note 18**). Chill the beaker in cold water and mix the solution with 2 μl SYBR safe DNA stain. Pour gel in an RNAse-free 7×10 cm gel tray using a ten-well comb.
2. Mix 1 μg of isolated RNA from each subcellular fraction with Glycerol loading Dye, load on the gel, and run electrophoresis at 120 V for 20–25 min.
3. Visualize RNA in a gel documentation instrument (Fig. 1).

4 Notes

1. A minimum cell number of 3×10^6 ensures that a pellet is clearly visible throughout all steps of the fractionation process. Depending on the cell type, 3×10^6 cells will yield tens of micrograms of RNA from each of the three subcellular fractions, which is enough for most RNA-based applications such as quantitative real-time PCR or deep sequencing.

2. Especially when preparing multiple samples simultaneously, a large number of microcentrifuge tubes are required to perform the subsequent centrifugation steps and to collect all cellular fractions. In order to ensure fast processing of the samples, it is advisable to prepare and label all required tubes and reagents before starting with the protocol.
3. This step is omitted when suspension cells are used.
4. Reduce the deceleration setting to three in order not to disturb the cell pellet when the centrifuge run stops.
5. Don't aspirate the supernatant completely in order not to disturb the cell pellet.
6. Incubation with NP-40 disrupts the plasma membrane while leaving the nuclear membrane intact. Alternative protocols use hypotonic buffer to swell the cell osmotically and then disrupt the cells mechanically in a dounce homogenizer. We have found that both methods are equally efficient and use the simpler protocol.
7. The use of low binding tubes prevents the nuclei from sticking to the wall of the tube during the subsequent centrifugation.
8. The sucrose layer thus cleans the separation of nuclei and cytoplasm. It should be noted however that mitochondria will cosediment into the pellet under these conditions. Mitochondrial membranes will be lysed together with the nuclear envelope so that matrix components will be found in the nucleoplasmic fraction.
9. Previous protocols suggested centrifugation at 14,000 rpm [6]. However, we found that high speed centrifugation can lead to nuclear lysis at this step, depending on the cell type. If premature nuclear lysis is observed, centrifugation speed can be further reduced to 1000 × *g*.
10. After this step, you may still observe a minor fraction of floating material that has not readily sedimented. This material should simply be transferred to a new tube together with the rest of the supernatant and cleared from the cytoplasmic extract by an additional high speed centrifugation.
11. Other methods of RNA isolation are proteinase K digest followed by Phenol-Chloroform or the use of resin-based columns. We use the TRIzol method due to its good reproducibility and cost-effectiveness.
12. RNA is recovered from Trizol by vigorous mixing with 200 µl chloroform, followed by 15 min centrifugation at 12,000 × *g*. The aqueous phase is collected and RNA precipitated by mixing with the same volume of isopropanol and 1 µl GlycoBlue to support RNA precipitation. After a 10 min spin at 12,000 × *g*, the RNA pellet is washed with 75 % ethanol and spun again for 5 min at 7500 × *g*. Precipitated RNA is finally resuspended in 30–50 µl RNAse-free water.

To obtain very pure RNA from Trizol samples, the aqueous phase from the first centrifugation step can be extracted a second time with 500 µl phenol at pH 4.5–5, followed by another extraction with 500 µl chloroform to remove traces of phenol. Here, the acidic pH of the phenol is critical since DNA remains in the organic phase under these conditions.

13. The purpose of this brief PBS rinse is to remove remaining cytoplasmic material and increase the purity of nuclear fractions. This step can be omitted if aggregation of nuclei is observed during subsequent resuspension in glycerol buffer.
14. Due to the high amount of DNA, nuclear and chromatin pellets can be hard to dissolve. It is usually best to first pipette the pellet up and down extensively with a 1 ml pipette tip, followed by several passages through a 21 gauge needle using a 2 ml syringe.
15. Intact nuclei are readily taken up into suspension at this stage. Extensive aggregation indicates premature nuclear lysis and leakage of DNA, which can depend on the cell type. In this case, the sedimentation speed through the sucrose cushion can be reduced to 1000 × *g* and the PBS rinse of the nuclear pellet can be omitted.
16. Additional clearing of the nucleoplasm by centrifugation is usually not necessary since the chromatin robustly precipitates into a solid white pellet.
17. This only serves as a fast verification that the experiment has worked in principle. The quality of the fractionation should be verified at high resolution by quantitative real-time PCR against cytoplasmic and chromatin-associated RNA species, such as the 7SL RNA or the 45S rRNA precursor (Fig. 2). A qPCR assay is more time-consuming compared to quick estimation on an agarose gel, but it also requires less starting material.

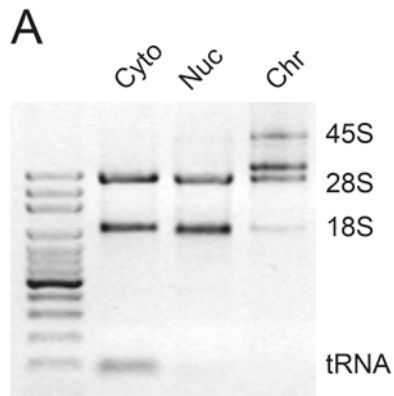


Fig. 2 Subcellular distribution of distinct RNA species shown by agarose gel electrophoresis and staining with SYBR safe. *Cyto* cytoplasm, *Nuc* nucleoplasm, *Chr* chromatin (reproduced from [2] under a CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/))

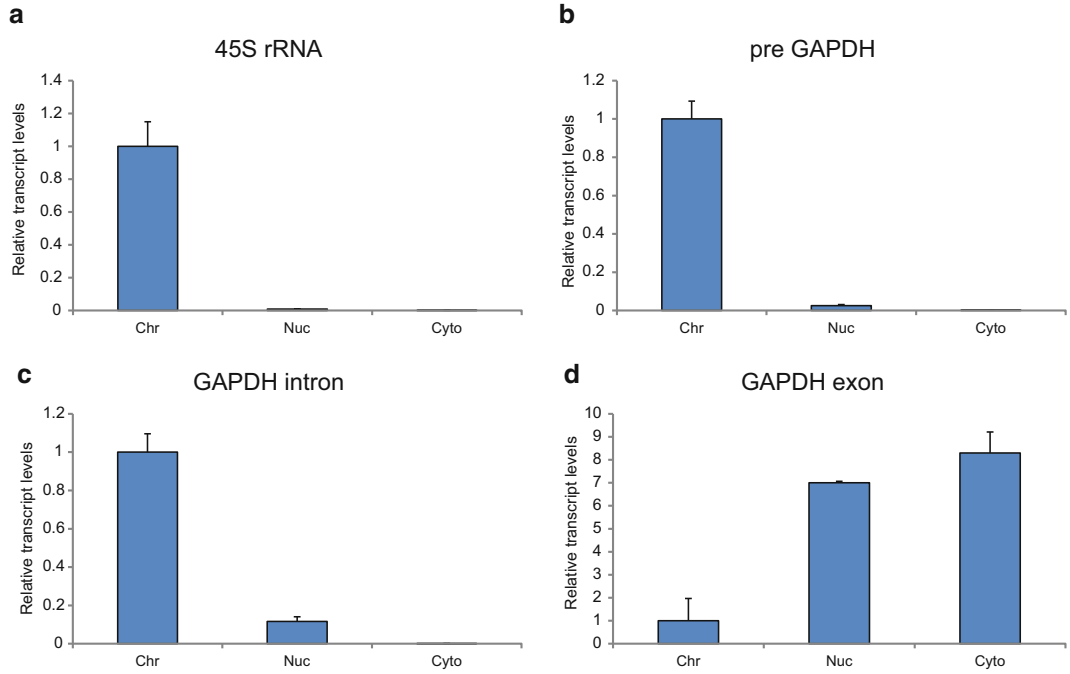


Fig. 3 RNA markers for cellular fractionation. Relative transcript levels within cellular subcompartments are shown. 1 μg of total RNA from the respective fraction was analyzed by cDNA synthesis and qPCR using primers against 45S rRNA (**a**); GAPDH pre-mRNA (**b**); GAPDH intronic sequence (**c**); GAPDH exonic sequence (**d**). Absolute expression values were then normalized to the chromatin fraction for each primer pair. Error bars show S.D. of three experiments. *Chr* chromatin, *Nuc* nucleoplasm, *Cyto* cytoplasmic (reproduced from [2] under a CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/))

18. A standard TBE agarose gel is sufficient for this purpose. In our experience, it is not necessary to use denaturing agents such as Urea or formaldehyde at this point, since the separation of rRNA and tRNA is readily visualized under nondenaturing conditions (Fig. 3).

References

1. Bhatt DM, Pandya-Jones A, Tong AJ et al (2012) Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell 150:279–290
2. Conrad T, Marsico A, Gehre M et al (2014) Microprocessor activity controls differential miRNA biogenesis in vivo. Cell Rep 9:542–554
3. Mayer A, Di Iulio J, Maleri S et al (2015) Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell 161:541–554
4. Wuarin J, Schibler U (1994) Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol Cell Biol 14:7219–7225
5. Cai H, Luse DS (1987) Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. J Biol Chem 262:298–304
6. Pandya-Jones A, Black DL (2009) Co-transcriptional splicing of constitutive and alternative exons. RNA 15:1896–1908

Chapter 2

Knockdown of Nuclear-Located Enhancer RNAs and Long ncRNAs Using Locked Nucleic Acid GapmeRs

Benoit T. Roux, Mark A. Lindsay, and James A. Heward

Abstract

The human genome is widely transcribed outside of protein-coding genes, producing thousands of noncoding RNAs from different subfamilies including enhancer RNAs. Functional studies to determine the role of individual genes are challenging with noncoding RNAs appearing to be more difficult to knockdown than mRNAs. One factor that may have hindered progress is that the majority of noncoding RNAs are thought to be located within the nucleus, where the efficiency of traditional RNA interference techniques is debatable. Here we present an alternative RNA interference technique utilizing Locked Nucleic Acids, which is able to efficiently knockdown noncoding RNAs irrespective of intracellular location.

Key words eRNA, Long ncRNA, LNA, GapmeR, Nuclear, RNAi, Enhancer

1 Introduction

The discovery of RNA interference (RNAi) in *C. elegans* [1] rapidly led to the widespread use of small interfering RNAs (siRNA) to determine the function of individual genes. siRNAs are typically 20–25 nucleotide double-stranded sequences that once transfected into cells are processed into a single strand and incorporated into the RNA-induced silencing complex (RISC), where they are able to bind target RNA through complementary base pairing to induce degradation of the transcript. Although siRNA-based functional studies are not without their caveats, in particular the occurrence of off-target effects at higher concentrations of siRNA and the difficulty in transfecting certain cell types (e.g., primary cells), they have proved to be an effective tool for determining the function of protein-coding genes.

In recent years, it has become apparent that various families of noncoding RNAs (ncRNA) play a crucial role in the regulation of protein-coding gene expression [2]. These families included the well-described microRNAs (miRNA) and the heterogeneous long noncoding RNAs (long ncRNA). It is clear that the long ncRNA family consists of a number of subfamilies, often defined on

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_2, © Springer Science+Business Media New York 2017

genomic position rather than function; these subfamilies are broadly split between those transcribed from the same locus as a protein-coding gene (including antisense, intronic, and promoter-associated long ncRNAs) and those transcribed from stand-alone genes located within intergenic regions of the genome [3]. Another recently identified family of long ncRNAs is the enhancer RNAs (eRNAs) that can be produced as mono or bidirectional transcripts and polyadenylated or non-polyadenylated transcripts. These are transcribed from enhancers and believed to regulate their action through stabilization of chromatin looping [4–6].

Functional characterisation of eRNAs and other long ncRNAs has proven to be challenging. One factor that might explain this is the observation that eRNAs, along with the majority of other long ncRNA species, are predominantly located within the nucleus [4]. While RNAi was traditionally thought to be limited to the cytoplasm, recent reports have presented evidence of RNAi occurring within nuclei, although the composition of the RISC complex may be different between the two compartments [7, 8]. In support of this contention, a number of groups have reported successful knockdown of nuclear-located long ncRNAs using siRNAs [9, 10]. However, the success rate with these nuclear transcripts appears to be significantly lower than when targeting cytoplasmic RNA, and while there may be an unknown feature of certain nuclear-long ncRNAs that renders them susceptible to RNAi, new techniques optimized for the knockdown of nuclear restricted RNA are required.

Here, we describe an alternative approach using Locked Nucleic Acid (LNA) technology to successfully target nuclear-located eRNAs and long ncRNAs, as well as cytoplasmic long ncRNAs. LNAs are modified nucleotides where the 2′C and 4′C atoms are linked by an oxymethylene bridge, increasing the affinity of the LNA for complementary sequences and thus decreasing off-target effects. LNAs are also more resistant to exo- and endonucleases, increasing in vitro and in vivo stability [11, 12]. LNA GapmeRs are single-stranded oligonucleotides antisense to the targeted RNA, normally around 15 nucleotides in length, with the most 5′ and 3′ stretch of nucleotides "locked," leaving the middle stretch of the oligonucleotide as unmodified DNA nucleotides. Upon binding of the LNA GapmeR to RNA, the central unmodified nucleotides form a DNA/RNA duplex that is recognized and cleaved by RNase H, effectively degrading the target RNA [13, 14] (Fig. 1). Given that RNase H is ubiquitously expressed throughout the cell, including within the nucleus, LNAs are able to target any RNA molecule, regardless of intracellular location.

Indeed, we have been able to use LNA GapmeRs to target both nuclear and cytoplasmic long ncRNAs (Fig. 2). As part of a prior publication, we sought to determine whether two eRNAs expressed from enhancers proximal to *IL1β* regulated its transcription upon LPS stimulation [15]. Although both eRNAs were

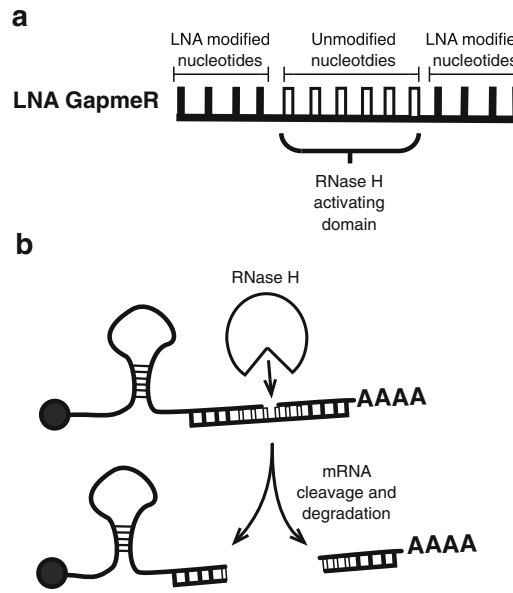


Fig. 1 Structure of LNA GapmeRs and mechanism of action. (**a**) LNA GapmeRs consist of two outer blocks of modified LNA nucleotides flanking an unmodified central stretch of bases. (**b**) LNA GapmeRs bind to RNA transcripts by complementary base paring and are only able to bind in regions devoid of secondary structure. The unmodified central bases form an RNA/DNA duplex that is recognized and cleaved by RNase H through the RNA strand, resulting in degradation of the target RNA [17]

nuclear restricted and only expressed after LPS exposure, we were able to identify one functioning LNA out of the five screened for *IL1β-RBT46*(+) and one out of three for *IL1β-eRNA*, with the third also showing partial activity (Fig. 2a). In contrast, under the same experimental conditions, four siRNAs plus a pool of the individual siRNAs were all unable to knockdown *IL1β-eRNA* (Fig. 2b). In separate studies, we targeted the constitutively expressed long ncRNA *OIP5-AS1*, which is expressed in both the cytoplasm and nucleus (data not shown), and were able to achieve knockdown with all five LNA GapmeRs examined (Fig. 2c).

Although LNA GapmeRs have the potential to induce cytotoxicity and stimulate an immune response in cells, LNA GapmeRs are less likely to be immunogenic than unmodified oligonucleotides while toxicity is highly dependent on the sequence of LNA GapmeRs [16–19]. Both these issues can be ameliorated through careful initial screening of LNA GapmeRs while the use of the lowest effective concentration and the inclusion of non-targeting negative control LNAs should prevent false-positives occurring as a result of off-target effects.

LNA GapmeRs therefore represent an excellent alternative to siRNAs and are able to target all long ncRNAs, including challenging nuclear-located transcripts such as eRNAs.

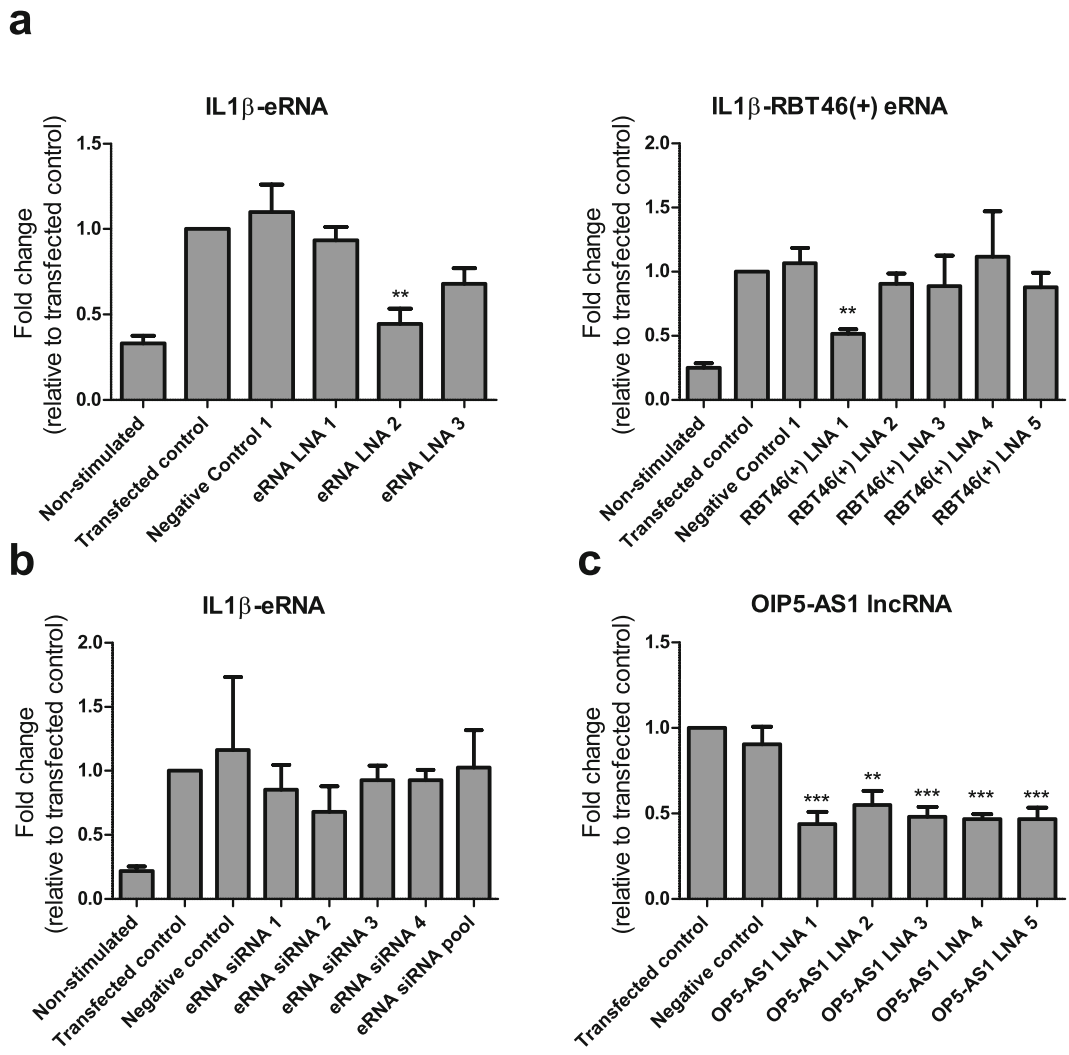


Fig. 2 Knockdown of eRNA and long ncRNAs by LNA GapmeRs. (**a**) THP-1 cells were transfected with 3 LNA GapmeRs targeting IL1β-eRNA and five targeting IL1β-RBT46(+) at a final concentration of 30 nM. THP-1 cells were then treated with buffer or 1 μM LPS for 2 h, prior to quantification of eRNA expression by qRT-PCR. Data are the mean ± SEM of five independent experiments [15]. (**b**) THP-1 cells were transfected with 4 siRNAs targeting IL1β-eRNA, plus a pool of the 4 siRNAs, at a final concentration of 30 nM. THP-1 cells were then treated with buffer or 1 μM LPS for 2 h, prior to quantification of eRNA expression by qRT-PCR. Data are the mean ± SEM of two independent experiments. (**c**) THP-1 cells were transfected with 5 LNA GapmeRs targeting the constitutively expressed long ncRNA OIP5-AS1 at a final concentration of 30 nM, prior to quantification of eRNA expression by qRT-PCR. Data are the mean ± SEM of three independent experiments. Statistical significance was determined using a one-way analysis of variance with a Dunnett's post test, where $^{**}P < 0.01$ and $^{***}P < 0.001$

2 Materials

1. Complete growth medium.
2. Antibiotic-free growth medium.
3. Serum- and antibiotic-free growth medium.
4. HiPerFect (Qiagen) *or* alterative lipid delivery reagent optimized for the cell line of interest.
5. Microcentrifuge tubes (e.g., 1.5 ml).
6. 24-Well tissue culture plates.
7. Nuclease-Free Water (non DEPC-treated).
8. LNAs GapmeRs (Exiqon).

LNAs were reconstituted with Nuclease-Free Water at 20 μM, diluted into 30 μl working stocks at 2 μM, and stored at −80 °C.

3 Methods

3.1 Transfection for Suspension Cells

3.1.1 Cell Preparation

1. 24 h before transfection, split cells to 3×10^5 cells/ml.
2. On the day of transfection, pellet the cells and resuspend at 2.5×10^6 cells per ml (*see* **Note 1**) in antibiotic-free growth medium (*see* **Note 2**).
3. Prepare sufficient cells for controls. We recommend untreated cells (No RNA or HiPerFect), HiPerFect alone (transfected control) and a minimum of one negative (scrambled) non-targeting control.
4. Seed 100 μl of the cell suspension per well (2.5×10^5 total cells), shake to ensure the cells disperse throughout the well, and incubate at 37 °C while preparing LNA transfection mixtures (*see* **Note 3**).

3.1.2 LNAs Transfection Mixture Preparation

1. Prepare a mixture of 100 μl serum- and antibiotic-free growth medium (*see* **Note 4**) and 5 μl of HiPerFect (vortex before use) for each sample (*see* **Notes 5**–**7**).
2. Add 3 μl of 2 μM LNA per tube (Final concentration 30 nM in 200 μl; *see* **Note 8**).
3. Vortex the mixtures and incubate at room temperature for 10 min.
4. Remove the prepared plate from the incubator and add 100 μl of each mixture dropwise per well, rocking the plate gently between each well and at the end (*see* **Note 9**).
5. Incubate the cells for 6–15 h at 37 °C (*see* **Note 10**).
6. Add 400 μl of complete growth medium and continue incubating cells for the desired total length of transfection (e.g., 24 h).

3.2 "Reverse" Transfection for Adherent Cells

We have found this protocol to be effective for transfecting adherent cell lines; however, the previous suspension cell transfection protocol can also be applied to adherent cell lines. *See* **Note 3** for details on how to apply the previous protocol to adherent cells.

3.2.1 Cell Preparation

1. Detach cells (e.g., by scraping or trypsin incubation), pellet, and resuspend in antibiotic-free growth medium at a concentration of 2×10^6/ml (*see* **Note 11**).
2. Prepare sufficient cells for controls. We recommend untreated cells (No RNA or HiPerFect), HiPerFect alone (transfected control) and a minimum of one negative (scrambled) nontargeting control.
3. Keep cell suspension in a conical centrifuge tube in a 37 °C incubator until ready. Ensure that the tube is mixed regularly to prevent the cells from clumping or adhering to the plastic.

3.2.2 LNAs Solution Preparation

1. Prepare a master-mix of 100 µl serum- and antibiotic-free growth medium (*see* **Note 2**) and 5 µl of HiPerFect (vortex before use) per sample (*see* **Notes 4** and **5**).
2. Pipette 3 µl of 2 µM LNA onto the center of the required wells of a 24-well plate.
3. Vortex and add 100 µl of the transfection mix directly onto the spotted siRNA for the required wells (*see* **Note 3**).
4. Incubate at room temperature for 10 min. Rock plate periodically to ensure the well is covered (*see* **Note 12**).
5. Remove cell suspension from the incubator and add 100 µl of the suspension (2×10^5 cells per well) to each well dropwise, rocking the plate gently between each well (*see* **Note 9**).
6. Incubate the cells in a 37 °C-humidified incubator for 6–15 h (overnight) at 37 °C (*see* **Note 10**).
7. Add 400 µl of complete growth medium and continue incubating cells for the desired total length of transfection (e.g., 24 h).

4 Notes

1. This concentration is for 2.5×10^5 cells per well, a number we have found effective when using monocytic cell lines. This may be adjusted accordingly for different cell lines; resuspend the cells at a concentration 10× higher than the desired so that the required number of cells is seeded in 100 µl.
2. We suggest using antibiotic-free medium in order to reduce any potential toxicity caused by inadvertent transfection of the antibiotics.

3. When using this protocol for adherent cells, seed the appropriate number of cells 24 h before transfection so that the cells will be 70–80% confluent on the day of transfection. Aspirate the medium, replace with 100 μl of antibiotic-free growth medium, and then proceed with the remainder of the protocol.
4. Proteins contained within the serum can interfere with the formation of HiPerFect/LNA micelles; use serum- and antibiotic-free growth medium at this stage to prevent this.
5. We have found HiPerFect to be an effective reagent for transfecting monocytic cell lines; however, lipid-based transfection reagents display starkly different transfection efficiencies and toxicity upon different cell lines. Choice of the most appropriate reagent is crucial for success in transfecting LNAs.
6. There is generally a trade-off between using higher volumes of transfection reagents and increasing transfection rates, with corresponding increases in toxicity induced by the transfection reagent. We have found 5 μl of HiPerFect to provide a sensible compromise between effective transfection rates and toxicity. This may be adjusted between 3 and 9 μl; however, it is vital to determine the degree of toxicity and not just transfection efficiency.
7. The Nucleofector has proved to be an effective alternative to lipid-based reagents when transfecting challenging cell lines, e.g., primary cells. Although further work is required to determine the efficiency of using the Nucleofector to knockdown nuclear-located eRNAs and long ncRNAs, it has been reported to deliver nucleic acids directly to the nucleus and to be able to knockdown other species of nuclear ncRNAs through the transfection of antisense oligonucleotides modified in a similar fashion to LNA GapmeRs [20].
8. Although LNAs are reported to have lower off-target effects than siRNAs, they are generally active at similar concentrations. We therefore recommend not exceeding guidelines for siRNAs to avoid any risk of off-target effects. 30 nM (in 200 μl) should be sufficient to induce knockdown without off-target effects [21–23].
9. We suggest "rocking" or "shaking" the plate side to side, rather than "swirling," in order to prevent cells from being drawn and clumping to the edge of the wells. This is especially important when using adherent cells.
10. When targeting constitutively expressed genes, we typically transfect cells for 6 h in 200 μl before adding 400 μl of complete growth medium for an additional 18 h. To knockdown inducible genes, our standard protocol is to transfect the cells overnight (~15 h) in 200 μl and then add 400 μl of complete growth medium the following morning, with the stimuli added at this point for the required length of time.

11. This number of cells is typically used for epithelial cell lines that grow in a monolayer. Other cell types may require adjustment according to size and rate of growth. Cells should be seeded to achieve 50–80% confluence on the day of transfection.
12. The volume of liquid (media, transfectection reagent and LNA) at this stage (100 µl) will only just about cover the surface of the well. Regularly, rock the plate to ensure the LNA/micelles coat the entire well and do not pool around the edges.

References

1. Fire A, Xu S, Montgomery MK et al (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811
2. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157:77–94
3. Iyer MK, Niknafs YS, Malik R et al (2015) The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47:199–208
4. 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
5. Natoli G, Andrau J-C (2012) Noncoding transcription at enhancers: general principles and functional models. http://dxdoiorg/101146/annurev-genet-110711-155459 46:1–19
6. Ørom UA, Shiekhattar R (2013) Long noncoding RNAs usher in a new era in the biology of enhancers. Cell 154:1190–1193
7. Gagnon KT, Li L, Chu Y et al (2014) RNAi factors are present and active in human cell nuclei. Cell Rep 6:211–221
8. Robb GB, Brown KM, Khurana J, Rana TM (2005) Specific and potent RNAi in the nucleus of human cells. Nat Struct Mol Biol 12:133–137
9. Lam MTY, Cho H, Lesch HP et al (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature 498(7455):511–515
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
11. Braasch DA, Corey DR (2001) Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. Chem Biol 8:1–7
12. Veedu RN, Wengel J (2010) Locked nucleic acids: promising nucleic acid analogs for therapeutic applications. Chem Biodivers 7: 536–542
13. Kurreck J, Wyszko E, Gillen C, Erdmann VA (2002) Design of antisense oligonucleotides stabilized by locked nucleic acids. Nucleic Acids Res 30:1911–1918
14. Wahlestedt C, Salmi P, Good L et al (2000) Potent and nontoxic antisense oligonucleotides containing locked nucleic acids. Proc Natl Acad Sci U S A 97:5633–5638
15. IIott NE, Heward JA, Roux BT et al (2014) Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. Nat Commun 5:3979
16. Whitehead KA, Dahlman JE, Langer RS, Anderson DG (2011) Silencing or stimulation? siRNA delivery and the immune system. Ann Rev Chem Biomol Eng 2:77–96
17. Deleavey GF, Damha MJ (2012) Designing chemically modified oligonucleotides for targeted gene silencing. Chem Biol 19:937–954
18. Kole R, Krainer AR, Altman S (2012) RNA therapeutics: beyond RNA interference and antisense oligonucleotides. Nat Rev Drug Discov 11:125–140
19. Stanton R, Sciabola S, Salatto C et al (2012) Chemical modification study of antisense gapmers. Nucleic Acid Ther 22:344–359
20. Ideue T, Hino K, Kitao S et al (2009) Efficient oligonucleotide-mediated degradation of nuclear noncoding RNAs in mammalian cultured cells. RNA 15:1578–1587
21. Semizarov D, Frost L, Sarthy A et al (2003) Specificity of short interfering RNA determined through gene expression signatures. Proc Natl Acad Sci U S A 100:6347–6352
22. Persengiev SP, Zhu X, Green MR (2004) Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). RNA 10:12–18
23. Jackson AL, Bartz SR, Schelter J et al (2003) Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol 21:635–637

Chapter 3

Visualization of Enhancer-Derived Noncoding RNA

Youtaro Shibayama, Stephanie Fanucchi, and Musa M. Mhlanga

Abstract

Enhancers are principal regulators that allow spatiotemporal tissue-specific control of gene expression. While mounting evidence suggests that enhancer-derived long noncoding RNAs (long ncRNAs), including enhancer RNAs (eRNAs), are an important component of enhancer function, their expression has not been broadly analyzed at a single cell level via imaging techniques. This protocol describes a method to image eRNA in single cells by in situ hybridization followed by tyramide signal amplification (TSA). The procedure can be multiplexed to simultaneously visualize both eRNA and protein-coding transcript at the site of transcriptional elongation, thereby permitting analysis of dynamics between the two transcript species in single cells. Our approach is not limited to eRNAs, but can be implemented on other transcripts.

Key words RNA visualization, Fluorescence in situ hybridization, Tyramide signal amplification, Enhancers, Long ncRNA, eRNA, Single cell analysis

1 Introduction

An enhancer element was first observed when a piece of SV40 DNA remotely activated the beta-globin gene in-cis as far as thousands of bases away [1]. The viral DNA could be placed in either orientation and could act in many positions, both upstream and downstream, relative to the activated gene. The regulatory function of enhancers has now been demonstrated by decades of research, which has established that enhancers are critical for metazoan cells to generate cell- and tissue-type-specific gene expression programs in response to developmental and environmental cues. The number of putative enhancers, as predicted by chromatin marks and transcription factor binding, vastly outnumber that of protein-coding genes in the human genome, suggesting a highly complex usage of enhancers in achieving a precise temporal and spatial control of gene expression [2].

Transcription of an active enhancer was first reported at the beta-globin locus [3, 4], but it was the recent advances in sequencing technology that surprisingly revealed that enhancer sites are

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_3, © Springer Science+Business Media New York 2017

pervasively transcribed [5]. Transcription at enhancers occurs bi-directionally by RNA Pol II to produce a class of long ncRNA termed eRNA. eRNA transcripts are mostly unspliced and non-polyadenylated and have a median length of 346 nucleotides [6]. Levels of eRNA production have been correlated to those of induced nearby genes, suggesting that enhancer transcription may be used to gauge enhancer activity [5, 6]. Measuring eRNA levels by the highly sensitive CAGE sequencing to identify active enhancers, a recent enhancer atlas has included 43,011 such elements across the majority of human cell types and tissues [6]. Remarkably, the abundance of enhancer transcripts is 19- to 34-fold lower than that of gene transcripts [6].

Owing to advances in techniques based on chromosome conformation capture (3C), it is now widely accepted that enhancers exert their effect on target promoters by existing in close proximity in three-dimensional space [2]. This raises the possibility that eRNAs are merely transcriptional noise that correlates to the activation of nearby genes. A number of studies, however, suggest that eRNAs functionally contribute to the induction of target genes. For example, enhancer knockdown experiments have resulted in the reduction in transcription of specific nearby genes [7–10]. In addition, eRNA tethering experiments using fusion constructs targeted to reporter genes have shown that the eRNA itself, rather than the act of enhancer transcription, is necessary for the activation of the reporter [7, 9]. Furthermore, an inversion of enhancer sequence, producing an eRNA with a completely different sequence, abolished enhancer activity, suggesting that a specific eRNA sequence is necessary for its function [8].

It is important to note that all reported studies on eRNA thus far, to the best of our knowledge, have been conducted on bulk populations of cells. As the observed widespread transcription of enhancers and their correlation with induced coding genes have been from a cell population, it remains unknown whether this same picture exists in single cells or it is merely an "averaged out" view of the population, leaving a serious gap in our understanding of eRNA function. This is especially true in cases where multiple enhancers have been observed to simultaneously regulate the expression of a single gene in a cell population, as we cannot rule out the possibility that each cell produces only one of the multiple eRNAs. It has been suggested that each large cluster of enhancers, called stretch- or super-enhancer [11, 12], known to harbor multiple eRNAs, acts as a single regulatory unit within which all eRNAs are uniformly up- or down-regulated to control gene expression [13]. Such claims, however, can only be verified by single-cell studies.

One way to study transcription in single cells is by fluorescence in situ hybridization (FISH) and more specifically single molecule FISH (smFISH). While RNA FISH has been robustly applied to visualize mRNA in a wide variety of cells and tissues, the technique

is not easily implemented on transcripts of short length such as eRNAs, due to the limitation in the number of signal-generating probes that can hybridize to the target. The problem is further compounded by the supposed low copy-number of eRNAs.

Here we introduce a protocol that we have used to overcome these limitations to successfully visualize an eRNA of average length in primary Human Umbilical Vein Endothelial Cells (HUVECs) stimulated with TNFα. The method is based on using singly biotinylated short (20 nt each) probes, which, post hybridization, allows the binding of multiple copies of fluorophore per probe. Probe design, preparation and hybridization, and mounting of coverslips are adopted from the method used for smFISH [14, 15]. High-density fluorescence labeling of each probe is mediated by tyramide signal amplification (TSA) [16, 17], which relies on the catalytic activity of horseradish peroxidase (HRP) to activate dye-labeled tyramide (Fig. 1). Briefly, biotinylated probes are first hybridized to

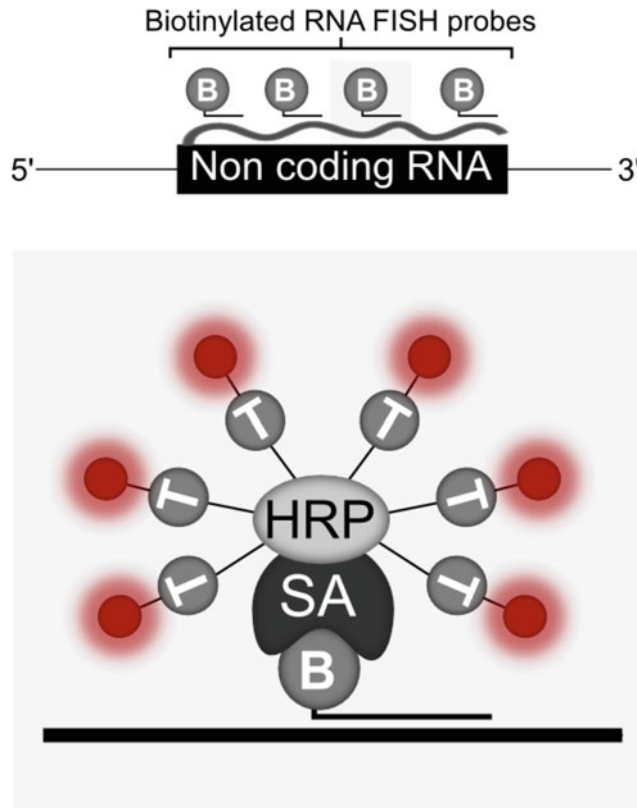


Fig. 1 Schematic representation of the TSA method. Biotinylated probes bind to target RNA, after which streptavidin–HRP conjugate is introduced to bind the biotin. The HRP then activates dye-labeled tyramide, resulting in the accumulation of multiple copies of dye per probe. *B* biotin, *SA* streptavidin, *HRP* horseradish peroxidase, *T* tyramide

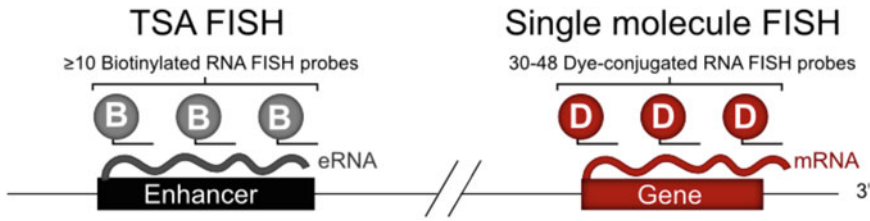


Fig. 2 Schematic representation of the TSA method multiplexed with smFISH. Enhancer transcript is targeted by a small number (~10) of biotinylated probes, while intronic RNA from a protein-coding gene is targeted by a larger number (~30–50) of dye-labeled probes. *B* biotin, *D* dye

the target eRNA in fixed cells, followed by the introduction of streptavidin–HRP conjugate which binds to biotin. Dye-labeled tyramide is then added, which is activated by HRP to multiply bind the probe–biotin–streptavidin–HRP complex. The signal is readily detected by a wide-field fluorescence microscope. Using the long ncRNA HOTTIP (which in itself is an enhancer-like long ncRNA, although much larger in length compared to typical eRNAs) [18] in HeLa cells as a control, we first show that the TSA method reliably produces similar results to smFISH in terms of spot counts when the exonic portion is targeted. We then target an eRNA in HUVECs by signal amplification from 12 probes, resulting in clear punctate spots. Importantly, TSA can be multiplexed with intronic smFISH (Fig. 2) to achieve simultaneous detection of both eRNA and the induced nascent gene transcript in the same nucleus. As introns are degraded shortly after transcriptional elongation, the intronic FISH spots represent sites of active transcription [19, 20]. We observe diffraction-limited co-localization of eRNA and pre-mRNA spots, suggesting that the observed TSA spots are derived from bona-fide enhancer transcription. This concurrently illustrates that eRNA occurs at the active site of gene transcription. Fundamentally, the method described here can be used to visualize not only eRNA but also other long ncRNA or mRNAs.

2 Materials

2.1 Equipment

1. High-pressure liquid chromatograph (HPLC) equipped with a reverse-phase C-18 column and a dual wavelength detector (*see* **Note 1**).
2. CO_2 incubator and laminar flow-hood.
3. Wide-field fluorescence microscope equipped with a mercury lamp, appropriate filters, 100× objective of high numerical aperture (>1.3) and CCD camera.

2.2 Reagents

Prepare all reagents with RNase-free DEPC-treated water.

2.2.1 Preparation of Probe

1. Oligos targeting eRNA and coding gene transcript (*see* Subheading 3.1).
2. Amino-reactive biotin and dye (*see* **Note 2**).
3. TE buffer, pH 8.0.
4. 3 M sodium acetate, pH 5.2.
5. Dimethyl sulfoxide (DMSO).
6. 0.1 M sodium tetraborate.
7. HPLC Buffer A: 0.1 M triethylammonium acetate, pH 6.0, sterilized using 0.2 μm filter.
8. HPLC buffer B: 0.1 M triethylammonium acetate in 70% (v/v) acetonitrile, pH 6.0, sterilized using 0.2 μm filter.

2.2.2 Cell Culture

1. 24-Well cell culture dish.
2. 10–12 mm No. 1 cover slips, sterilized by successive washes in 70 and 100% ethanol, followed by exposure to UV for 15 min.
3. Appropriate cell culture medium and related reagents.

2.2.3 Fixation, Permeabilization, and Hybridization

1. 100% Methanol.
2. PBS.
3. 70% Ethanol.
4. Washing buffer: 2× SSC and 10% deionized formamide.
5. Hybridization buffer: 1 μg/μl *Escherichia coli* tRNA, 10% formamide, 2 mM vanadyl ribonucleoside complex, 10% (w/v) dextran sulfate, 0.02% RNase-free BSA. Filter in 0.2 μm filters. Store in aliquots at −20 °C.

2.2.4 Washing, TSA and Mounting of Cover Slips

1. 0.5% Triton X-100 in PBS.
2. RNase inhibitor.
3. TSA kit with HRP–streptavidin (Molecular Probes, T20936 for Alexa Fluor 647). Prepare all reagents in the kit as per instruction manual. When preparing the fresh blocking buffer, add the above RNase inhibitor at 2 units/μl and DTT at 1 mM.
4. Equilibration buffer: 2× SSC and 0.4% (w/v) glucose.
5. DAPI. Dissolved in dimethylformamide to 5 mg/ml and stored in aliquots at −20 °C.
6. Catalase from *Aspergillus niger*. Store at 4 °C.
7. Glucose oxidase (Type VII) from *A. niger*. Diluted to 3.7 mg/ml in water and stored at 4 °C.
8. Deoxygenated mounting medium: Mix equilibration buffer, catalase, and glucose oxidase preparations in 100:1:1 ratio. Prepare fresh. Store at 4 °C for a day.

3 Methods

3.1 Probe Design

Here we describe the strategies for designing probes against both the eRNA and its partner protein-coding gene transcript at its intron. These strategies are adapted from those previously described for smFISH [14, 15]. We use a publicly accessible computer program that generates a list of probes for a particular target sequence (https://www.biosearchtech.com/stellarisdesigner/). For targeting the intron of a protein-coding gene, the same rules as smFISH apply. Nearly fifty probes of ~20 nt, spaced out by 2 nt at minimum, should ideally be generated within a single intron to produce a strong fluorescence signal, but decent signals can still be obtained from as few as thirty probes. Masking level, or probe specificity against the background genome sequence, should be as high as possible (between 5 and 3) without sacrificing the number of probes.

For targeting eRNA, we advise designing at least around ten probes (*see* **Note 3**). Design parameter stringency may need to be loosened to achieve this number of probes. However, probe length should be at least 18 nt and masking level never below 3. If possible, we advise increasing probe spacing to 3–5 nt, which could potentially lower steric hindrance caused by the binding of streptavidin–HRP.

All oligos should be synthesized with a 3′ amino group modification for the subsequent labeling with dye or biotin. We order our probes from Biosearch Technologies (Novato, CA, USA), but many other manufacturers can synthesize oligos of decent quality. We usually order probes at 5 nmol scale in a 96-well format.

3.2 Probe Labeling

Probes targeting eRNA need to be labeled with biotin, while probes targeting coding transcript with dye. Fluorophores need to be chosen according to the available light source and filter set on the widefield microscope. Three dyes we routinely use for probe labeling in our lab are Atto488 (green), Atto565 (orange), and Atto647N (far red). These dyes are stable when used with the deoxygenated mounting medium and can also be easily multiplexed as their fluorescence spectra do not overlap. In the particular example given here, we have used Atto565 for the coding gene and Alexa Fluor 647 (far red; included in the kit) for the TSA.

1. Dissolve all oligos in TE buffer in equimolar concentrations. For 5 nmol scale, we dissolve each oligo in 100 μl.
2. Pool all oligos for each target. For fifty oligos, we take 10 μl of each oligo in each well. For much fewer oligos (i.e., for eRNA), it is recommended to take more. The remaining oligos can be stored at −20 °C.

3. Precipitate the pooled oligos by adding 1/10 volume of 3 M sodium acetate, pH 5.2, and 2.5 volume of 100% ethanol. Incubate at −20 °C for at least 1 h and centrifuge to pellet the DNA. Dissolve the pelleted DNA in freshly prepared 200 µl 0.1 M sodium tetraborate.
4. Take a tiny amount (0.1–1.0 mg) of dye (for coding gene) or biotin (for eRNA) into a fresh 2 ml tube. It does not have to be weighed; using a pipette tip usually works fine. Dissolve the dye/biotin in ~20 µl DMSO. Once dissolved, add 200 µl 0.1 M sodium tetraborate.
5. Mix the DNA and the dye/biotin together. Incubate at 37 °C for 6 h to overnight to allow conjugation to the DNA.
6. Precipitate the conjugated oligos by adding 1/10 volume of 3 M sodium acetate, pH 5.2, and 2.5 volume of 100% ethanol. Incubate at −20 °C for at least 1 h and centrifuge to pellet the DNA. Dissolve the pellet in 100–300 µl HPLC Buffer A. This can be stably stored at −20 °C or preferably −80 °C for years.

3.3 Probe Purification

Labeled oligos need to be purified from unlabeled oligos as well as free dye or biotin. Create an HPLC program that initiates with 2% Buffer B that rises linearly to 98% over 20 min. Labeled oligos are more hydrophobic than unlabeled ones, causing longer retention on a reverse-phase column. All pooled oligos are usually eluted in a single peak. It is uncommon but possible for this peak to be divided into smaller sub-peaks, all of which should be collected. For an example of how these peaks usually appear, refer to Batish et al. [15].

1. Load the sample onto the HPLC column that is initially equilibrated with 2% Buffer B and start the program. Monitor the absorption at both 260 nm and the maximal absorption wavelength for the dye. Both unlabeled oligos and biotin-labeled oligos will create a peak of absorption at 260 nm only. However, unlabeled oligos will be eluted earlier from the column and will therefore create a peak before the biotin-labeled ones. Dye-labeled oligos will absorb at both 260 nm as well as the dye-specific wavelength.
2. Two peaks should be observed. The first peak always corresponds to unlabeled oligos. Collect all samples from the second peak only. Depending on the dye used, the time between the first and the second peaks will vary. A peak from oligos labeled with biotin usually appears very rapidly after the first peak.
3. Salt-ethanol precipitate the collected sample as before. Dissolve the pellet in TE buffer and measure DNA concentration, which ideally should be at least 50 ng/µl. Store the purified labeled probes in aliquots at −20 °C.

3.4 Cell Culture, Fixation, and Permeabilization

Follow regular procedures for cell culture. We grow HUVECs to ~80% confluency on sterile No.1 cover slips (10–12 mm) in 24-well dishes. The cells are serum-starved for 18 h and then stimulated with TNFα for 1 h to induce the expression of TNF responsive genes. While 4% formaldehyde is a common choice of fixative for mRNA FISH, we do our fixation in methanol when eRNA or other long ncRNA is involved, as it denatures proteins and as a result improves probe hybridization. Although methanol fixation causes the nuclear stain to become blurry, it still permits the outline of the nucleus to be observed.

1. Aspirate the culture medium and gently wash cells with PBS.
2. Fix cells in ice-cold 100% methanol for 10 min at −20 °C. Wash twice in PBS.
3. Add 70% ethanol and incubate at 4 °C for at least 2 h to permeabilize the cells. Cells on coverslips can be stored in 70% ethanol at 4 °C for a few days.

3.5 Hybridization

The protocol below describes multiplexing of hybridization to target both the eRNA and protein-coding transcript, but the same procedure applies when targeting just a single species. Since our protocol uses short probes, formamide concentration is kept low at 10% during both hybridization and washes (*see* **Note 4**) [14, 15]. A good starting point for probe concentration is 1 ng/μl. Keeping the hybridization temperature constant at 37 °C, probe concentration can be lowered when high background noise occurs (*see* **Note 4**).

1. Make a platform for the hybridization reaction by covering a glass plate with a clean sheet of parafilm. Press the parafilm firmly onto the glass so that it does not peel off.
2. Place the parafilm-covered glass plate into a hybridization chamber. An empty tip box makes a good chamber. The glass plate can be rested on the stage of the tip box where the tips usually sit. Pour some water into the bottom compartment of the tip box to keep the entire chamber humidified during hybridization.
3. Add appropriate amount of probe for both eRNA and protein-coding transcript to 50 μl hybridization buffer to make a final concentration of 1 ng/μl for each probe set. Probe stock should be concentrated enough to not considerably dilute the hybridization buffer. Mix by pipetting.
4. Equilibrate the cells by replacing the 70% ethanol in the well containing the cover slip with wash buffer and incubate at room temperature for at least 2 min.
5. Place the entire volume of probe in hybridization buffer on the parafilm covering the glass plate to form a droplet. Using fine forceps, pick up the cover slip and carefully blot the edge on paper towel to remove excess wash buffer. Gently place the cover slip onto the droplet of hybridization solution, cell side facing down. Avoid air bubbles.

6. Close the lid of the hybridization chamber and incubate at 37 °C overnight. Cover slips should be protected from light. For extra humidity, we place the hybridization chamber on a platform rising just above the water level inside a 37 °C water bath.

3.6 Washes

1. Pick up the cover slip and gently blot the edge on paper towel to remove excess hybridization solution. Place it in a well containing wash buffer, cell side facing up. Gently rotate on a shaker for 30 min at room temperature. Protect from light.
2. Repeat the wash with fresh wash buffer.
3. For multiplexed TSA FISH and smFISH, carry on with the TSA step below. If only the protein-coding gene was targeted using dye-labeled probes (i.e., smFISH only), skip the TSA step below and continue with counterstaining.

3.7 Tyramide Signal Amplification

For signal amplification from a small number of biotinylated probes, we use the TSA kit from Molecular Probes (catalogue number T20936, containing HRP–streptavidin and Alexa Fluor 647 tyramide). Reagents are prepared exactly as according to the kit instruction manual, with a single minor modification. As a precautionary measure, when preparing the fresh 1% blocking reagent, we add RNase inhibitor (RNaseOUT, Invitrogen, catalogue number 10777-019) to a final concentration of 2 units/μl and DTT to a final concentration of 1 mM. The protocol is adopted from the kit manual, with another modification. We include a second permeabilization step to improve the nuclear penetration of streptavidin–HRP complexes. Importantly, signal from the fluorescently labeled probe will persist through the TSA procedure. Remember to keep the cover slip protected from light at all stages.

1. Replace the wash buffer with 0.5% Triton X-100 in PBS and incubate at room temperature for 15 min. Briefly rinse twice in PBS.
2. Place the cover slips in 2× SSC and incubate at room temperature for 2 min.
3. In the box that was used as the hybridization chamber, replace the parafilm covering the glass plate with a clean one. Place a droplet of 50 μl blocking reagent containing RNase inhibitor and DTT on the parafilm. Pick up the cover slip, blot the edge gently on paper towel to remove excess solution, and place it onto the droplet, cell side facing down. Avoid air bubbles. Close the lid to retain humidity and incubate for 30 min at room temperature.
4. Prepare a working solution of streptavidin–HRP by diluting the stock solution 1:100 in blocking reagent containing RNase inhibitor and DTT. 50 μl is more than sufficient per cover slip.
5. On a clean spot on the same parafilm, place a droplet of 50 μl streptavidin–HRP working solution. Pick up the cover slip, blot the edge on paper towel, and transfer it onto the droplet.

Avoid air bubbles. Incubate in the humidified box for 30 min at room temperature.

6. Wash the cover slip three times by immersion in PBS for 5 min each at 37 °C.
7. Prepare an Alexa Fluor 647-tyramide working solution by diluting the stock solution 1:100 in amplification buffer/0.0015% H_2O_2.
8. On a clean spot on the same parafilm, place a droplet of 50 μl Alexa Fluor 647-tyramide working solution. Pick up the cover slip, blot the edge, and place it on the droplet. Avoid air bubbles. Incubate in the humidified box for 10 min at room temperature.
9. Repeat the wash described in **step 6**.

3.8 Counterstaining and Mounting

It is crucial to counterstain the nucleus when targeting intronic RNA or eRNA, as signal is only expected to be observed in the nucleus (*see* **Note 5**). Mounting is done as previously described [14, 15] in deoxygenated mounting medium (*see* **Note 6**).

1. Counterstain the nuclei by immersion of the cover slip in 15 nM DAPI in PBS for at least 2 min at room temperature. Rinse briefly twice with PBS.
2. Immerse the cover slip in equilibration solution and allow the cover slips to equilibrate for at least 2 min at room temperature.
3. Place 5 μl freshly prepared deoxygenated mounting medium on a clean glass slide. Pick up the cover slip, blot the edge, and place it onto the mounting medium, cell side facing down. Avoid air bubbles. Remove excess mounting medium by gentle blotting with paper towel.
4. Seal the cover slip by applying a thin coat of nail polish around the edge. Do not let the nail polish enter into the medium below the cover slip.

3.9 Imaging and Image Processing

As a proof of principle to show the robustness of the TSA method, we have included images comparing smFISH and TSA FISH on HOTTIP long ncRNA (Fig. 3). The same oligos were used to prepare the two sets of probes, except the TSA set included half the number of probes as the smFISH set (24 vs. 48 probes). Both methods show roughly equal frequency of spot counts of ~55% per allele.

Unlike exonic smFISH where spots of near equal intensity are generally observed throughout the cytoplasm and nucleus, intronic smFISH produces one or two intense spots in the nucleus only (*see* **Note 7**).

Although most eRNAs are unspliced [6], current research points to these transcripts working in cis, at sites of active gene transcription. Co-localization of FISH spots from eRNA and intronic portion of gene transcript is therefore expected, which is what we observe in a HUVEC nucleus (Fig. 4).

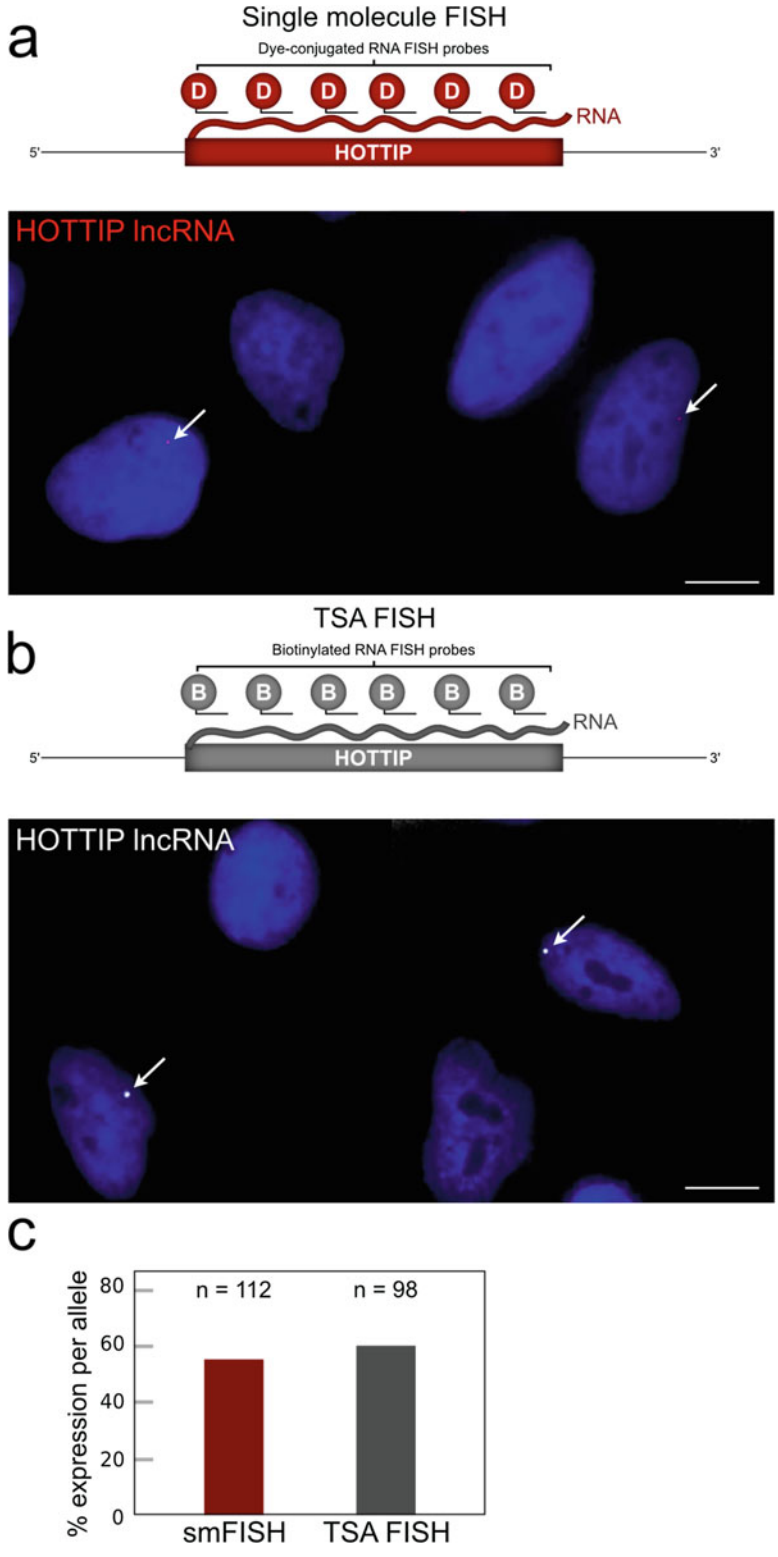


Fig. 3 smFISH and TSA FISH produce similar spot counts of HOTTIP long ncRNA, demonstrating that the TSA method is as robust as intronic smFISH. (**a**) HOTTIP exon visualized by smFISH in HeLa cells. (**b**) HOTTIP visualized by TSA FISH in HeLa cells using half the number of probes. (**c**) Both methods produce similar spot counts of ~55 %/allele. Arrows point to FISH spots. Bar = 10 μm

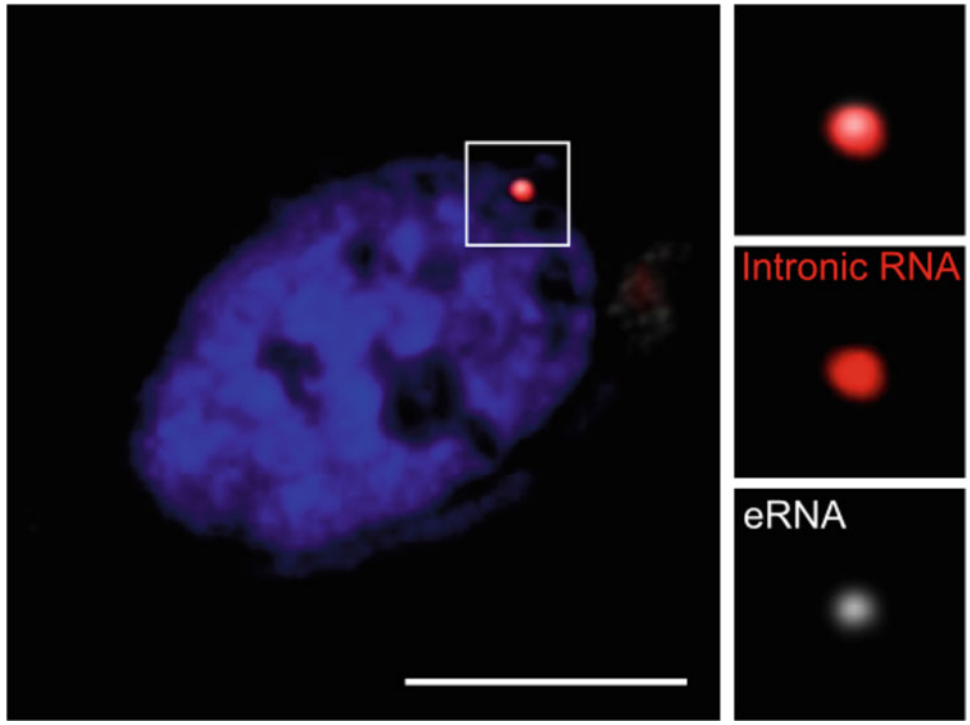


Fig. 4 Multiplexed FISH image showing eRNA and intron of a protein-coding gene at its site of active transcription. Bar = 10 μm

1. For each fluorescence channel, obtain *z*-stacks of 0.2–0.3 μm spacing with the 100× objective. The entire thickness of the nucleus should be covered with the *z*-stacks.
2. Go through the *z*-stacks to make sure the observed spots are within the nucleus. A good way to present images is by maximum intensity projection of the stacks. ImageJ/Fiji is a good publicly available software for processing.

4 Notes

1. Although we conjugate our probes and subsequently purify them using HPLC, it is possible to purchase probes that are prelabeled with either dye or biotin.
2. We frequently purchase our dyes from Atto-Tec (Siegen, Germany), but other companies such as Invitrogen and Amersham Bioscience also provide similar products of good quality. Three dyes used routinely in our lab are Atto488, Atto565, and Atto647N. Be sure to obtain amino-reactive versions (succinimidyl ester or thiocyanate) of both these dyes and biotin for subsequent conjugation to oligos. An example of biotin we use is from Molecular Probes (B6352).

3. It is a good idea to start by first checking the particular enhancer is transcribed in the population of specific cell-type to be tested. For primary cells, we have used the publicly available enhancer atlas (http://enhancer.binf.ku.dk/enhancers.php) which was constructed based on the transcriptional activity of these elements in a vast range of cell types [6]. It is important to note that some enhancers included there are transcribed bidirectionally and that the given genomic coordinates include transcripts from both the top and bottom strands of DNA. The mid-position of each bidirectional eRNA can be acquired from the BED files listing the eRNAs for each cell type. Depending on the length of the particular eRNA, probes may be designed for either the top or bottom strand, or both, in order to attain sufficient number of probes. When both strands of eRNA are targeted simultaneously in the same color, single punctate spots can still be expected, as the two transcriptional events take place in close proximity in a coordinated manner. It is also a good idea to confirm transcription of the particular enhancer in the cell population by RT-qPCR.
4. Hybridization and washing stringencies can be controlled by adjusting the formamide concentration. For mRNA FISH using many short probes (i.e., for the coding gene), 10% formamide usually works well. Off-target binding of some probes do not cause considerable background, as the fluorescence signal caused by the accumulation of nearly fifty probes at the actual target is much stronger. However, for few number of probes, as is the case for eRNA, off-target binding may pose a serious issue due to the lower ratio of probes binding between the target and off-target sites. If significant background is observed, probe concentration may be lowered. Alternatively, increase the formamide concentration or the hybridization/washing temperatures.
5. eRNA has been reported to exist in the cytoplasm [6], but currently we are only interested in nuclear transcripts.
6. It is important to use the deoxygenated mounting medium to minimize photobleaching during illumination. Light-mediated degradation of fluorophores requires oxygen, which can be enzymatically removed by glucose oxidase and catalase in the presence of glucose [21].
7. These are sites of active transcription, where multiple rounds of transcriptional elongation result in many copies of the target intron at one or both alleles. Spot intensity will therefore correspond to locus activity, which may not necessarily be identical between different alleles or cells. As splicing and intron degradation typically occurs rapidly following transcriptional elongation, it is rare to observe signal elsewhere in the cell. More than two spots may be observed in polyploid cells.

Acknowledgments

We acknowledge contributions from Robyn Brackin who helped develop the TSA technique in our lab. This work was supported by grant PG-V2KYPO7 from the Council for Industrial and Scientific Research (CSIR, South Africa) and by a grant from the Emerging Research Area Program of The Department of Science and Technology (DST, South Africa) and grant PTDC/SAU-GMG/115652/2009 from the Fundação para a Ciência e a Tecnologia (FCT, Portugal), all to Musa M. Mhlanga.

References

1. Banerji J, Rusconi S, Schaffner W (1981) Expression of beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27:299–308
2. Levine M, Cattoglio C, Tjian R (2014) Looping back to leap forward: transcription enters a new era. Cell 157:13–25
3. Collins P, Antoniou M, Grosveld F (1990) Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high level expression. EMBO J 9:233–240
4. Tuan D, Kong S, Hu K (1992) Transcription of the hypersensitive site HS2 enhancer in erythroid cells. Proc Natl Acad Sci U S A 89:11219–11223
5. Kim TK, Hemberg M, Gray JM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
6. 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
7. Li W, Notani D, Ma Q et al (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498:516–520
8. Lam MTY, Cho H, Lesch HP et al (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature 498:511–515
9. Melo CA, Drost J, Wijchers PJ (2013) eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell 49:524–535
10. Mousavi K, Zare H, Dell'orso S et al (2013) eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. Mol Cell 51:606–617
11. Parker SC, Stitzel ML, Taylor DL et al (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A 110:17921–17926
12. Whyte WA, Orlando DA, Hnisz D et al (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153:307–319
13. Hah N, Benner C, Chong LW et al (2015) Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. Proc Natl Acad Sci U S A 112:E297–E302
14. Raj A, van den Bogaard P, Rifkin SA et al (2008) Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods 5:877–879
15. Batish M, Raj A, Tyagi S (2011) Single molecule imaging of RNA in situ. In: Gerst JE (ed) RNA detection and visualization: methods and protocols, vol 714, Methods in molecular biology. Springer, Heidelberg, pp 3–13
16. Bobrow MN, Harris TD, Krista H et al (1989) Catalyzed reporter deposition, a novel method of signal amplification application to immunoassays. J Immunol Methods 125:279–285
17. van Gijlswijk RPM, Zijlmans HJMAA, Wiegant J et al (1997) Fluorochrome-labeled tyramides: use in immunocytochemistry and fluorescence in situ hybridization. J Histochem Cytochem 45:375–382
18. Wang KC, Yang YW, Liu B et al (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472:120–124
19. Levesque MJ, Raj A (2013) Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nat Methods 10:246–248
20. Fanucchi S, Shibayama Y, Burd S et al (2013) Chromosomal contact permits transcription between coregulated genes. Cell 155:606–620
21. Yildiz A, Forkey JN, McKinney SA et al (2003) Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. Science 300:2061–2065

Chapter 4

UV-RNA Immunoprecipitation (UV-RIP) Protocol in Neurons

Katie Schaukowitch, Jae-Yeol Joo, and Tae-Kyung Kim

Abstract

With the many advances in genome-wide sequencing, it has been discovered that much more of the genome is transcribed into RNA than previously appreciated. These nonprotein-coding RNAs (ncRNAs) come in many different forms, and they have been shown to have a variety of functions within the cell, influencing processes such as gene expression, mRNA splicing, and transport, just as a few examples. As we delve deeper into studying their mechanisms of action, it becomes important to understand how they play these roles, in particular by understanding what proteins these ncRNAs interact with. This protocol describes one technique that can be used to study this, ultra-violet light cross-linking RNA immunoprecipitation (UV-RIP), which uses an antibody to pull down a specific protein of interest and then detects RNA that is bound to it. This technique utilizes UV light to cross-link the cells, which takes advantage of the fact that UV light will only cross-link proteins and nucleic acids that are directly interacting. This approach can provide key mechanistic insight into the function of these newly identified ncRNAs.

Key words UV cross-linking, RNA immunoprecipitation, lncRNA, RNA-binding proteins

1 Introduction

New roles for long ncRNAs (lncRNAs) are constantly being discovered. However, in order to move past the initial identification and loss of function studies of long ncRNAs, it will be necessary to understand the mechanism by which these long ncRNAs are working. To this end, identifying interacting partners and the complexes in which they work is a crucial step in understanding the role they are playing. Protein interactions with DNA or RNA can be dynamic and transient, which imposes difficulty in preserving the interactions during purifications. Additionally, strong caution should be taken due to the possibility of detecting nonspecific protein–nucleic acid interactions.

This protocol describes a technique that uses UV light to cross-link proteins and nucleic acids, thereby preserving the interactions between proteins and DNA or RNA that have occurred in an intact cell. UV cross-linking utilizes the natural photoreactivity of the RNA bases, especially pyrimidines, and does not induce protein–protein cross-linking [1–3]. Another benefit of using UV

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_4, © Springer Science+Business Media New York 2017

cross-linking comes from the fact that the UV light is able to form a covalent bond between protein and nucleic acids only when they are within Angstroms of each other, thereby detecting only direct protein–nucleic acid interactions [4]. The strength of the covalent bond allows for harsher washing conditions, while the distance constraint ensures detection of direct interaction partners. Together, these properties allow for greater specificity in the pull down, which is advantageous as nonspecificity can be an issue with more traditional cross-linking methods. Formaldehyde in particular is known to be able to form cross-links not only between proteins and nucleic acids but also protein–protein interactions, which can lead to the pull down of whole complexes that may be interacting indirectly. This UV-RIP technique can then provide evidence that the components in question are directly interacting.

Our system uses primary cortical neuronal cultures from embryonic mice to study long ncRNAs that are induced in response to depolarization of the neurons. These long ncRNAs are expressed from enhancers of activity-induced genes, termed eRNAs, and therefore the cells need to be depolarized to stimulate expression [5]. eRNAs are expressed at very low levels within the cell, as compared to their target protein-coding mRNAs, and are not very stable, contributing to their transient induction. UV-RIP has successfully been used to identify binding partners of eRNAs as well as another class of activating long ncRNAs (ncRNA-a) [6, 7]. Other cell culture or tissue conditions can be used based on the system of choice. While this protocol allows one to identify specific RNA transcripts that are bound to a protein of interest, other protocols have been designed to look at interacting transcripts on a genome-wide scale, such as CLIP (cross-linking immunoprecipitation)-based methods, in which an RNA-binding protein is pulled down and then all associated transcripts are sequenced [8–10]. One advantage of CLIP is that due to the protein adduct left behind, it is possible to determine the exact site of binding along the transcript.

2 Materials

2.1 Buffers

For all buffers, protease inhibitors and RNase inhibitor should be added right before use.

1. KCl Depolarization Buffer: 170 mM KCl, 2 mM $CaCl_2$, 1 mM $MgCl_2$, 10 mM Hepes, pH 7.4.
2. Low-salt Lysis Buffer: 50 mM Hepes KOH, pH 7.5, 10 mM NaCl, 1 mM EDTA, pH 8.0, 10% glycerol, 0.2% NP-40, 1% Triton X-100, Protease Inhibitors, RNasin Plus (50 U/mL).
3. High-salt Lysis Buffer: 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 10 mM Tris–HCl, pH 8.0, 600 mM NaCl, 1% Triton X-100, 0.1% Sodium Deoxycholate (DOC), Protease Inhibitors, RNasin Plus (50 U/mL).

4. IP Buffer: 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 10 mM Tris–HCl, pH 8.0, 1% Triton X-100, 0.1% DOC, Protease Inhibitors, RNasin Plus (50 U/mL).
5. Low-salt Wash Bufkfer: 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8.1, 150 mM NaCl.
6. High-salt Wash Buffer: 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8.1, 500 mM NaCl.
7. LiCl Wash Buffer: 0.25 M LiCl, 1% IGEPAL CA630, 1% deoxycholic acid (sodium salt), 1 mM EDTA, 10 mM Tris, pH 8.1.
8. Elution Buffer: 10 mM Tris, pH 8.0, 1 mM EDTA, pH 8.0, 1% SDS, RNasin Plus (50 U/mL).

2.2 Tissue Culture

1. Poly-D-lysine (50 mg/mL).
2. Plating media: Advanced DMEM supplemented with 10% FBS and Glutamax.
3. Neurobasal media supplemented with B-27 and Glutamax.
4. Tetrodotoxin (TTX; 1 μM).

2.3 Other (See Note 1)

1. Protease Inhibitors: cOmplete Protease Inhibitor Cocktail Tablets (Roche). A 50× stock solution can be made by dissolving 1 tablet in 1 mL H_2O.
2. RNAse Inhibitor: RNasin Plus (Promega), 40 U/μL stock.
3. PBS: Phosphate-buffered saline, tissue culture grade.
4. Protein A/G plus beads.
5. Proteinase K (20 mg/mL).
6. Phenol:Chloroform:Isoamyl alcohol (25:24:1, v/v).
7. 3 M sodium acetate (NaOAc), pH 5.2.
8. Glycogen (20 μg/μL).
9. 100% Molecular Biology Grade Ethanol.
10. Nuclease-free water.
11. DNAse I (2 U/μL).
12. cDNA Reverse Transcription kit.

3 Methods

3.1 Cell Culture

1. Primary cortical neurons are plated on 150 mM × 25 mM dishes coated with Poly-D-lysine in plating media, at 3×10^7 cells/dish (*see* **Note 2**).
2. Change plating media to Neurobasal/B-27/Glutamax after 2 h.
3. Neurons at days in vitro (DIV) 5 are made quiescent with 1 μM TTX overnight. On DIV 6, neurons are stimulated with 55 mM KCl depolarization buffer for 30 min. 5×10^7 cells were used per IP.

3.2 UV Cross-linking

1. Remove culture media and add 10 mL ice-cold PBS with proteinase inhibitors.
2. Place dish, on ice, in Stratalinker UV-light box without the lid.
3. UV cross-link at 400 mJ/cm^2.
4. Collect cells by cell scraper and transfer to a conical tube on ice.
5. Spin cells in a benchtop centrifuge at 2000 rpm (872 × *g*) for 5 min at 4 °C, then remove PBS.

It is possible to snap freeze the cell pellet and store at −80 °C, but if doing so, RNase inhibitor should be added in addition to protease inhibitors.

3.3 Cell Lysis and RNA Immunoprecipitation (See Note 3)

1. Resuspend cell pellet in ice-cold Low-salt lysis buffer (10 mL for 100 M cells).
2. Incubate on a rotating platform at 4 °C for 10 min.
3. Centrifuge at 2000 rpm (872 × *g*) for 10 min at 4 °C.
4. Remove supernatant, resuspend the nuclei pellet in 1 mL High-salt lysis buffer, and rotate at 4 °C for 1 h.
5. Centrifuge at 2000 rpm (872 × *g*) for 10 min at 4 °C, remove supernatant, and then transfer to new 1.5 mL tube.
6. Add 1 mL of IP buffer (containing protease inhibitor and RNase inhibitor) and mix, and then aliquot 1 mL for IgG and 1 mL for IP.
7. Save 1/20th of the lysate for Input. Store at 4 °C overnight.
8. Add 10 μg of antibody/IP to the pre-cleared lysate and incubate on a rotating platform overnight at 4 °C.
9. The next day, wash protein A/G plus beads twice with cold PBS and once with diluted high-salt lysis buffer (300 mM NaCl final). Use 20 μL final bead volume per IP.
10. Add washed beads to the lysate–antibody mix and incubate for at least 2 h at 4 °C.
11. Spin down the beads in a microcentrifuge at 6000 rpm (3300 × *g*) for 1 min at 4 °C.
12. Remove the supernatant and wash as follows: 2× low-salt Wash Buffer, 2× high-salt Wash Buffer, 2× LiCl Wash Buffer, 1× TE. For each wash, rotate beads for 5–10 min at 4 °C before spinning down at 6000 rpm (3300 × *g*) for 1 min at 4 °C.

3.4 Elution

1. Add 150 μL Elution Buffer to the beads.
2. Place tubes in a 65 °C heat block for 10 min, with gentle vortexing every 2 min.
3. Spin down the beads at 6000 rpm (3300 × *g*) for 1 min at 25 °C. Transfer the supernatant to a new tube.

4. Repeat the elution on more time for a final elution volume of 300 µL.
5. Bring up input samples to 300 µL with elution buffer.
6. Add 7 µL Proteinase K to each sample and incubate 2 h at 50 °C.
7. Extract RNA by phenol:chloroform extraction, followed by ethanol precipitation.

Add an equal volume of phenol:chloroform:isoamyl alcohol (300 µL) to each sample and shake vigorously for 30 s or vortex briefly and incubate for 3 min at room temperature.

8. Centrifuge at max speed in a microcentrifuge (>16,100 × *g*) for 15 min at 4 °C.
9. Transfer the aqueous phase to a new 1.5 mL tube and add 1/10th volume 3 M NaOAc (30 µL) and 1 µL Glycogen (20 µg/µL) as carrier.
10. Add 3× volume 100% EtOH (900 µL).
11. Gently mix and then precipitate overnight at −80 °C.
12. Centrifuge at max speed for 15 min at 4 °C and then discard supernatant.
13. Wash with 1 mL 75% EtOH.
14. Centrifuge at max speed for 10 min at 4 °C and then discard EtOH.
15. Let RNA pellet air dry.
16. Resuspend in 10–20 µL nuclease-free water.

3.5 Detecting RNA (See Note 4)

1. Treat 10 µL of RNA with 1 µL of DNaseI and incubate for 20 min at 37 °C.
2. Reverse transcribe RNA using a reverse transcription kit, with a final reaction volume of 20 µL.
3. Dilute cDNA 1:2.
4. Perform RT-qPCR using primers targeting your RNA of interest. Use 2 µL of diluted RT product in a 10 µL reaction volume.

All RT-qPCR products should be checked by running on a 6% PAGE gel for the expected sizes. Sequencing can also be performed to ensure detection of the correct target.

4 Notes

1. As the readout for the interaction is RNA, special care should be taken to avoid RNase contamination. RNase inhibitor (RNasin) should be added whenever indicated, gloves should always be worn, and RNase-free tubes and pipette tips used. RNaseAway (Ambion) can also be used.

2. Amounts of cells and antibody needed vary depending on the cell type and individual antibody (*see* Subheadings 3.1 and 3.3). These conditions should be optimized accordingly.
3. Although UV cross-linking allows for stringent purification methods, it is still subject to nonspecific association of uncross-linked RNA and/or protein during immunoprecipitation. Proper negative controls should be carefully designed and performed in parallel. Independent validation would also be desired.
4. Because of the covalent and irreversible nature of the bond formed during UV cross-linking, a protein adduct will be left even after Proteinase K treatment. This has the potential to cause inefficient reverse transcription of the pulled down RNA, although the methods are still compatible [8] (*see* Subheading 3.5). Other methods of detection include Northern blotting and RNAse protection assays [11].

References

1. Hockensmith JW, Kubasek WL, Vorachek WR et al (1986) Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. J Biol Chem 261(8):3512–3518
2. Greenberg JR (1979) Ultraviolet light-induced crosslinking of mRNA to proteins. Nucleic Acids Res 6(2):715–732
3. Brimacombe R, Stiege W, Kyriatsoulis A et al (1988) Intra-RNA and RNA-protein cross-linking techniques in Escherichia coli ribosomes. Methods Enzymol 164:287–309
4. Darnell RB (2010) HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip Rev RNA 1(2):266–286
5. Kim TK, Hemberg M, Gray JM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465(7295):182–187
6. Schaukowitch K, Joo JY, Liu X et al (2014) Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell 56(1):29–42
7. Lai F, Orom UA, Cesaroni M et al (2013) Activating RNAs associate with mediator to enhance chromatin architecture and transcription. Nature 494(7438):497–501
8. Ule J, Jensen K, Mele A, Darnell RB (2005) CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods 37(4):376–386
9. Licatalosi DD, Mele A, Fak JJ et al (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456(7221):464–469
10. Huppertz I, Attig J, D'Ambrogio A et al (2014) iCLIP: protein-RNA interactions at nucleotide resolution. Methods 65(3):274–287
11. Sei E, Conrad NK (2014) UV cross-linking of interacting RNA and protein in cultured cells. Methods Enzymol 539:53–66

Chapter 5

Mapping Long Noncoding RNA Chromatin Occupancy Using Capture Hybridization Analysis of RNA Targets (CHART)

Keith W. Vance

Abstract

Capture Hybridization Analysis of RNA Targets (CHART) has recently been developed to map the genome-wide binding profile of chromatin-associated RNAs. This protocol uses a small number of 22–28 nucleotide biotinylated antisense oligonucleotides, complementary to regions of the target RNA that are accessible for hybridization, to purify RNAs from a cross-linked chromatin extract. RNA–chromatin complexes are next immobilized on beads, washed, and specifically eluted using RNase H. Associated genomic DNA is then sequenced using high-throughput sequencing technologies and mapped to the genome to identify RNA–chromatin associations on a large scale. CHART-based strategies can be applied to determine the nature and extent of long noncoding RNA (long ncRNA) association with chromatin genome-wide and identify direct long ncRNA transcriptional targets.

Key words CHART, Long noncoding RNA, Chromatin, Genome-wide binding, Oligonucleotide capture

1 Introduction

Long noncoding RNAs (long ncRNAs) have emerged as important transcription and chromatin regulatory molecules that can function locally to regulate the expression of nearby genes and also at more distal locations, away from their sites of synthesis, to regulate genome-wide expression programs. Although detailed *cis*-acting modes of action have been described for numerous long ncRNAs in the regulation of local gene expression, the molecular mechanisms used by chromatin-associated long ncRNAs to regulate global gene expression programs are not as well characterized. However, recently developed experimental techniques to identify and map the association of long ncRNAs with chromatin on a large scale are beginning to shed light on the mechanisms of long ncRNA genomic targeting and distal transcriptional regulation (reviewed in [1]).

In this chapter, I will describe the Capture Hybridization Analysis of RNA Targets (CHART) protocol that can be applied

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_5, © Springer Science+Business Media New York 2017

to identify long ncRNA genomic binding sites. This method was originally developed to determine the binding profile of the *roX2* ncRNA regulator of dosage compensation in *Drosophila* S2 cells and was later extended to map genome-wide chromatin occupancy for several mammalian long ncRNAs, including *Xist* [2, 3]. More recently, CHART has been combined with mass spectrophotometry to identify proteins associating with the *Neat1* and *Malat-1* long ncRNAs [4]. The protocol described here has been used to map the genome-wide binding locations of *Paupar* and *Dali*, two vertebrate-conserved long ncRNAs that function in the control of neural growth and differentiation, and showed that these transcripts preferentially associate with regions of active chromatin in *trans* across multiple chromosomes [5, 6]. Furthermore, both *Paupar* and *Dali* are co-expressed with their adjacent transcription factor genes during neural differentiation and directly bind the protein product of these genes to target a subset of their genomic binding sites. This shared mechanism of genome targeting and chromatin interaction may be a general feature of a larger family of long ncRNAs whose DNA loci lie in close proximity to transcription factor genes.

In CHART, target long ncRNAs are enriched from a cross-linked chromatin extract using antisense biotinylated oligonucleotides. After immobilization on beads and extensive washing, RNA–chromatin complexes are eluted using RNase H. Genomic DNA associated with these long ncRNAs can be purified and analyzed using real-time PCR in a locus-specific manner or with high-throughput sequencing technologies to map long ncRNA association with chromatin genome-wide. There are a number of key differences between CHART and similar methods, such as ChIRP and RAP, which have also been developed for affinity purification of RNA–chromatin complexes [7, 8]. Firstly, CHART uses a two-step formaldehyde cross-linking approach to fix nuclei. Secondly, an RNase H sensitivity assay is used to identify regions in the target RNA that are accessible for hybridization with antisense oligonucleotides. A small number of short oligonucleotides that have been predetermined to interact with the RNA target are then used in CHART to enrich for RNA–chromatin complexes. Thirdly, antisense oligonucleotide bound RNA–chromatin complexes are eluted using RNase H. This reduces nonspecific false positive binding events generated by direct binding of antisense oligonucleotide probes to DNA [3]. Although, CHART has now been used to map chromatin occupancy for a number of long ncRNAs, the length, stability, and cellular localization of target RNAs will all influence the efficacy of antisense oligonucleotide-based approaches to purify long ncRNA-associated complexes. These factors should be taken into consideration during experimental design.

2 Materials

Wear gloves and take care to avoid RNase and DNase contamination during all stages of this protocol. Before starting, treat all surfaces and pipettes with an RNase decontamination solution such as RNaseZap and then wipe clean with sterile diethylpyrocarbonate (DEPC)-treated water.

2.1 Equipment

1. Nuclease-free 1.7 ml microcentrifuge tubes, 200 μl thin-walled PCR tubes, 15 and 50 ml centrifuge tubes.
2. 1.5 ml Phase Lock Gel Light tubes.
3. Refrigerated microcentrifuge and centrifuge.
4. Glass Dounce homogenizer.
5. Rotator (for 15 and 50 ml tubes).
6. Sonicator (a Bioruptor or similar machine).
7. Magnetic racks to use with 1.7 and 15 ml tubes.
8. Thermomixer (heat block with shaker).
9. Spectrophotometer (NanoDrop).
10. Real-time PCR machine.

2.2 Solutions

Use molecular biology grade reagents and nuclease-free water to make up all buffers and solutions. Solutions should be stored at 4 °C unless stated otherwise.

1. Glycerol Buffer: 25% glycerol, 10 mM HEPES pH 7.5, 1 mM EDTA, 0.1 mM EGTA, and 100 mM KOAc. Just before use, add 0.5 mM spermidine, 0.15 mM spermine, 1× cOmplete EDTA-free protease inhibitors, 1 mM DTT, and 20 U/ml SUPERase-In.
2. Sucrose Buffer: 300 mM sucrose, 1% Triton X-100, 10 mM HEPES pH 7.5, 0.1 mM EGTA, and 100 mM KOAc. Immediately before use, add 0.5 mM spermidine, 0.15 mM spermine, 1× cOmplete EDTA-free protease inhibitors, 1 mM DTT, and 20 U/ml SUPERase-In.
3. Nuclei Rinse Buffer: 50 mM HEPES pH 7.5, 75 mM NaCl, and 0.1 mM EGTA. Add 1× cOmplete EDTA-free protease inhibitors, 1 M DTT and 20 U/ml SUPERase-In immediately before use.
4. 2.5 M glycine solution.
5. Sonication Buffer: 50 mM HEPES pH 7.5, 75 mM NaCl, 0.1 mM EGTA, 0.125% *N*-lauroylsarcosine, and 0.025% sodium deoxycholate. Store at room temperature. Add 1× cOmplete EDTA-free protease inhibitors, 1 M DTT, and 20 U/ml SUPERase-In immediately before use.

6. Quenching Buffer: 250 mM Tris pH 7.2, 125 mM EDTA, 2.5% SDS, and 5 μg/μl proteinase K. Make fresh.
7. Wash Buffer 100 (WB100): 100 mM NaCl, 10 mM HEPES pH 7.5, 2 mM EDTA, 1 mM EGTA, 0.2% SDS, and 0.1% *N*-lauroylsarcosine solution. Store at room temperature.
8. Wash Buffer 250 (WB250): 250 mM NaCl, 10 mM HEPES pH 7.5, 2 mM EDTA, 1 mM EGTA, 0.2% SDS, and 0.1% *N*-lauroylsarcosine solution. Store at room temperature.
9. Denaturant Buffer: 8 M urea, 100 mM HEPES pH 7.5, 200 mM NaCl, and 2% SDS.
10. 2× Hybridization Buffer: 1.5 M NaCl, 1.12 M urea, 10 mM EDTA, and 10× Denhardt's solution.
11. RNase H Elution Buffer (HEB): 50 mM HEPES pH 7.5, 75 mM NaCl, 3 mM $MgCL_2$, 0.125% N-lauroylsarcosine, 0.025% sodium deoxycholate, 10 mM DTT, and 20 U/ml SUPERase-In. Make fresh.
12. Nucleic Acid Buffer (XLR): 250 mM Tris pH 7.5, 2.5% SDS and 5 μg/μl proteinase K. Make fresh.

2.3 Reagents and Consumables

1. RNaseZap.
2. DEPC.
3. Nuclease-free H_2O.
4. Phosphate-Buffered Saline (PBS).
5. 37% Formaldehyde solution.
6. 20 U/μl SUPERase-In.
7. cOmplete EDTA-free protease inhibitor cocktail.
8. 1 U/μl RQ1 RNase-Free DNase.
9. 20 μg/μl Proteinase K solution RNA grade.
10. GlycoBlue Co-precipitant.
11. SYBR Green PCR Master Mix.
12. Streptavidin C1 beads.
13. Commercially synthesized 24 mer oligonucleotides with an 18-atom hexa-ethyleneglycol spacer followed by a biotin-TEG modification at the 3′ end (-HEG-BIOTIN-TEG).
14. Standard unmodified desalted oligonucleotides.
15. 5 U/μl RNase H.
16. Reverse Transcription Kit.
17. miRNeasy purification kit (Qiagen) or Trizol LS (Ambion).
18. Phenol:CHCl3:isoamyl alcohol saturated with 10 mM Tris pH 8.0, 1 mM EDTA.
19. Chloroform.

3 Methods

The protocol described here has been used to map the genomic occupancy of two lowly expressed, chromatin-associated long ncRNAs in mouse N2A neuroblastoma cells [5, 6]. This protocol can be applied to cell lines from different species, including mouse, human, and *Drosophila*, to identify long ncRNA targets. For cell lines that are grown in suspension, pellet the required number of cells and start the protocol at Subheading 3.1, **step 3**.

3.1 Preparation of Nuclei from Cross-Linked Cells

1. Grow approximately 8×10^7 cells for each individual CHART pull down (*see* **Note 1**). We typically use five 15 cm dishes of adherent mammalian cells, of fibroblast size, grown to approximately 80–90% confluency for each antisense oligonucleotide cocktail used.
2. Wash each dish twice with PBS and harvest cells by trypsinization.
3. Resuspend cell pellets in a total of 20 ml ice-cold PBS, pool together into a single 50 ml tube, and pellet at $2000 \times g$, 5 min, 4 °C.
4. Resuspend cells in 20 ml PBS, taking care to obtain a single cell suspension, add formaldehyde to a final concentration of 1%, and rotate at room temperature for 10 min.
5. Add glycine to 0.125 M final concentration, rotate cell suspension for a further 5 min, and pellet at $2000 \times g$, 5 min, 4 °C.
6. Resuspend pellet and wash further three times using 10 ml ice-cold PBS.
7. Resuspend cell pellet in 4 ml ice-cold Sucrose Buffer and transfer to a chilled glass Dounce homogenizer (*see* **Note 2**). Disperse cells with ten strokes of the pestle; incubate for 10 min on ice and then dounce another ten times.
8. Add 4 ml ice-cold Glycerol Buffer to a 15 ml tube, mix another 4 ml ice-cold Glycerol Buffer into the cell suspension in the Dounce homogenizer and then gently pipette this mixture onto the top of the Glycerol Buffer in the 15 ml tube.
9. Pellet at $1000 \times g$, 10 min, 4 °C to isolate the nuclei and remove the supernatant (*see* **Note 3**).
10. Isolated nuclei can be stored at –70 °C or used immediately for RNase H mapping (*see* Subheading 3.2) or for a CHART pull down (*see* Subheading 3.3).

3.2 RNase H Mapping to Identify Regions of the Long ncRNA That Are Accessible for Hybridization

1. Use nuclei purified from approximately 2×10^7 cells to prepare an extract for RNase H mapping experiments. This should be sufficient to test RNase H sensitivity for 50 antisense oligonucleotides (*see* **Note 4**).
2. Resuspend nuclei in 5 ml ice-cold PBS in a 15 ml tube and pellet at $2000 \times g$, 5 min, 4 °C to wash. Repeat wash one more time.

3. Wash pellet once more using 500 µl sonication buffer and transfer to a 1.7 ml microcentrifuge tube.
4. Resuspend pellet in 500 µl sonication buffer, separate into two 250 µl aliquots so as not to exceed the maximum sonication volume allowed per tube, and sonicate for 20 cycles (30 s on, 30 s off) in a Bioruptor Plus (Diagenode) at full power.
5. Pellet debris for 20 min at 16,000 × *g*, 4 °C and remove the supernatant, i.e., the soluble chromatin extract to a fresh tube. Cleared extract can be aliquoted and frozen at −70 °C or used immediately. 500 µl extract is sufficient for 50 × 10 µl RNase H mapping reactions.
6. Prepare a master mix based on the number of oligonucleotides that are being tested for RNase H sensitivity. For each reaction, use 10 µl extract, 0.3 µl $MgCl_2$ (100 mM stock), 0.1 µl DTT (1 M stock), 1 µl RNase H, and 0.5 µl SUPERase-In.
7. Aliquot 11.9 µl master mix into the required number of thin-walled PCR tubes and add 1 µl of each antisense oligonucleotide to be tested into separate tubes (*see* **Note 5**). Include at least two control "no oligonucleotide" reactions using 1 µl H_2O to make reaction volumes equal.
8. Mix by pipetting, pulse tubes in a centrifuge, and then incubate reactions for 1 h at 37 °C.
9. Pulse tubes in a microcentrifuge, add 1 µl RQ1 RNase-Free DNase, and 1 µl $CaCl_2$ (6 mM stock) to each reaction and incubate for 10 min at 30 °C. DNase solution can be prepared as a master mix for the required number of tubes immediately before use.
10. Pulse tubes, add 2 µl Quenching Buffer, pipette up and down to mix, and incubate at 55 °C for 1 h followed by 65 °C for 30 min.
11. Purify RNA using a commercially available kit such as the Qiagen miRNeasy kit, elute in 30 µl nuclease-free H_2O, and quantify RNA using a NanoDrop spectrophotometer.
12. Perform quantitative RT-PCR (RT-qPCR) with 1 µg purified input RNA per RT reaction. Use the formula, $2^{(\text{Ct oligo}-\text{Ct no oligo})}$ for the RNA target/$2^{(\text{Ct oligo}-\text{Ct no oligo})}$ for a control RNA to calculate relative RNase H sensitivity for each antisense oligonucleotide (*see* **Note 6**).
13. Synthesize RNase H selected antisense 24 mer oligonucleotides, containing a -HEG-BIOTIN-TEG modification at the 3′ end, for use in CHART (*see* **Note 7**).

3.3 CHART Pull down

1. Wash nuclei (prepared from five 15 cm dishes) twice with 10 ml ice-cold PBS and transfer to a 50 ml falcon tube.

2. Resuspend nuclei in 20 ml PBS and add formaldehyde to a 3% final concentration. Incubate for 30 min with rotation at room temperature to further cross-link the nuclei.
3. Add 0.125 M final concentration glycine to quench the reaction, incubate for an additional 5 min, and pellet nuclei at 2000 × *g*, 5 min, 4 °C.
4. Wash nuclei twice with 5 ml ice-cold PBS and twice with ice-cold WB100.
5. Resuspend cross-linked nuclei in 1 ml WB100 containing SUPERase-In RNAse inhibitor and EDTA-free protease inhibitor cocktail.
6. Transfer 250 µl aliquots into separate 1.7 ml microcentrifuge tubes and sonicate for 20 cycles (30 s on, 30 s off) in a Bioruptor Plus (Diagenode) at full power to shear the chromatin (*see* **Note 8**).
7. Pellet debris for 20 min at 16,000 × *g*, 4 °C, pool aliquots together, and adjust final volume of extract to 1.5 ml using WB100. Add 30 µl SUPERase-In, 15 µl DTT (1 M stock), and 30 µl protease inhibitor cocktail (50× stock).
8. Transfer 6 × 250 µl sample volumes into separate 1.7 ml microcentrifuge tubes for hybridization and add 125 µl Denaturant Buffer and 375 µl 2× Hybridization Buffer to each 250 µl extract to make a total of 750 µl for each pull down. 1% volume can be removed at this stage for the input sample.
9. Add 6.75 µl of a 25 µM stock solution of CHART oligonucleotide cocktail mix to each 750 µl CHART pull down reaction (*see* **Note 9**) and hybridize overnight at room temperature with rotation.
10. Clear samples by centrifuging for 20 min, 16,000 × *g* at room temperature and remove supernatant.
11. Use a total of 250 µl MyOneC1 streptavidin beads to capture long ncRNA chromatin complexes for each pull down. Rinse beads twice with 500 µl nuclease-free H_2O, resuspend in 167 µl nuclease-free H_2O, and add 83 µl Denaturant Buffer to make a 500 µl bead suspension.
12. Transfer 83 µl bead suspension into six separate microcentrifuge tubes, add 1/6 volume cleared CHART reaction to each tube, and incubate overnight at room temperature with rotation to capture complexes.
13. Pellet beads using a magnet and resuspend in 250 µl WB250. Pool together the six reactions from each CHART pull down into a single 15 ml tube containing 5 ml WB250.
14. Wash beads a total of four times with 5 ml WB250 in a 15 ml tube (*see* **Note 10**).

15. Resuspend beads in 500 µl HEB and transfer to a new 1.7 ml microcentrifuge tube. Repeat this process using another 500 µl HEB so that no beads are left behind and transfer solution to the same tube.
16. Separate beads using a magnet, remove supernatant, and resuspend in 400 µl HEB. Aliquot into four separate 100 µl reactions for RNase H elution.
17. Add 2 µl RNase H to each tube and incubate at 37 °C for 30 min with gentle shaking to elute RNA–chromatin complexes (*see* **Note 11**).
18. Pulse samples in a microcentrifuge and then use a magnet to isolate the beads. Pool samples together into a single 1.7 ml microcentrifuge tube.

3.4 Nucleic Acid Purification to Test for Enrichment of Target Long ncRNA and Associated DNA Targets

1. Add 100 µl XLR Buffer containing Proteinase K to each 400 µl pooled CHART pull down sample (from Subheading 3.3, **step 18**). Also, prepare an input chromatin sample (from Subheading 3.3, **step 8**). Adjust input volume to 400 µl before adding 100 µl XLR Buffer containing Proteinase K. Purify RNA and DNA in parallel as described.
2. Incubate at 55 °C for 1 h followed by 65 °C for 30 min to reverse cross-links (*see* **Note 12**).
   (a) Remove 100 µl for RNA purification to test for target enrichment (*see* **Note 13**).
   (b) Purify DNA associated with each CHART pull down from the remaining 400 µl sample using Phenol:CHCl3:isoamyl extraction and ethanol precipitation.
3. Add an equal volume (400 µl) Phenol:CHCl3:isoamyl alcohol to each sample and mix by shaking vigorously for 15 s.
4. Centrifuge at 12,000 × *g* for 5 min at room temperature to separate the phases. Remove the upper aqueous phase to a new tube taking care not to remove the interphase (*see* **Note 14**).
5. Add equal volume of chloroform to the aqueous phase, mix by shaking, and centrifuge for 3 min at 12,000 × *g*, room temperature. Remove the aqueous layer and repeat this step to ensure that residual phenol is removed.
6. Transfer the upper aqueous layer to a new microcentrifuge tube, add 40 µl (1/10 volume) 3 M NaOAc pH 5.5, 1 µl GlycoBlue co-precipitant and 1 ml (2 1/2 volumes) ethanol. Mix and incubate overnight at −20 °C to precipitate the DNA.
7. Pellet DNA by centrifugation at 16,000 × *g*, 30 min, 4 °C. Remove supernatant and wash pellet with 500 µl 70% ethanol.
8. Centrifuge at 16,000 × *g*, 5 min, 4 °C. Remove ethanol and resuspend the air-dried DNA pellet in 100 µl TE.

9. CHART-enriched genomic DNA can be analyzed using real-time PCR to test for enrichment at specific loci (*see* **Note 15**) or with high-throughput sequencing technologies (*see* **Note 16**) to map long ncRNA associations with chromatin genome-wide.

4 Notes

1. We used approximately 8×10^7 cells per pull down to prepare a CHART extract to enrich for a long ncRNA expressed at an average level of 15 copies per cell. However, it may be necessary to adjust the amount of starting material based on the expression level and subcellular localization of the RNA target to improve the signal versus noise ratio. A single CHART experiment consists of 1–2 specific pull downs and a non-targeting control so it is necessary to grow up 2–3 times this number of cells per experiment depending on the number of pull down reactions. The volume of different solutions detailed in the protocol is for a single CHART pull down using extract prepared from five 15 cm dishes. This will need to be scaled up for each experiment depending on the number of separate pull down reactions.
2. The Dounce homogenizer should be left on ice for more than 15 min before use so that it is suitably chilled. Use a "tight" or "type B" pestle.
3. Take care to remove the upper layer first to minimize contamination from the cytoplasmic fraction.
4. RNase H specifically degrades RNA that is hybridized to DNA. To identify antisense oligonucleotides mapping to accessible regions of target transcripts using RNase H sensitivity, we first design a series of 24 mer oligonucleotides spanning the length of the target transcript. We typically perform two rounds of screening to identify regions in target RNAs that are accessible for hybridization. In initial "low resolution screens," we space oligonucleotides approximately 200–300 nucleotides apart so that we can tile over large transcript distances. We then focus on the most accessible transcript regions for a second "high resolution" screen using oligonucleotides spaced much more closely together (approximately 25 nucleotides apart) to identify oligonucleotides to use in CHART experiments. It should be noted that we have also had success using pools of antisense oligonucleotides designed against evolutionary conserved transcript regions in CHART experiments to successfully enrich target long ncRNAs, completely omitting the RNase H mapping step [5].
5. Standard PCR grade unmodified desalted oligonucleotides are used in RNase H mapping experiments.
6. We routinely use the QuantiTect Reverse Transcription Kit (Qiagen) for this step but other kits are available. We use

qPCR primer pairs to amplify an approximately 100–250 nucleotide region of the target as well as primers to amplify a control unrelated cDNA, such as *Gapdh*, to normalize for input levels. The qPCR primers used to amplify the target should span the antisense oligonucleotide that is being assayed for accessibility and therefore multiple different primer pairs are needed to tile along the length of a RNA target. A standard curve should be performed for all primer pairs to confirm linear and specific amplification.

7. We recommend using two separate pools of 4–5 non-overlapping antisense oligonucleotides to independently enrich a single target long ncRNA and identify binding sites that are common to both CHART pull down experiments in order to reduce false positive binding events [5]. All antisense oligonucleotide probes should be tested using BLAT searches to ensure that they uniquely align to the RNA target (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). If multiple hits against the target transcriptome and/or genome are found an oligonucleotide should not be used. The melting temperature of oligonucleotides should be between 55 and 65 °C.
8. Over-sonication of the samples can lead to degradation of the RNA target. We therefore fragment chromatin to an average size of approximately 1000–1500 bp for CHART as opposed to the 500 bp average fragment size that we routinely use for ChIP.
9. We use a mix of 4–5 separate RNase H mapped antisense oligonucleotides for each CHART pull down. A control CHART reaction using a separate oligonucleotide mix is also prepared in parallel. We have used oligonucleotides against *LacZ* that don't target the mammalian genome as well as sense oligonucleotides as controls. Based on these analyses, we recommend the use of sense oligonucleotides as controls: they work better to eliminate false positive binding events generated as a result of non-specific hybridization of probes directly to genomic DNA, and the GC content of sense probes exactly matches that of the antisense oligonucleotides used in the CHART pull down.
10. We use a DynaMag-15 (Life Technologies) to isolate beads in a 15 ml tube.
11. RNase H elution of RNA–chromatin complexes enriched using antisense DNA oligonucleotides has been identified as a key step to reduce nonspecific associations between antisense oligonucleotides and complementary genomic sequences [3]. Use a fresh aliquot of RNase H for each experiment. Elution efficiency can be analysed by RT-qPCR using RNA prepared from the beads as well as the eluate.
12. We used a short 30 min 65 °C incubation to reverse cross-links. Although longer incubation times improve the efficiency

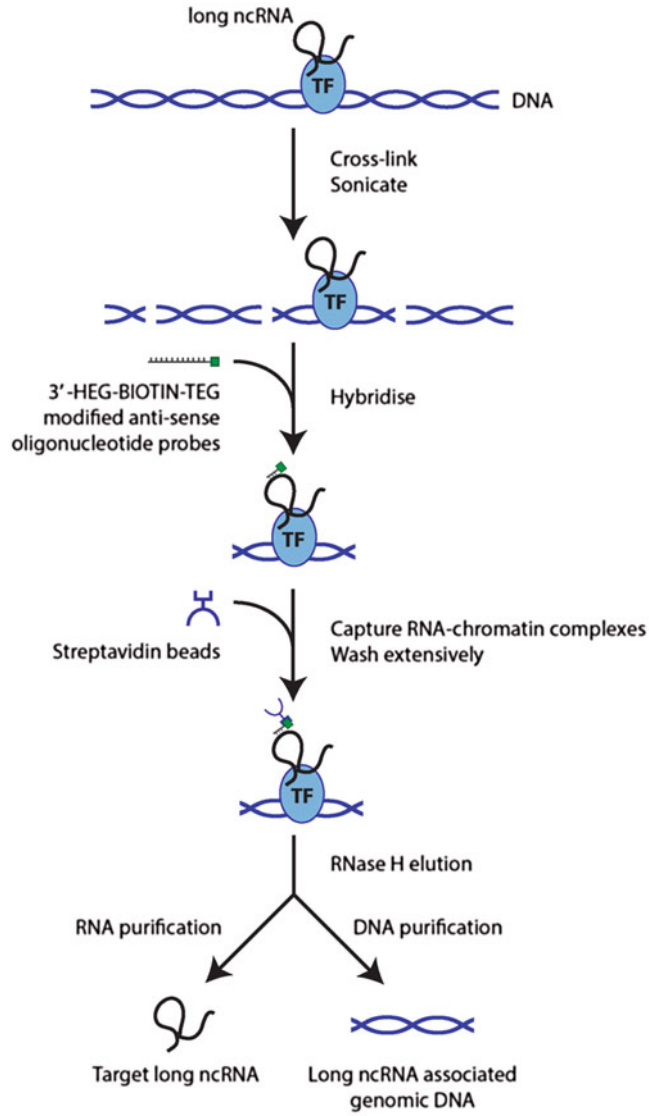


Fig. 1 Overview diagram showing CHART workflow. In this example, the target RNA is shown to indirectly interact with genomic regulatory sequences through association with a sequence-specific DNA-binding transcription factor (TF)

of cross-link reversal, this leads to increased RNA degradation. This incubation may be extended for different RNA targets.

13. We have successfully used the Qiagen miRNeasy kit as well as TRIzoL LS Reagent (Life Technologies) to purify CHART-enriched RNA targets. We elute RNA in 12 μl nuclease-free H_2O and generate cDNA using the QuantiTect Reverse Transcription Kit. We use multiple different primer pairs along the length of the target molecule in a qPCR reaction to assess target RNA recovery. We test for specific RNA enrichment

compared to control pull downs and also assay additional nuclear-enriched chromatin-associated transcripts, such as *Malat-1*, as well as *Gapdh* mRNA to assess specificity of the pull down. If target RNA enrichment in the CHART pull down is poor, a higher concentration of antisense oligonucleotides can be tested in further CHART experiments (Fig. 1).

14. We use Phase Lock Gel tubes to ensure maximum recovery of DNA and to reduce protein contamination from the interface.
15. As the distal-binding locations of target long ncRNAs are most likely unknown, we use real-time PCR to identify enrichment of the endogenous DNA locus of CHART-enriched target long ncRNAs as a positive control to test for associated genomic DNA before high-throughput sequencing.
16. We sequence four different CHART-enriched DNA samples on a single lane of an Illumina HiSeq2500 (50 bp, paired-end sequencing). This generates at least 25 million uniquely mapped reads per sample in a successful CHART experiment. CHART DNA is treated the same as ChIP DNA for library preparation and sequencing using standard protocols.

Acknowledgment

I would like to thank Dr. Mike Clark (Oxford) and Dr Lovorka Stojic (Cambridge) for critically reading the manuscript.

References

1. Vance KW, Ponting CP (2014) Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet 30:348–355
2. Simon MD, Pinter SF, Fang R et al (2013) High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature 504:465–469
3. Simon MD, Wang CI, Kharchenko PV et al (2011) The genomic binding sites of a noncoding RNA. Proc Natl Acad Sci U S A 108:20497–20502
4. West JA, Davis CP, Sunwoo H et al (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol Cell 55:791–802
5. Chalei V, Sansom SN, Kong L et al (2014) The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. Elife 3:e04530
6. Vance KW, Sansom SN, Lee S et al (2014) The long non-coding RNA Paupar regulates the expression of both local and distal genes. EMBO J 33:296–311
7. Chu C, Qu K, Zhong FL et al (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA chromatin interactions. Mol Cell 44:667–678
8. Engreitz JM, Pandya-Jones A, McDonel P et al (2013) The Xist long ncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341:1237973

Chapter 6

Detecting Long-Range Enhancer–Promoter Interactions by Quantitative Chromosome Conformation Capture

Wulan Deng and Gerd A. Blobel

Abstract

Chromosome conformation capture (3C) technology and its derivatives are currently the primary methodologies measuring contacts among genomic elements. In fact, the lion share of what is currently known about chromosome folding is based on 3C-related approaches. For example, distal enhancers are commonly in physically proximity with their target genes, forming chromatin loops. Additional layers of chromatin organization have been described using 3C-based techniques, including topological domains (TADs) and sub-TADs. Finally, inter-chromosomal interactions have been reported although they are much less frequent. 3C is becoming increasingly widespread in its use for understanding genome organization. Here we provide a protocol for quantitative 3C using real-time PCR analysis, along with essential quality controls and normalization methods.

Key words 3C, Chromosome conformation capture, Chromatin looping, Enhancer, Promoter, Restriction enzyme, Ligation, qPCR, TaqMan

1 Introduction

Chromatin is folded in complex but nonrandom patterns. In higher eukaryotic cells, regulatory elements can reside up to hundreds of kilobases away from the genes they control but can be brought into spatial proximity through chromatin loop formation. Moreover, spread out enhancer elements can also interact with each other to form chromatin hubs. Fluorescence in situ hybridization (FISH) assays can be used to measure distances among genomic regions in individual cells but are limited by their spatial resolution and low throughput. In comparison, the chromosome conformation capture (3C) technology [1] can effectively detect a wide range of chromatin interactions, and can be combined with sequencing technology for high throughput studies. Many such high-throughput 3C derivative technologies have been developed in recent years, including Hi-C [2], ChIA-PET [3], 4C-seq [4], and Capture-C [5]. Even though single cell Hi-C [6] promises insights into chromatin folding of

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_6, © Springer Science+Business Media New York 2017

individual cells, almost all reported 3C-based studies provide population averages. In spite of this limitation, 3C and its derivative technologies are still gaining in popularity due to ever improving throughput and resolution. With this increase in use, there has also been a widening in the standards and rigor applied to the execution and interpretation of 3C experiments. Here we provide a basic protocol for quantitative 3C along with controls and normalization standards as a primer for those getting started.

The principle of 3C technology is based on assessing the contact frequency between any chosen pair of genomic DNA segments by measuring their ability to be ligated to each other after cross-linked chromatin has been restriction digested and re-ligated. Thus, 3C involves four major steps: cross-linking, digestion, ligation, and quantification of the ligation products (Fig. 1). Formaldehyde is commonly used to cross-link DNA and protein complexes and thus stabilize chromatin contacts. The cross-linked

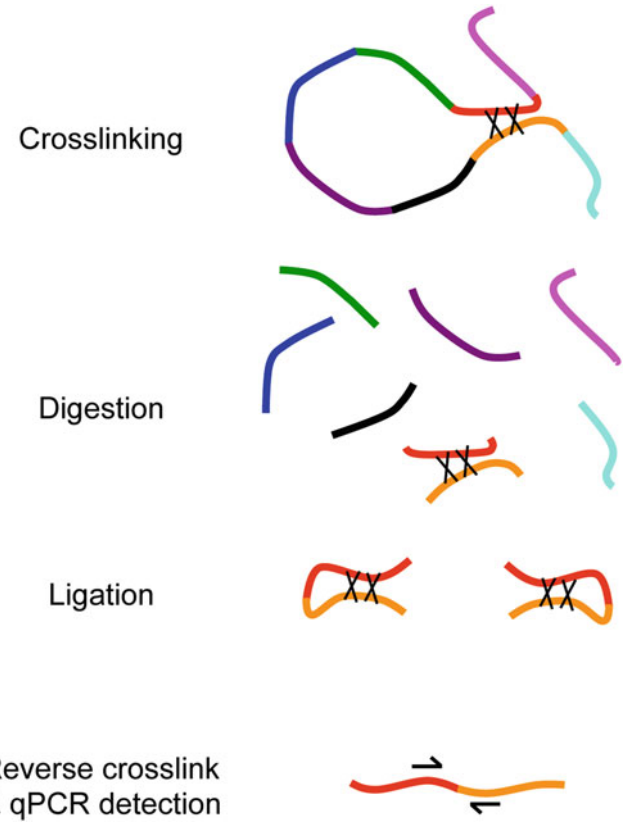


Fig. 1 Schematic illustration of 3C procedures. Genomic fragments digested by restriction enzymes are pictured as lines in different colors. Formaldehyde is used to captures the chromatin interactions via cross-linking protein–protein and protein–DNA interactions, as denoted by "X." Sequence-specific primers are used for qPCR analysis

chromatin is then digested with a restriction enzyme of choice. The digested chromatin is then ligated under diluted condition to facilitate interactions among DNA fragments that are part of the same cross-linked complexes. It is possible to perform the DNA-DNA proximity ligation in intact nuclei as described in in situ Hi-C method [7]. The chance of two genomic fragments being ligated to each other is therefore a function of their spatial proximity. Following ligation, crosslinks are reversed and the DNA is purified. Ligation products of interest are quantified by real-time PCR using sequence specific primers and TaqMan probes. In the end, the abundance of ligation products correlates inversely with the distance between the two fragments. While random ligation events are also inversely correlated with the distance between two DNA segments, specific interactions are defined as "local peaks" above the baseline of random interactions (Fig. 2). In other words, an interaction is defined as contact frequencies among two regions that are higher than intervening fragments or random interactions. There is an ongoing debate as to whether local or genome wide background signals should be used as standards, and how high the enrichment over background needs to be in order to qualify as an "interaction." Regardless, a meaningful 3C analysis requires careful quality control as well as accurate quantification that need to be detailed in any resulting publication.

2 Materials

2.1 Reagents and Solutions

1. 37% Formaldehyde.
2. Protease inhibitor cocktail (Roche).
3. PMSF (Thermo scientific).

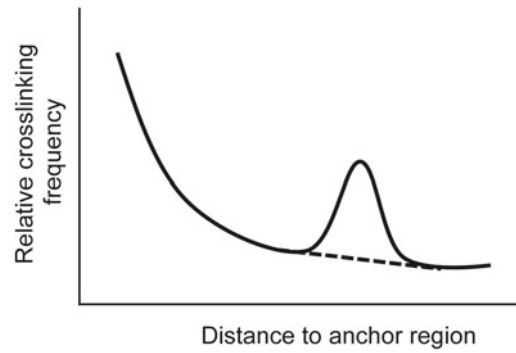


Fig. 2 Virtual 3C data showing relative cross-link efficiency as a function of distance between the anchor fragment and the candidate fragments. The solid line represents the plot for a looped conformation with the presence of a "local peak," whereas the dotted line represents the plot for a linear conformation with decreasing interaction over distance

4. Cell Lysis Buffer: 10 mM Tris pH 8, 10 mM NaCl, 0.2% NP-40 (Igepal CA-630), add protease inhibitor and 1 mM PMSF right before use.
5. Trypan blue solution.
6. SDS (Fisher).
7. Triton X-100 (Sigma).
8. Restriction enzyme and digestion buffer (NEB).
9. 10× Ligation buffer: 500 mM Tris pH 7.5, 100 mM $MgCl_2$, 100 mM DTT.
10. ATP (Sigma).
11. T4 ligase (NEB).
12. PK buffer: 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.5% SDS.
13. Proteinase K (Sigma).
14. Rnase (Dnase free) (Roche).
15. Phenol/chloroform/isoamyl alcohol.
16. Chloroform.
17. 100% Ethanol.
18. Glycogen.
19. 3 M Sodium Acetate.
20. Genomic DNA Extraction Kit (Qiagen).
21. Probes and primer oligos (IDT).
22. Sybr-green master mix (ABI).

2.2 Equipment and Software

1. Optical microscope for assessing cell lysis.
2. Dounce homogenizer.
3. Thermomixer.
4. Water bath.
5. Centrifuge for 15 and 50 ml Falcon tubes.
6. Microcentrifuge.
7. Quantitative Real-time PCR system (for example ABI 7900HT, ViiA 7).
8. Probe and primer design tool (for example Primer Express Software).

3 Methods

3.1 Experimental Design

1. Restriction enzymes. A restriction enzyme should be selected to dissect the locus of interest such that relevant regulatory elements are separated in distinct genomic fragments, and several intervening fragments that can serve as controls. Six base

cutters such as EcoRI, BglII, or HindIII are commonly used for analyzing the topology of a large locus. When analyzing a small locus or short-distance interactions (<10–20 kb), four base cutters such as DpnII or NlaIII can be used to increase the resolution. Variation in digestion efficiencies increases the risk for experimental artifacts. Since the thermostability and digestion efficiency of restriction enzymes vary, the fraction of digested chromatin needs to be measured at every locus under study. This can be done by quantitative PCR with primers against digested and undigested DNA sequences (see below).

2. Design of 3C primers and probes (Fig. 3). A restriction fragment containing a region of interest, such as an enhancer or promoter, is designated as the anchor region or point of view. The 3C analysis quantifies the products generated from the ligation of the anchor fragment with any other candidate interacting fragments. Since the amount of the ligation product of any two interacting DNA fragments is minuscule among the vast and complex mixture of ligation products within the 3C library, highly specific and sensitive TaqMan qPCR is required for quantification. The reaction requires a TaqMan probe and an "anchor primer" located on the opposite strand of the anchor fragment, and a set of "test primers" targeting the candidate interacting fragments. For efficient amplification, all primers should be designed with similar melting temperatures and all amplicons should be small (100–200 bp) and of similar sizes. The annealing temperature of the TaqMan probe should be 8–10 °C higher than that of the primers. The probes contain

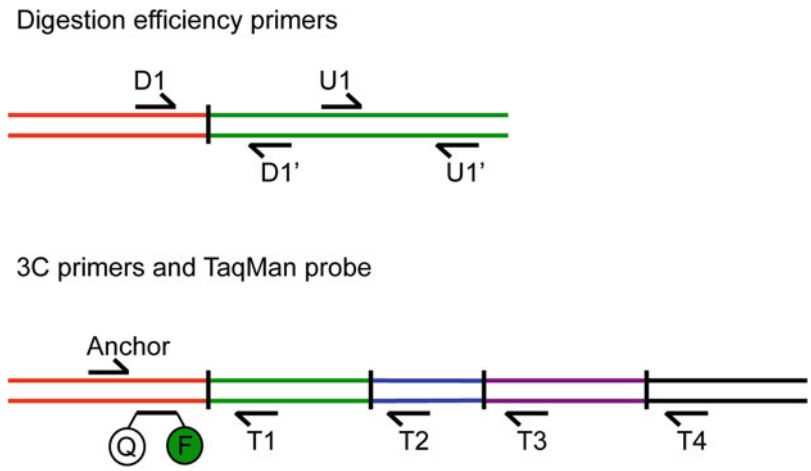


Fig. 3 Schematic diagram of designing primers and probe for analyzing digestion efficiency (*upper*) and association frequency (*bottom*). The D1 and D1′, U1 and U1′ primer pairs detect the DNA sequences that are digested or undigested by the restriction enzyme. The anchor primer and test primers against the candidate fragments (T1, T2, T3, T4) are designed as pictured. The 3C probe is designed to contain a 5′ fluorescein and a 3′ quencher group

a 5′ fluorophore (e.g. FAM), and a 3′ quencher (e.g. TAMRA, BHQ). All primers and probes should be tested with PCR reactions using the standard DNA (see below) as template to validate linear amplification of the templates.

3. Control template for 3C analysis. To account for variability in amplification efficiencies between primer pairs, standard curves need to be generated from a control template that contains the ligation product of interest. For this purpose, bacterial artificial chromosomes (BACs) can be used that contain the all regions of interest of a given locus. Purified BAC DNA is digested with the same enzyme used to generate the 3C library and subsequently re-ligated to generate a mixture of all possible ligation products. An alternative approach is to generate by preparative PCR the ligation product of interest from the 3C library, verify the correctness of the product by sequencing, and then use defined amounts of the ligation product to generate the standard curve. Below we describe the preparation of control template from BAC DNA.

   (a) Culture 2 × 500 ml bacteria with 12.5 μg/ml chloramphenicol. Use Qiagen large-construct Kit to purify BAC DNA.

   (b) Digest ~20 μg of BAC DNA with restriction enzyme overnight and confirm completeness of the digestion by agarose gel electrophoresis.

   (c) Purify the digested BAC DNA by phenol chloroform extraction followed by ethanol precipitation. Dissolve the DNA pellet in 160 μl H_2O.

   (d) Combine 160 μl of digested BAC DNA, 20 μl of 10× ligation buffer, 2 μl of BSA (10 mg/ml), 2 μl of ATP (100 mM), and 7600 cohesive end units of T4 DNA ligase, and adjust to 200 μl final volume. Incubate overnight at 16 °C.

   (e) Inactivate the ligase by incubating the solution for 15 min at 65 °C.

   (f) Purify the ligated BAC DNA by phenol chloroform extraction followed by ethanol precipitation. Dissolve the DNA pellet in TE buffer to obtain a DNA solution with the final concentration of ~100 ng/μl.

   (g) Remove any RNA by adding DNase-free RNase followed by incubation for 15 min at 37 °C. The DNA obtained is termed the control template.

3.2 Cross-Linking and Digestion

1. Collect 10×10^6 suspension cells that are growing at log phase. Spin down cells and resuspend cells into single cell suspension in 20 ml room temperature PBS. Transfer cells to 50 ml conical tube.
2. Cross-link cells by adding 845 μl of 37% formaldehyde (the final concentration is 1.5%) and gently shaking at room temperature for 10 min (*see* **Note 1**).

3. Quench by adding 0.4 g Glycine (the final concentration is about 0.25 M). Gently shake at room temperature for 5 min, then store on ice for 10 min. Adherent cells can be cross-linked and quenched on plates, and scraped off plates.
4. Spin down cells by swing bucket centrifugation, 2000 rpm (805 × *g*), 5 min, 4 °C.
5. Carefully remove the supernatant without disturbing the cell pellet, wash with 10 ml ice-cold PBS. Spin down cells by swing bucket centrifugation, 2000 rpm (805 × *g*), 5 min, 4 °C. Remove supernatant.
6. Resuspend cell pellet in 1.5 ml cold cell lysis buffer containing protease inhibitor (*see* **Note 2**). Store on ice for 20 min.
7. Lyse cells with pre-chilled dounce homogenizer, using ten strokes of Dounce Pestle A (*see* **Note 3**).
8. Transfer cells to 1.7 ml tube. Collect nuclei by swing bucket centrifuge, 2000 rpm (805 × *g*), 5 min, 4 °C.
9. Resuspend the collected nuclei in 800 µl of cold appropriate 1.2× restriction enzyme digestion buffer (RE buffer).
10. Collect nuclei by swinging bucket centrifugation at 2000 rpm (805 × *g*), 5 min, 4 °C.
11. Pipette up and down to resuspend cells in 500 µl of 1.2× RE buffer, avoid air bubbles (*see* **Note 4**).
12. Add 7.5 µl of 20% SDS (the final concentration is 0.3%, *see* **Note 5**). Incubate in a Thermomixer for 1 h at 37 °C, with shaking at 950 rpm.
13. Add 50 µl of 20% Triton X-100 (the final concentration is 1.8%, *see* **Note 6**). Incubate in a Thermomixer for 1 h at 37 °C, with shaking at 950 rpm.
14. Resuspend cells well (*see* **Note 7**).
15. Add 400–1600 U of restriction enzyme (*see* **Note 8**). Incubate in a Thermomixer overnight at 37 °C, with shaking at 950 rpm.

3.3 Ligation and Reverse Cross-Linking

1. Add 40 µl of 20% SDS (the final concentration is 1.6%, *see* **Note 9**). Incubate in a Thermomixer for 25 min at 65 °C, with shaking at 950 rpm.
2. Transfer the sample to a 15 ml centrifuge tube and add 750 µl 10× Ligase buffer, 375 µl 20% Triton X-100 (the final concentration is 1%), 75 µl 10 mg/ml BSA, and 5.7 ml H_2O (*see* **Note 10**). Incubate in a 37 °C water bath for 1 h. The total volume is 7.5 ml.
3. Chill samples on ice. Add 75 µl of 100 mM ATP (the final concentration is 1 mM, *see* **Note 11**).
4. Add 4000 U (10 µl × 400 U/µl) T4 DNA ligase, gently mix, and incubate at 16 °C water bath for 4 h.

5. Take out sample and put at room temperature for 30 min.
6. Add 160 μl of 0.5 M EDTA to stop reaction.
7. To reverse cross-linking, add 50 μl of 20 mg/ml proteinase K, incubate in 65 °C water bath overnight.

3.4 DNA Purification

1. Add additional 25 μl of 20 mg/ml proteinase K and incubate in 55 °C water bath for 2 h (*see* **Note 12**).
2. Transfer the solution to a 50 ml conical tube and cool to room temperature. Add 10 ml pH 8.0 phenol/chloroform/isoamyl alcohol (PCI) and mix vigorously. Then centrifuge at 3500 rpm (2465 × *g*) for 10 min at room temperature. Take out the aqueous phase and extract with PCI again (*see* **Note 13**). Take out the aqueous phase.
3. Add 8 ml chloroform and mix vigorously. Then centrifuge at 3500 rpm (2465 × *g*) for 10 min at room temperature.
4. Transfer the supernatant to a new 50 ml tube and ~3 ml H_2O to bring final volume to 10 ml. Add 1 ml 3 M sodium acetate (pH 5.2), invert to mix, add 25 ml of 100 % ethanol, and invert to mix. Place at −80 °C for 20 min.
5. Pellet DNA by spinning in a swinging bucket rotor at 3500 rpm (2465 × *g*) for 30 min, at 4 °C (*see* **Note 14**).
6. Wash with 20 ml 70 % ethanol, spinning at 3500 rpm (2465 × *g*) for 20 min, at 4 °C.
7. Air-dry pellet for 5 min. Resuspend the pellet in 400 μl 10 mM Tris pH 8 buffer. Add 2 μl RNase, incubate in a 37 °C water bath for 30 min (*see* **Note 15**).
8. Extract with 400 μl PCI. Vortex for 30 s. Spin in a bench top centrifuge at 16000 × *g* for 5 min. Recover the aqueous phase, repeat the PCI extraction.
9. Extract with 400 μl chloroform. Vortex for 30 s. Spin in a bench top centrifuge at 16000 × *g* for 5 min.
10. Recover aqueous phase, add 40 μl of 3 M sodium Acetate, followed with 1.1 ml 100 % ethanol, invert to mix. Put on ice for 30 min.
11. Spin at 16000 × *g* for 30 min at 4 °C.
12. Wash with 1 ml cold 70 % ethanol for three times (*see* **Note 16**).
13. Air-dry the pellet. Resuspend the pellet in 200 μl 10 mM Tris pH 8 buffer. To help DNA to dissolve, incubate in 55 °C water bath for 10 min, and leave at 4 °C overnight. This 3C library is ready for analysis.

3.5 Analysis of 3C Library

3.5.1 Assessment of Digestion Efficiency

Since the digestion efficiency can vary between experiments, it is necessary to measure it in each experiment.

1. Take 10 µl of sample after the digestion step, add 190 µl PK buffer, 3 µl 20 µg/µl proteinase K, and 2 µl RNase. Incubate in 65 °C overnight.
2. Add 200 µl TE buffer and 10 µl (1 µg/µl) Glycogen. Cool to room temperature.
3. Add 400 µl PCI. Vortex for 30 s. Centrifuge at 16000 × *g* for 5 min.
4. Recover the supernatant and add 400 µl Chloroform. Vortex for 30 s. Centrifuge at 16000 × *g* for 5 min.
5. Recover the supernatant and add 40 µl 3 M Sodium Acetate, briefly mix. Add 1 ml cold 100% ethanol. Invert several times to mix. Put in –20 °C for at least 30 min.
6. Centrifuge at 16000 × *g* for 30 min at 4 °C. A small pellet should be visible.
7. Remove ethanol carefully, watch not lose DNA pellet.
8. Add cold 1 ml 70% ethanol, invert several times. Centrifuge at 16000 × *g* for 10 min at 4 °C.
9. Remove supernatant. Air-dry pellet for about 10 min.
10. Add 60 µl water to dissolve DNA.
11. Prepare genomic DNA of the cells used for 3C studies, following instructions of genomic DNA extraction kit.
12. Serial dilutions of genomic DNA (0.1–10 ng/µl) are used as reference DNA for qPCR quantification.
13. Set up 10 µl qPCR reactions using the primer pairs against digested or undigested regions, and using diluted digested DNA sample or the genomic DNA as template.

2× Sybr-green master mix	5 µl
Forward primer (10 µM)	1 µl
Reverse primer (10 µM)	1 µl
DNA template	2 µl
H_2O	1 µl

14. Carry out qPCR and calculate the quantity of digested DNA template for each primer pairs relative to undigested genomic DNA.
15. Calculate digestion efficiency using following function,

$$\%\ \text{of digestion} = \left[1 \quad Q(D)/Q(U)\right] \times 100$$

$Q(D)$ and $Q(U)$ are the relative quantities of DNA template for primer pairs that amplify DNA sequences either containing digested (D) or undigested (U) cutting sites. The optimal digestion efficiency of a 3C library is >70%. Samples with poor digestion efficiency should be discarded.

3.5.2 TaqMan qPCR Analysis of 3C Library

1. Quantify the concentration of 3C library using Sybr-green qPCR with primer pair against undigested regions. The qPCR reaction is the same as that in **step 13** of Subheading 3.5.1.
2. Set up 10 μl reactions as below. 50–200 ng of 3C DNA is typically used for each reaction. The anchor primer is paired with desired test primers to quantify ligation products of the anchor fragment. It is recommended to test at least two different concentrations of DNA template (*see* **Note 17**). Serial dilutions of prepared BAC control template (*see* Subheading 3.1) are used as reference DNA for qPCR quantification.

2× Taqman master mix	5 μl
Anchor primer (10 μM)	1 μl
Test primer (10 μM)	1 μl
Probe (2.5 μM)	1 μl
DNA	2 μl

3. Carry out qPCR. Analyze the data and calculate the relative amounts of ligation products.
4. To allow comparison between experiments, results need to be normalized to a control interaction. Typical controls are housekeeping loci such as ERCC3 and GAPDH. These control interactions are used when comparing different conditions but only if it has been verified that they are invariable.

4 Notes

1. The optimal cross-linking condition (formaldehyde concentration, reaction temperature, and reaction time) should be empirically determined for different cell types. For a typical experiment with mammalian cells, use 1–2% of formaldehyde and cross-link cells at room temperature for 5–15 min. The nature of the long-range interaction under investigation should also be taken into consideration for the optimal cross-linking condition. The detection of stable interactions may require less cross-linking than weak or transient interactions. More intense cross-linking might reduce digestion efficiency while weak cross-linking might fail to capture an interaction. Balancing these criteria requires varying cross-linking conditions.

2. Cross-linked cells tend to stick the sidewall of conical tubes. Wash the sidewall with lysis buffer to collect any remaining cells. Cell lysis buffer contains 0.2% NP-40, which lyses the cytoplasmic membrane but not nuclear membrane.
3. Generating homogenous nuclei at the lysis step is essential for the next digestion step. The choice of lysis protocol depends on the cell type used. For experiments using a new cell type, take a few microliter of homogenized cells and stain them with trypan blue. A successful lysis should give homogenously intact and blue nuclei under microscope. Otherwise, repeat **steps 6** and 7.
4. Single nuclei suspension is very important for efficient digestion. Aggregation of nuclei profoundly interferes with digestion efficiency.
5. This step is to solubilize the nuclear membrane.
6. The presence of SDS from **step 12** inhibits restriction digestion. TritonX-100 is a non-denaturing detergent that doesn't interfere with enzyme activity and is used to sequester the SDS. TritonX-100 is light sensitive, so it is recommended to keep the stock solution in the dark and prepare a fresh working solution.
7. Single nuclei suspension is important for efficient digestion. Aggregation of nuclei lowers digestion efficiency. Therefore it is important resuspend cells well before digestion. If aggregates are formed during digestion, it is recommended to resuspend cells again.
8. The amount of restriction enzyme should be determined empirically by assessing the digestion efficiency. If the digestion efficiency is low, the amount of enzyme and reaction time can be increased. Enzyme can be added multiple times for enzymes with a short life.
9. SDS is used to deactivate Bgl II, which is heat resistant.
10. Triton is to sequester the SDS from the last step. Acetylated BSA confers stability to T4 DNA ligase.
11. The ATP stock is adjusted to pH 7.5 and stored in –20 °C. Care should be taken to avoid ATP degradation by keeping ATP on ice once thawed. Degradation of ATP can lead to inefficient ligation.
12. Proteinase K retains high activity from pH 6.5–9.5, temperature 20–60 °C and in the presence of up to 0.5% SDS.
13. DNA purity is important for accurate quantification by PCR. If the aqueous phase is still very turbid, repeat PCI extraction.
14. Pellet after this step could be white and big, presumably due to DTT precipitation. Pellet size will shrink in the subsequent steps.
15. RNA digestion is carried out after the first round of DNA precipitation because it minimizes reaction volume making the reaction more efficient.

16. These washing steps are necessary to remove extra salt from the DNA pellet. Pellet size should reduce visibly with ethanol washes.
17. The 3C library has impurities that can inhibit the qPCR reaction. Therefore, using less material in the reaction (which means higher dilution) might reduce inhibition and actually increase PCR product. Of course, given the low abundance of a given ligation product, excessive dilution of the library in the reaction leads to loss of signal. Therefore, it is recommended to test a few dilutions each time to determine the optimal concentration.

References

1. Dekker J (2002) Capturing chromosome conformation. Science 295:1306–1311
2. Lieberman-Aiden E, van Berkum NL, Williams L et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293
3. Li G, Fullwood MJ, Xu H et al (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol 11:R22
4. van de Werken HJG, Landan G, Holwerda SJB et al (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Meth 9:969–972
5. Hughes JR, Roberts N, McGowan S et al (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet 46:205–212
6. Nagano T, Lubling Y, Stevens TJ et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502:59–64
7. Rao SSP, Huntley MH, Durand NC et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159:1665–1680

Chapter 7

Deciphering Noncoding RNA and Chromatin Interactions: Multiplex Chromatin Interaction Analysis by Paired-End Tag Sequencing (mChIA-PET)

Jocelyn Choy and Melissa J. Fullwood

Abstract

Genomic DNA is dynamically associated with protein factors and folded to form chromatin fibers. The 3-dimensional (3D) configuration of the chromatin will enable the distal genetic elements to come into close proximity, allowing transcriptional regulation. Noncoding RNA can mediate the 3D structure of chromatin. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) is a valuable and powerful technique in molecular biology which allows the study of unbiased, genome-wide de novo chromatin interactions with paired-end tags. Here, we describe the standard version of ChIA-PET and a Multiplex ChIA-PET version.

Key words Multiplex ChIA-PET, Chromatin immunoprecipitation (ChIP), Chromatin interactions, Proximity ligation, Protein, Antibodies, Barcoded half-linkers, High-throughput sequencing, Paired-end tags

1 Introduction

Chromatin interactions are two or more genomic regions in close spatial proximity and are observed between enhancer elements separated from their target genes by hundreds or thousands of base pairs, thereby leading to gene regulation [1]. Besides mRNA, chromatin interactions play a role in controlling the expression of long noncoding RNA (long ncRNA) through mechanisms involving super enhancers [2]. At the same time, long ncRNA can control chromatin interactions. Activating RNAs can associate with mediator to enhance chromatin interactions [3]. As an example of specific long ncRNAs, *CTCF*, and *CCAT1-L*, long ncRNA may participate in a positive regulatory network in control of *C-MYC* transcription by regulating the higher chromatin organization of 8q24 surrounding the *C-MYC* locus, contributing in part to the aberrant expression of *C-MYC* in human colorectal cancer pathogenesis [4] (Fig. 1).

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_7, © Springer Science+Business Media New York 2017

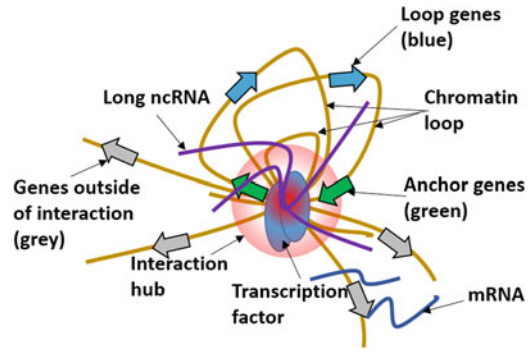


Fig. 1 Chromatin interactions and RNA

Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) was introduced in 2009 and involves the conversion of functional chromatin structure into millions of short tag sequences [5]. It combines Chromatin Immunoprecipitation (ChIP), proximity ligation, and high-throughput sequencing, thus allowing us to look at higher-order chromatin structures associated with specific protein factors [6]. Proximity ligation refers to the method of greatly diluting cross-linked complexes in solution, followed by adding ligase, which leads to intra-complex ligations instead of inter-complex ligations. From the mapping results of the reads to the genomic sequence, the genomic distance between the two mapped tags will reveal whether a PET is derived from a self-ligation product of a single DNA fragment (short genomic distance) or an inter-ligation product of two DNA fragments (long genomic distance or inter-chromosomal) [7]. In this way, ChIA-PET can reveal interactions between enhancer elements and their associated genes. RNA Polymerase II-associated ChIA-PET analysis has enabled the study of the effects of 3D chromatin interactions on the transcription regulation of mRNAs, revealing that genes are regulated in large multigene complexes [8], as well as the effects of 3D chromatin interactions on miRNAs by RNA Polymerase II [9]. Hence, ChIA-PET is a very powerful technique which enables us to understand signaling networks and cell states.

Singleplex ChIA-PET was introduced to allow the study of chromatin interactions involving a single transcription factor of interest (Fig. 5a). Singleplex ChIA-PET involves the use of ChIP against a protein of interest to create a library which harbors information about interactions between the genomic DNA regions that are bound to the protein of interest. This method allows the study of only the chromatin interactions associated with one transcription factor per ChIA-PET library constructed. Hence, our group decided to optimize the ChIA-PET protocol to allow more transcription factors and cell types to be analyzed in a single ChIA-PET library by the introduction of 6 half-linkers during ChIA-PET library construction (Figs. 2 and 5b). The ChIA-PET library is

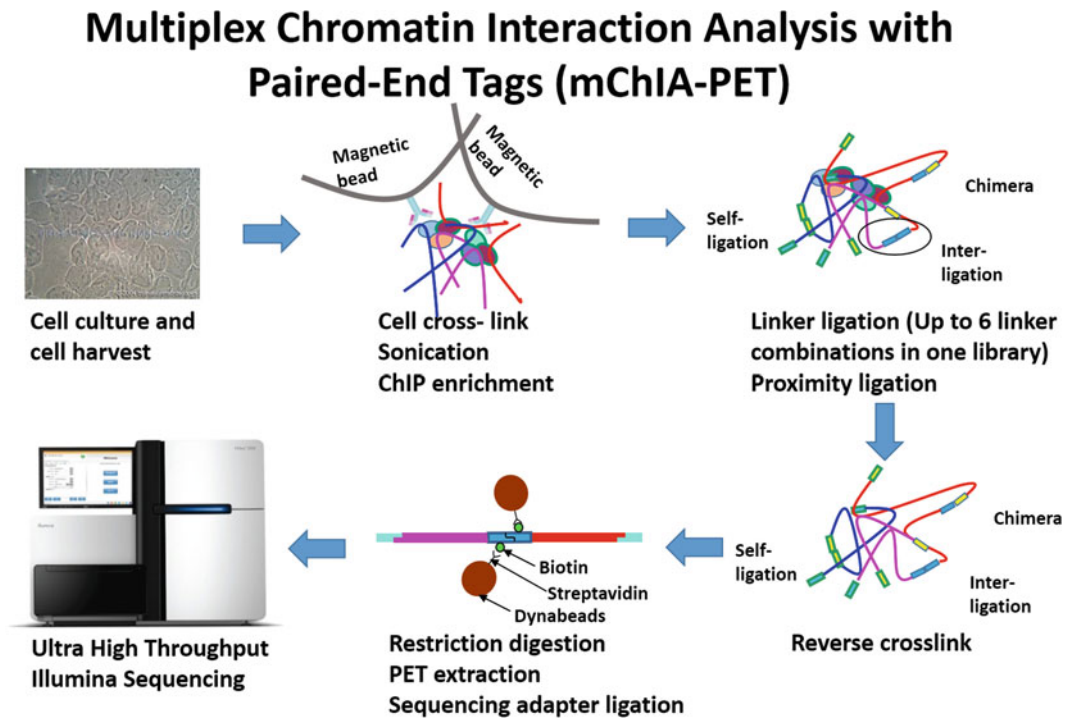


Fig. 2 Multiplex ChIA-PET procedure outline

multiplexed at the half-linker ligation step, whereby the different barcodes within each half-linker distinguished one ChIP from another. The 6 linkers allow the chromatin interaction analysis of up to 6 different ChIPs in one library. This contributes to the flexibility of Multiplex ChIA-PET as there is a choice to analyze one or more ChIPs in the same amount of time. Multiplex ChIA-PET allows more chromatin interaction information to be obtained from single library purification, making it more robust and less time consuming. Multiplex ChIA-PET is also cheaper in terms of labor costs and reagents costs (Table 1).

2 Materials

2.1 Cell Culture

1. Cell culture media: RPMI, 10% Fetal Bovine Serum, 100 μg/ml penicillin-streptomycin. Add 57 ml of Fetal Bovine Serum and 5.7 ml of Penicillin-streptomycin to one bottle of 500 ml RPMI to get a final concentration of 10% and 100 μg/ml, respectively. Store at 4 °C.
2. Phosphate-buffered saline: 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4. Dissolve 8 g of NaCl, 0.2 g of KCl, 1.44 g of Na_2HPO_4, and 0.24 g of KH_2PO_4 in 800 ml of distilled water. Adjust the pH to 7.4 with HCl.

Table 1

Comparison between singleplex ChIA-PET and multiplex ChIA-PET

Singleplex ChIA-PET	Multiplex ChIA-PET
More time consuming	Less time consuming
Need a higher starting amount of ChIP material for each protein factor of interest	Need a lower starting amount of ChIP material for each protein factor of interest
Lesser chimera combinations	More chimera combinations
More labor and resources required to construct libraries.	Less labor and resources required to construct libraries
Less cost-effective	More cost-effective
Short read tags give rise to multi-mapping	
Sequencing library is not Illumina-compatible. Need customized sequencing primers	
Unbiased, whole genome and de novo approach for long range chromatin interaction analysis	

Adjust the volume to 1 l with the addition of distilled water. Sterilize the buffer by autoclaving.

3. Cells from American Type Culture Collection (ATCC).
4. T25 tissue culture flasks.
5. T175 tissue culture flasks.
6. Haemocytometer.
7. 37 °C water bath.
8. 37 °C incubator with 5% CO_2.

2.2 ChIP

Buffers (*see* **Note 1**).

1. Beads wash buffer: 1× PBS, 0.1% Triton X-100. Add 0.2 ml of Triton X-100 to 199.8 ml 1× PBS. Store at 4 °C.
2. High-salt wash buffer: 50 mM HEPES pH 7.5, 350 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Sodium Deoxycholate, 0.1% SDS. Mix together 10 ml of 1 M HEPES pH 7.5, 14 ml of 5 M NaCl, 400 µl of 0.5 M EDTA (pH 8.0), 2 ml of 100% Triton X-100, 2 ml of 10% sodium deoxycholate, 2 ml of 10% SDS, and 169.6 ml of water. Store at room temperature.
3. Lithium chloride wash buffer: 10 mM Tris pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Igepal-CA-630, 0.5% Sodium Deoxycholate. Mix together 2 ml of 1 M Tris pH 8.0, 2.12 g of LiCl (42.39 g/mol), 400 µl of 0.5 M EDTA pH 8.0, 10 ml of 10% IgePal CA-630, 10 ml of 10% sodium deoxycholate, and 177.6 ml of water. Store at 4 °C.

4. TE buffer: 10 mM Tris (pH 8.0), 1 mM EDTA pH 8.0. Mix together 2 ml of 1 M Tris (pH 8.0), 400 μl of 0.5 M EDTA pH 8.0, and 197.6 ml of water. Store at room temperature.
5. ChIP elution buffer: 50 mM Tris pH 8.0, 10 mM EDTA, 1% SDS. Mix together 10 ml of 1 M Tris pH 8.0, 4 ml of 0.5 M EDTA pH 8.0, 20 ml of 10% SDS, and 166 ml of water. Store at room temperature.
6. 0.1% SDS lysis buffer: 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Sodium Deoxycholate, 0.1% SDS. Mix together 25 ml of 1 M HEPES-KOH (pH 7.5), 15 ml of 5 M NaCl, 1 ml of 0.5 M EDTA, 5 ml of Triton X-100, 5 ml of 10% sodium deoxycholate, 5 ml of 10% SDS, and 444 ml of water. Sodium deoxycholate is light sensitive, hence protect it from light. Store at room temperature.
7. 1× TAE buffer: 40 mM Tris, 20 mM acetic acid, 1 mM EDTA. Dissolve 48.5 g of Tris in about 800 ml of water. Add 11.4 ml of acetic acid and 20 ml of 0.5 M EDTA (pH 8.0). Top up to 1 l with water to get a 10× stock solution. Dilute stock solution 1:10 to make a 1× working solution. Store at room temperature.
8. 37% Formaldehyde.
9. 2 M glycine: Dissolve 75.07 g Glycine (75.07 MW) in 0.5 l of water. Sterilize by filtration. Store at room temperature.
10. 50 ml Falcon tubes.
11. 1.5 and 2 ml microcentrifuge tubes.
12. EDTA-free Protease Inhibitor tablet.
13. Intellimixer.
14. Bioruptor Plus Sonicator (Diagenode).
15. Centrifuge for 15 and 50 ml tubes.
16. Centrifuge for 1.5 and 2 ml tubes.
17. Dry incubator.
18. Heat block.
19. Power supply for gel electrophoresis.
20. Gel electrophoresis tank.
21. ChIP- grade antibodies.
22. Dynabeads Protein G for Immunoprecipitation (Life Technologies).
23. Magnetic Particle Concentrator.
24. RNase A.
25. Proteinase K.
26. Agarose powder.

27. PCR purification kit.
28. qPCR Master Mix.
29. Picogreen kit for DNA quantification.

2.3 Multiplex ChIA-PET

Buffers (*see* **Note 1**)

1. TE buffer: 10 mM Tris (pH 8.0), 1 mM EDTA pH 8.0. Mix together 2 ml of 1 M Tris (pH 8.0), 400 µl of 0.5 M EDTA pH 8.0, and 197.6 ml of water. Store at room temperature.
2. Wash buffer: 10 mM Tris–Cl pH 7.5, 1 mM EDTA, 500 mM NaCl. Mix together 5 ml of 1 M Tris–Cl, 1 ml 0.5 M EDTA, 50 ml 5 M NaCl, and 444 ml of water. Store at room temperature.
3. Elution buffer: 1× TE buffer, 1% (w/v) SDS. Mix together 100 µl of 10% SDS with 900 µl of 1× TE. Elution buffer is prepared fresh.
4. 2× B&W buffer: 10 mM Tris–HCl pH 7.5, 1 mM EDTA, 2 M NaCl. Mix together 5 ml of 1 M Tris–HCl, 1 ml of 0.5 M EDTA, 200 ml of 5 M NaCl, and 294 ml of water. Store at room temperature.
5. 1× B&W buffer: 5 mM Tris–HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl. Mix together 2.5 ml of 1 M Tris–HCl, 0.5 ml of 0.5 M EDTA, 100 ml of 5 M NaCl, and 397 ml of water. Store at room temperature.
6. 1× TBE buffer: 100 mM Tris base, 100 mM boric acid, and 2 mM EDTA. Dissolve 121.1 g of Tris base, 61.8 g of boric acid and 7.4 g of EDTA in 1 l of water to make a 10× stock. Dilute stock solution 1:10 to make a 1× working solution. Store at room temperature.
7. Qiagen buffer EB.
8. 10× buffer for T4 DNA polymerase.
9. 10 mM dNTPs.
10. T4 DNA Polymerase.
11. Biotinylated half-linkers A to F and non-biotinylated half-linker H (*see* Subheading 5).
12. 5× T4 DNA ligase buffer with PEG.
13. T4 DNA ligase.
14. 10× T4 DNA ligase buffer.
15. T4 DNA polynucleotide kinase.
16. SpinX columns.
17. 20% Triton X-100.
18. 10× T4 DNA ligase buffer.
19. Proteinase K.
20. MaXtract high-density tubes.

21. Nalgene FEP tubes.
22. Phenol/chloroform (pH 7.9).
23. 3 M sodium acetate pH 5.5.
24. Glycoblue.
25. Molecular biology grade isopropanol.
26. Molecular biology grade ethanol.
27. 10× SAM.
28. 10× NEBuffer 4.
29. MmeI.
30. M280 Streptavidin Dynabeads.
31. MmeI Adapter A and B (*see* Subheading 5).
32. 10× T4 DNA ligase buffer.
33. 10× NEBuffer 2.
34. *E. coli* DNA Polymerase I.
35. Phusion PCR Master Mix.
36. ChIA-PET PCR primers (*see* Subheading 5).
37. 4–20% TBE PAGE gel.
38. 6× gel loading dye.
39. 25 bp DNA ladder.
40. 100 bp DNA ladder.
41. SYBR Gold nucleic acid gel stain.
42. Scalpels.
43. 6% TBE PAGE gel.
44. 21G needle.
45. 0.2, 0.6, 1.5, and 2 ml microcentrifuge tubes.
46. Magnetic Particle Concentrator.
47. Agilent Bioanalyser DNA 1000 kit.
48. SYBR FAST qPCR Master Mix.
49. Gel electrophoresis chamber.
50. Intellimixer.
51. Centrifuge for 1.5 and 2 ml tubes.
52. Centrifuge for 15 and 50 ml tubes.
53. Ultracentrifuge.
54. Dry incubator.
55. Heat block.
56. Thermocycler.
57. Blue light transillumator.
58. Nuclease-free water.

3 Methods

3.1 Cell Culture

1. Use a haemocytometer to estimate the number of cells in each flask (*see* **Note 2**).
2. Resuspend the suspension cells by pipetting up and down using a 10 ml pipette. Transfer the recommended number of cells into the T175 flask and top up to a final volume of 25 ml. Repeat subculturing for the required number of culture flasks for the expansion of cells (*see* **Note 3**). K562 cells can be grown in flasks incubated in a flat position with vented caps.
3. Allow the cells to grow in a 37 °C incubator with 5% CO_2. Occasionally monitor the cells.

3.2 ChIP

3.2.1 Single Cross-linking of Chromatin-Bound Proteins and Cell Harvesting (Suspension Cell)

1. Usually, we grow approximately 2×10^7 cells in one T175 flask (estimate cell number with a haemocytometer).
2. Spin down cells, pool 2T175 flasks together, and resuspend cells in 30 ml of 1× PBS (*see* **Note 4**).
3. To approximately 30 ml of 1× PBS, add 833 µl of 37% formaldehyde to cross-link the cells (final concentration of formaldehyde in the media must be 1%). Incubate at room temperature for 10 min with rotation on the intellimixer (F1, 30) in the fume hood (*see* **Note 5**).
4. Add 2 ml of 2 M glycine and incubate it at room temperature for 5 min on the intellimixer (F1, 30) in the fume hood to quench the formaldehyde.
5. Spin down the cells at 1800 × *g* at 4 °C for 10 min and discard quenched cross-linkers into a waste bottle in the fume hood. Wash cells twice with ice-cold PBS, each time spinning down at 1800 × *g* at 4 °C for 10 min. Discard supernatant and proceed to cell lysis. Alternatively, store the pellet at –80 °C.

3.2.2 Cell Lysis

1. Always add Proteinase Inhibitor (PI) fresh to the 0.1% lysis buffer. Add 1 PI tablet into 10 ml of 0.1% SDS lysis buffer (for Mini EDTA-free tablets) or 1 PI tablet into 50 ml of 0.1% SDS lysis buffer (for Ultra EDTA-free tablets) (*see* **Note 6**).
2. For 10×10^7 cells, use 15 ml 0.1% SDS buffer.
3. Resuspend pellet in the respective volumes of 0.1% SDS lysis buffer + PI and mix thoroughly by pipetting. After adding 0.1% SDS lysis buffer + PI, shake and tap the tube to dislodge the pellet. Then incubate the solution at 4 °C for 1 h with shaking on the intellimixer (F1, 30) in the incubator.
4. Pellet lysed cells at 800 × *g* at 4 °C for 40 min.
5. Discard supernatant and repeat cell lysis once.
6. After the second wash, remove the supernatant. Then resuspend the pellet in 500 µl of 0.1% SDS lysis buffer + PI for every

4×10^7 cells. Transfer 500 µl of chromatin pellet from the 50 ml Falcon tube to the 1.5 ml tube (*see* **Note 7**).

3.2.3 Fragmentation of Chromatin

1. Shear chromatin–DNA to a size of 200–500 bp with Bioruptor Plus. (Sonication conditions: time: 10 min = 10 cycles, (30 s on, 30 s off), speed: high) (*see* **Note 8**).
2. Reverse cross-link an aliquot of chromatin and check fragmentation efficiency with the following steps.
   - Aliquot 6 µl of chromatin after sonication.
   - Centrifuge 6 µl of the sonicated chromatin at 16.1 k × *g* for 10 min at 4 °C.
   - Transfer supernatant to a new tube and add 2 µl of Proteinase K solution.
   - Incubate for 30 min at 50 °C in heat block or overnight at 37 °C in heat block.
   - Resolve reverse cross-linked chromatin on a 1% agarose gel.
3. Repeat sonication if the sizes of DNA are larger than expected (for around 2–5 min). If the sonicated size is ok, then proceed to preclear the chromatin (*see* **Note 9**).
4. Centrifuge remaining lysate at 16.1 k × *g*, 4 °C for 30 min or longer and transfer sonicated chromatin (supernatant) into a new tube. This step is to remove the cell debris. Combine the tubes from the same replicate together into one tube. If after spinning down and transferring the supernatant into a fresh tube, the supernatant still remains quite milky, then spin the tubes again at 4 °C for another 30 min to 1 h. Then transfer the supernatant from the second spin to another fresh tube. Then start the preclearing with Dynabeads Protein G beads. Alternatively, store sonicated chromatin at −80 °C until sufficient chromatin is collected to start a ChIP.

3.2.4 Washing and Preclearing of Chromatin

1. Assume 500 µl of sonicated chromatin (1 IP) contains 4×10^7 cells (500 µl may have more cells or fewer cells—it can accommodate $1–10 \times 10^7$ cells, so adjust the volumes accordingly if need be).
2. Aliquot out the respective amounts of chromatin to get 2×10^7 cells.
3. 2×10^7 cells are to be mixed with one kind of antibody.
4. Use 60 µl of Dynabeads Protein G beads for one ChIP of 2×10^7 cells.
5. Wash beads three times with 1 ml of beads wash buffer (*see* **Note 10**).
6. After the final wash, remove all the supernatant using the Magnetic Particle Concentrator (MPC). Then add the respective

amounts of sonicated chromatin to each tube containing pre-washed beads only. This step is to remove the background binding of chromatin to beads.

7. Keep 10% of sonicated chromatin as "input" for subsequent enrichment check by quantitative PCR (qPCR). Store the "input" at −80 °C.
8. Rotate overnight at 4 °C on the intellimixer (F1, 30) in the incubator.

3.2.5 Coating of Antibody onto Magnetic Beads

1. Aliquot 60 µl of magnetic Protein G beads per IP into a fresh tube.
2. Wash beads thrice with 1 ml of beads wash buffer (*see* **Note 10**).
3. After the final wash, add 30 µl of beads wash buffer to the beads.
4. For antibodies used (e.g., RNA Polymerase II), we use 14 µg of antibodies for 2×10^7 cells per IP (*see* **Note 11**).
5. Rotate overnight at 4 °C on the intellimixer (F1, 30) in the incubator.

3.2.6 Chromatin Immunoprecipitation

1. Wash antibody-coated beads 3× with 1 ml of beads wash buffer. After the last wash, leave the antibody-coated beads in 1 ml of beads wash buffer.
2. Briefly centrifuge the pre-cleared chromatin on the small benchtop centrifuge.
3. Place tube on the MPC.
4. Discard wash buffer from antibody-coated beads with the help of the MPC.
5. Transfer sonicated chromatin (supernatant) with the help of the MPC to antibody-coated beads.
6. Rotate overnight at 4 °C on the intellimixer (F1, 30) in the incubator.

3.2.7 Washing and Elution of Immunoprecipitated DNA–Protein Complexes

1. Wash chromatin-immunoprecipitated beads thrice with 1 ml of 0.1% SDS lysis buffer. Allow the tubes to rotate at 4 °C for 5 min on the intellimixer (F1, 30) for each wash.
2. Before removing the supernatant, briefly spin the tubes. Wash beads once with 1 ml of high-salt wash buffer. Allow the tubes to rotate at 4 °C for 5 min on the intellimixer (F1, 30) for each wash.
3. Before removing the supernatant, briefly spin the tubes. Wash beads once with 1 ml of Lithium chloride wash buffer. Allow the tubes to rotate at 4 °C for 5 min on the intellimixer (F1, 30) for each wash.

4. Before removing the supernatant, briefly spin the tubes. Discard wash buffer and resuspend washed beads with 1 ml of TE buffer. Allow the tubes to rotate at 4 °C for 5 min on the intellimixer (F1, 30) for each wash.
5. ChIP-enriched beads may be stored for up to 2 weeks at 4 °C.
6. Elute 20% of the ChIP-enriched beads (4×10^6 cells) with 100 µl ChIP elution buffer.
7. Elute the total input by topping up the volume to 100 µl with ChIP elution buffer.
8. Begin the reverse cross-linking procedure for both "input" and eluted ChIP complexes.
9. Add in 2 µl of RNase A (0.5 mg/ml). Incubate for 2 h in a 55 °C heat block (*see* **Note 12**).
10. Then, reverse cross-link "input" and eluted ChIP complexes with 2 µl Proteinase K (stock concentration at 20 mg/ml) (final concentration of 0.4 mg/ml) for 4 h at 55 °C or overnight at 37 °C (*see* **Note 13**).
11. Transfer the supernatant into a new 1.5 ml tube. Then purify the DNA using QIAQuick PCR purification kit according to the manufacturer's protocol. Resuspend the DNA in 20 µl of EB.
12. Quantitate "input" DNA and ChIP DNA by picogreen assay according to manufacturer's protocol (*see* **Note 15**).
13. Perform an enrichment check via quantitative PCR (qPCR) according to manufacturer's protocol (*see* **Note 15**).
14. Proceed to perform ChIA-PET if the ChIP enrichment is good.

3.3 Multiplex ChIA-PET

3.3.1 End Blunting of ChIP DNA Fragments

1. Pre-chill the wash buffer and TE buffer for about 15 min on ice before starting.
2. Take out the required volume of beads to get 200 ng of ChIP material. In this experimental setup, six different ChIPs are multiplexed into one library. These six ChIPs are two replicates of K562 H3K27ac ChIP, two replicates of K562 H3K27me3 ChIP, and two replicates of GM12878 H3K27ac ChIP. Hence, a total of 1.2 µg of ChIP material is used in this library. For a singleplex ChIA-PET library, 1 µg of a single ChIP is used to construct a ChIA-PET library.
3. Spin the tubes briefly at 0.1 k×*g*, 4 °C and put on the Magnetic Particle Concentrator (MPC) to separate the beads from the TE buffer. Discard the TE buffer (carefully without disturbing the beads). Wash the beads once with 700 µl of ice-cold TE buffer (*see* **Note 10**).
4. Prepare end blunting enzyme mix on ice (700 µl/sample) (*see* **Note 14**)

   End Blunting mix (+ T4 DNA polymerase)

Components	X1 (μl)	X1.2 (Singleplex ChIA-PET library) (μl)	X6.2 (Multiplex ChIA-PET library) (μl)
Nuclease-free water	615.8	738.96	3817.96
10× Buffer for T4 DNA pol	70	84	434
10 mM dNTPs	7	8.4	43.4

5. Put the tubes on the MPC to remove TE buffer (carefully without disturbing the beads) and add 692.8 μl of the end blunting enzyme mix to the beads by inverting the tube and flicking.
6. Add 7.2 μl of T4 DNA polymerase.
7. Mix and incubate at 37 °C for 40 min with rotation on the intellimixer (F8, 30 rpm; $U=50$, $u=60$).
8. After 40 min, take the tubes out from the 37 °C incubator. Briefly spin the tubes at 0.1 k×*g*, 4 °C and leave the tubes on the MPC. Discard the supernatant (carefully without disturbing the beads). Wash the beads three times with 700 μl of ice-cold wash buffer (*see* **Note 10**).

3.3.2 Ligation of Half-Linkers to Polished Ends (Refer to Subheading 5 for Sequence Information)

1. Make sure T4 DNA ligase IS NOT added to the ligation mix straight away. This is important! It is to prevent the linkers from ligating together.

   Linker ligation mix (without T4 DNA ligase) (Final volume: 200 μl) (*see* **Note 14**). The six different linkers used corresponded to the 6 different ChIPs constructed for K562 and GM12878 cells.

Components	X1 (μl)	X1.3 (μl)
Nuclease-free water	155	201.5
Linker A/B/C/D/E/F (Biotin, 200 ng/μl)	3	3.9
5× T4 DNA ligase buffer with PEG	40	52

2. Add 198 μl of master mix to the respective linker ligation mix (i.e., Linker A master mix to one tube and Linker B master mix to the other tube). For a multiplex ChIA-PET library, all linkers A to F are used. For a singleplex ChIA-PET library, only linkers A and B are used.
3. Add 2 μl of T4 DNA ligase (30 U/μl) to one tube and invert that tube to mix straightaway. Place that tubes on the intellimixer (F8, 30 rpm, $U=50$, $u=60$) at room temperature to mix and proceed to adding T4 DNA ligase to the next tube and so on.

The tubes are then incubated at 16 °C, overnight (at least 16 h) with rotation on the intellimixer (F8, 30 rpm, $U=50$, $u=60$).

3.3.3 Addition of Phosphate Group to 5′-Ends of the Linkered DNA Fragments

1. Place the tubes on the MPC for ~1 min and discard the ligation mix. Wash the beads thrice with 700 μl of ice-cold wash buffer (*see* **Note 10**). Repeat the washing procedures for two more times but for the third wash, keep the beads in the wash buffer on ice. Combine the beads in all six tubes together (for multiplex ChIA-PET) or combine the beads in two tubes together (for singleplex ChIA-PET).
2. Prepare enzyme mix on ice (700 μl/sample) (*see* **Note 14**)

Components	X1	X1.5
Nuclease-free water	616 μl	924 μl
10× T4 DNA ligase buffer	70 μl	105 μl
T4 DNA polynucleotide kinase	14 μl	(Added last)

3. Place the tube on the MPC for 1 min to remove wash buffer and resuspend 686 μl of the enzyme mix to the beads by inverting. 14 μl of T4 DNA polynucleotide kinase is added last. Incubate at 37 °C for 55 min with rotation on the intellimixer (F8, 30 rpm, $U=50$, $u=60$).

3.3.4 Elution of Chromatin–DNA Complex

1. Prepare elution buffer which is 1 % SDS (100 μl 10%SDS + 900 μl buffer TE) fresh.
2. Place the tube on the MPC for about 1 min. Discard the T4 DNA polynucleotide kinase enzyme mix.
3. Add 200 μl of elution buffer (buffer TE + 1 % SDS) to the beads. Flick gently to prevent the formation of bubbles.
4. Incubate at room temperature for 30 min with rotation on the intellimixer (F8, 30 rpm, $U=50$, $u=60$).
5. Place the tube on the MPC for about 1 min. Transfer the 200 μl elution buffer-containing chromatin–DNA complex to a fresh tube.
6. Wash the remaining beads with 900 μl of buffer EB by inverting the tube. Briefly spin at a benchtop centrifuge.
7. Place the tube on the MPC for about 1 min. Transfer the buffer EB to the same tube. Pass the eluate (total of 1100 μl), 600 μl through a SpinX column each time and spin the column at 16.1 k×*g*, 4 °C for 1 min (*see* **Note 16**).

8. Transfer the filtrate into a 1.5 ml tube and add 90 µl of 20% Triton X-100. Incubate at 37 °C for 1 h without rotation on the heat block. Total volume of DNA = (200 + 900 + 90) µl = 1190 µl.

3.3.5 Circularization of Linkered DNA Fragments

1. Prepare ligation mix (10 ml/sample) in a 50 ml Falcon tube (*see* **Note 14**)

Components	X1
Nuclease-free water	7776 µl
DNA	1190 µl
10× T4 DNA ligase buffer	1000 µl
T4 DNA ligase (30 U/µl)	33.4 µl (added last)

2. Incubate at 16 °C overnight (~20–24 h) without rotation.

3.3.6 Decross-linking of Chromatin–DNA Complex (Removal of Protein)

1. Add 100 µl of 20 mg/ml Proteinase K (Thermo Scientific). Mix by flicking, followed by a short spin.
2. Incubate at 37 °C overnight (~16 h) without rotation.

3.3.7 DNA Purification

1. Before using the MaXtract High Density, pellet the tube by centrifugation at 0.9 k×*g*, 4 °C for 5 min. Top up the volume of the ligation mix to 19 ml by adding 9 ml of nuclease-free water. (This is to make sure that the Nalgene FEP tube will be ≥80% full). Add 19 ml phenol/chloroform (pH 7.9) to each Maxtract High Density and mix by inverting vigorously for about 2 min. Centrifuge at 0.9 k×*g*, 4 °C for 5 min to separate the phases. Transfer the upper aqueous phase into a 50-ml transparent Nalgene FEP tube.
2. Precipitate the DNA by adding the following components to each tube.

Components	Volume
DNA solution	19 ml
3 M Sodium acetate pH 5.2	1.9 ml
GlycoBlue	5 µl
Molecular biology grade isopropanol	19 ml

3. Incubate at −80 °C for at least 45 min. Pre-chill the high speed centrifuge to 4 °C (pre-chill takes 30 min). After the sample is frozen, thaw the sample slightly before centrifugation. Weigh each tube to balance appropriately for centrifugation. Spin the DNA at 55 k×*g*, 4 °C for 30 min.

4. Decant the supernatant (carefully, pellet may be loose) and pool the blue pellets together in a 1.5 ml tube. Wash pellet twice with 1 ml of 75 % ethanol**.** Try to remove all the ethanol and air-dry the pellet. Resuspend the pellet in 34 μl of buffer EB (Qiagen).

3.3.8 MmeI Digestion to Release the Captured iPETs

1. Prepare MmeI enzyme master mix (*see* **Note 14**).

Components	X1 (μl)	X1.5 (μl)
10× NEBuffer 4	5	7.5
10× SAM (1 μl SAM + 63 μl dH_2O)	5	7.5
H linker (non-Biotin) to quench excess enzyme (200 ng/μl)	5	7.5

2. Add 15 μl of master mix to 34 μl of DNA. Mix by pipetting up and down.
3. Add 1 μl MmeI (NEB), mix by pipetting up and down and incubate at 37 °C for ≥2 h without rotation.

3.3.9 Preparation of Dynabeads and Immobilization of iPET DNA

1. Mix the M280 Streptavidin Dynabeads (Invitrogen) suspension well before transferring 50 μl Dynabeads suspension for each sample to a 1.5 ml tube. Wash the beads two times with 150 μl of ice-cold 2× B&W buffer and remove the buffer using the MPC (*see* **Note 10**). Resuspend the beads in 50 μl of 2× B&W buffer.
2. Transfer all 50 μl MmeI digested mix to the 50 μl washed dynabeads suspension (total volume is 100 μl). Incubate at room temperature for 45 min with rotation on the intellimixer (F8, 30 rpm, $U = 50$, $u = 60$). Wash the beads three times with 150 μl of ice-cold 1× B&W buffer and remove the buffer using the MPC (*see* **Note 10**). Leave the beads in 1× B&W buffer after the last wash on the MPC.

3.3.10 Ligation of Adaptors A and B to the Immobilized iPet-DNA

1. Make sure T4 DNA ligase IS NOT added to the ligation mix straightaway. This is important! It is to prevent the adaptors from ligating together.
2. Prepare ligation mix (50 μl/sample) (*see* **Note 14**).

Components	X1 (μl)	X1.5 (μl)
Nuclease-free water	36	54
MmeI Adaptor A (200 ng/μl)	4	6
MmeI Adaptor B (200 ng/μl)	4	6
10× T4 DNA ligase buffer	5	7.5

3. Remove the 1× B&W buffer using the MPC and resuspend the ligation mix to the beads.
4. Add 49 µl of master mix to beads and mix by pipetting up and down.
5. Add 1 µl T4 DNA ligase (30 U/µl) (Fermentas) and mix by pipetting up and down.
6. Incubate at 16 °C overnight (~16 h) with rotation on the intellimixer (F8, 30 rpm, $U=50$, $u=60$).

3.3.11 Nick Translation of Paired-End-Tag (PET) Constructs on Dynabeads

1. Wash the beads three times with 150 µl of 1× B&W buffer and remove the buffer using MPC (*see* **Note 10**). Leave the beads in 1× B&W buffer on ice after the last wash on the MPC.
2. Prepare enzyme mix (50 µl/sample) (*see* **Note 14**).

Components	X1 (µl)	X1.5 (µl)
Nuclease-free water	38.5	57.75
10× NEBuffer 2	5	7.5
10 mM dNTPs	2.5	3.75

3. Remove 1× B&W buffer using MPC and add 46 µl of master mix to beads and mix by pipetting.
4. Add 4 µl of *E.coli* DNA Polymerase I (NEB) and mix by pipetting. Incubate at room temperature for 2 h with rotation on the intellimixer (F8, 30 rpm, $U=50$, $u=60$).

3.3.12 QC PCR Amplification for Viewing the iPETs

1. Wash the beads three times with 150 µl of 1× B&W buffer (*see* **Note 10**).
2. Resuspend the beads in 50 µl of buffer EB (Qiagen).
3. Set up the following QC PCR reaction. Put 2 µl of beads suspension in a 0.2PCR tube. Add 48 µl of master mix to the beads by pipetting (*see* **Note 14**).

Components	X1 (µl)	X2.5 (master mix) (µl)
Beads suspension	2	2
2× Phusion master mix HF	25	62.5
Solexa 1-454 (25 µM)	1	2.5
Solexa 2-454 (25 µM)	1	2.5
Nuclease-free water	21	52.5

4. Set up the following PCR program. Pause the PCR program when it reaches 98 °C.

Step	Temperature (°C)	Time
1	98	30 s
Repeat steps 2–4 for a total of 18 or 20 cycles		
2	98	10 s
3	65	30 s
4	72	30 s
5	72	5 min
6	4	Forever

3.3.13 QC Gel Loading

1. Use a 10-well, 4–20 % pre-casted gradient PAGE gel.
2. Load 25 μl PCR product premixed with 5 μl of 6× loading dye, 250 ng of 25 bp ladder.
3. Run gel electrophoresis at 180 V for 55 min.
4. Post-stain the PAGE gel by adding 80 ml of 1× TBE with 8 μl of SYBRGold nucleic acid gel stain and shake it for 10 min at room temperature. View gel under the blue light transilluminator (*see* **Note 18**).

3.3.14 Scale-Up PCR Amplification for Viewing the iPETs

1. Set up the following QC PCR reaction. Put 2 μl of beads suspension in a 0.2PCR tube. Add 48 μl of master mix to the beads by pipetting (*see* **Note 14**).

Components	X1 (μl)	X24 (master mix) (μl)
Beads suspension	2	2
2× Phusion master mix HF	25	600
Solexa 1-454 (25 μM)	1	24
Solexa 2-454 (25 μM)	1	24
Nuclease-free water	21	504

2. Set up the following PCR program. Pause the PCR program when it reaches 98 °C.

Step	Temperature (°C)	Time
1	98	30 s
Repeat steps 2–4 for a total of n cycles (determined during QC PCR)		
2	98	10 s
3	65	30 s
4	72	30 s
5	72	5 min
6	4	Forever

3.3.15 Scale-Up Gel Loading

1. Pool the PCR reactions together.
2. Place the tubes on the MPC and transfer the supernatant into another tube.
3. Add 0.1 volume of sodium acetate pH 5.2, 1 volume of isopropanol and 2 μl of glycoblue.
4. Freeze and precipitate the DNA solution at −80 °C for at least 45 min.
5. Immediately spin at 16.1 k×*g*, 4 °C for 30 min.
6. Remove the supernatant and wash the pellet twice with 1 ml of ice-cold 75 % ethanol.
7. Air-dry the pellet for 5–10 min.
8. Resuspend pellet in 50 μl 1× TE and add 10 μl of 6× loading dye.
9. Use 6 % TBE pre-casted PAGE gel (non-urea) for scale-up.
10. Load 500 ng of 25 bp ladder. Load 30 μl of sample into each well.
11. Run electrophoresis at 180 V for 40 min.
12. Post-stain the PAGE gel by adding 80 ml of 1× TBE with 8 μl of SYBRGold nucleic acid gel stain and shake it for 10 min at room temperature. View gel under the blue light transilluminator (*see* **Note 18**).
13. Cut out gel and proceed to gel extraction (*see* **Note 18**).

3.3.16 Gel Extraction Protocol

1. Excise the 223 bp DNA into the 0.6 ml micro tubes that have been pierced at the bottom with a 21G needle (1 gel slice: 1 0.6 ml tube). The pieced tube is placed inside a 1.5 ml screw-cap micro tube and centrifuged at 16.1 k×*g*, 4 °C for 10 min. The gel slices are thus conveniently shredded and collected in the bottom of each 1.5 ml tube.
2. Add 200 μl of 1× TE buffer to each 1.5 ml screw-cap micro tube. Stir the gel pieces with the pipette tip. Make sure the gel pieces are immersed with the buffer.
3. Freeze the 1.5 ml screw-cap micro tubes containing the shredded gel at −80 °C for 1 h, and then transfer directly to 37 °C incubation. The shredded gel is thus macerated at 37 °C overnight.
4. Brief spin at room temperature. Transfer the gel pieces together with the buffer in each 1.5 ml tube to the filter cup of a SpinX column. Rinse the 1.5 ml tubes which have been used to macerate the shredded gel with 200 μl of 1× TE. Transfer this 200 μl of 1× TE to the SpinX column as well. Hence the total volume in the SpinX column is 400 μl. Centrifuge the SpinX column at 16.1 k×*g*, 4 °C for 10 min (*see* **Note 17**).

5. Transfer 400 μl of the spinned down solution into a 1.5 ml tube. Then perform isopropanol precipitation as follows.

Components	Volume (μl)
DNA solution	430
3 M Sodium acetate, pH 5.2	43
Glycoblue	2
Isopropanol	430

6. Incubate the tube at –80 °C for 1 h. Centrifuge at 16.1 k×*g*, 4 °C for 30 min. Wash the DNA pellet with 1000 μl 75% ethanol twice. Resuspend ChIA-PET DNA library in 12 μl of nuclease-free water after air-drying the DNA pellet for 5–10 min. Proceed to check the concentration of DNA using Agilant DNA 1000 and KAPA qPCR according to the manufacturer's protocol (Fig. 5) (*see* **Note 18**).
7. Sequencing was performed on Illumina MiSeq, using a sequence read length of 2x76bp. The library was prepared according to manufacturer's protocol (Part #15039740 Rev. D). However, customized sequencing primers (Solexa 3-454 and Solexa 4-454 sequencing primers) were used instead of the Illumina sequencing primers. Details of the customized sequencing primers could be found in Subheading 5. A final library concentration of 7.2 pM was loaded into the MiSeq cartridge to obtain a 60% loading density due to the low complexity in the half-linkers region. MiSeq v2 optimal loading is 950 k/mm^2 while MiSeq v3 optimal loading is 1200 k/mm^2. Hence a 60% loading density will give a loading of about 570 k/mm^2 for MiSeq v2 and a loading of about 720 k/mm^2 for MiSeq v3.
8. Sequencing was subsequently performed on Illumina HiSeq, using a sequence read length of 2x76bp. The library was prepared according to manufacturer's protocol (Part #15050107 Rev. C). Similarly, customized sequencing primers (Solexa 3-454 and Solexa 4-454 sequencing primers) were used instead of the Illumina sequencing primers. Details of the customized sequencing primers could be found in Subheading 5. MiSeq loading is 80% of HiSeq loading, hence the amount of library loaded onto HiSeq could be calculated based on the loading amount and cluster density generated from the MiSeq run. For example if a MiSeq loading of 7.2 pM gave a good 60% loading density of 720 k/mm^2 on MiSeq v3, then, the HiSeq loading should be 9 pM to give a good 60% loading density of 500 k/mm^2 on the HiSeq Rapid Run Sequencer.

4 Notes

1. During the preparation of buffers, stir the components on a magnetic stirrer at room temperature to get a homogenous mixture.
2. When using a haemocytometer to count the cells in each flask, load about 15 μl of cell solution into one chamber (top chamber). Count all the cells in the middle chamber consisting of 25 big squares; in each big square there are another 16 small squares. If the cells fall on the outer most triple gridlines, only count those cells that fall on the top and right triple gridlines; don't count the cells on the bottom and left triple gridlines. Counting the cells in all the 25 big squares will give you the number of cells in 1 mm^2. After counting, wipe the chambers using kimwipes and 75% ethanol. Wipe the exterior of the counting chamber using c-fold towel. Load cells into the bottom chamber also and take the average values of cells.
   - Total number of cells in one culture flask: Number of cells in 1 $mm^2 \times$ Volume of media in culture flask (in μl) $\times 10$.
3. Recommended subculturing cell numbers:
   - K562: Subculture at 1×10^6 cells/ml (Medium renewal: every 2–3 days).
   - GM12878: Use a minimum of 3×106 cells for seeding in each passage.
4. If cells form clumps, the cell strainer is used to disperse the clumps so that the cells will have uniform contact with formaldehyde during the subsequent cross-linking step.
5. Formaldehyde incubation is preferably to be performed for 10 min at room temperature to avoid over cross-link or under cross-link but should be optimized according to the antibody and protein of interest. Formaldehyde should be fresh. Discard formaldehyde bottles 3 months after opening.
6. Dissolve the Proteinase Inhibitor (PI) tablets by rotating the tubes on the intellimixer (F1, 30) in 4 °C.
7. When preparing chromatin for sonication, ensure that there are no bubbles in the tube of chromatin as bubbles will affect the efficiency of sonication. Remove bubbles by using a P200 to burst or suck out the bubbles. Note—sonication also breaks up the nuclear membrane. In several other protocols, a nuclear lysis step is performed. However, we've found this protocol works well without the nuclear lysis step for certain cell lines.
8. Sonication points to note: You need to ensure that there is water in the water bath. The water in the water bath is used to

maintain the sonicating conditions at 4 °C. Hence, you need to switch "ON" the sonicator and the water bath for at least 30 min before use to allow the temperature to drop to 4 °C before you can start sonicating your chromatin. Sonication is done using high frequency sound waves to break the chromatin. No probe is required. When putting the 1.5 ml tubes into the adapter, you need to ensure that the tubes are balanced, because the adapter will rotate in the water during sonication to allow the heat generated during sonication to spread evenly in the water. The 4 °C water bath and the rotation of the adapter will help to keep the sonication conditions at 4 °C.

9. The ideal sonicated chromatin size ranges from 200 to 500 bp. Resonicate for another 2–5 cycles, each cycle consisting of 30 s on, 30 s off, high speed if majority of the sonicated chromatin falls above 500 bp. Discard the chromatin and sonicate a fresh tube of chromatin using 5–8 cycles, each cycle consisting of 30 s on, 30 s off, high speed if the majority of the sonicated chromatin falls below 300 bp. An example of an ideal sonication result is shown below in Fig. 3.
10. Washing of magnetic beads is done by tapping the tube to resuspend the beads. Brief spin the tubes at 4 °C to collect the magnetic beads at the bottom of the tube. Do not centrifuge the beads at high speed. Place the tubes on the Magnetic Particle Concentrator (MPC) and remove the supernatant. Repeat the washing steps or proceed to the next step as stated in Subheading 3. Ensure that the beads do not become dry.
11. Protein G binds to the antibodies' (Ab) constant region. Different animal antibodies bind differently to Protein G. Hence, you need to check the compatibility of the antibody to Protein G before coating the Protein G magnetic beads to the antibody. For every ChIP experiment that we perform, we

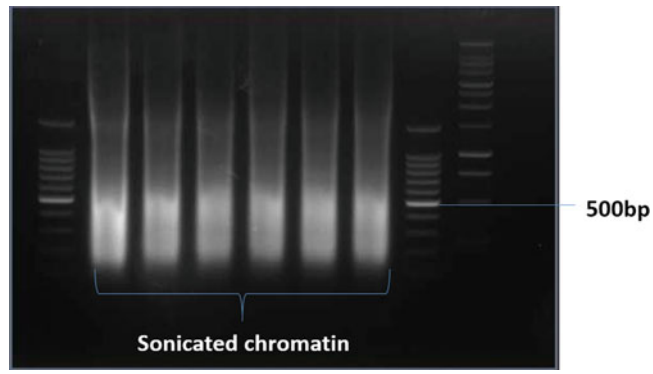


Fig. 3 Ideal gel electrophoresis profile of sonicated chromatin. The majority of the sonicated chromatin should have a size of between 200 and 500 bp

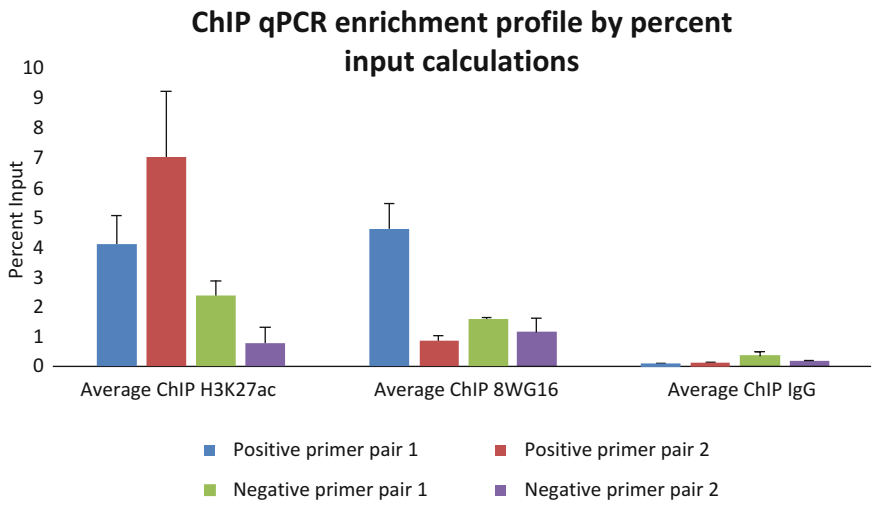


Fig. 4 Ideal ChIP qPCR enrichment profile. The enrichment observed in the positive qPCR primers should be higher than the negative control or "background" qPCR primers. The enrichment in the factors of interest (H3K27ac and RNA Polymerase II) should be higher than the negative control factor (IgG)

will include a positive control using RNA Polymerase II ChIP and a negative control using IgG ChIP.

12. The purpose of incubating with RNase A is to remove all the RNA that might be present to ensure that only the ChIP DNA is analyzed.
13. If incubating Proteinase K overnight, incubate at 37 °C to reduce the formation of nicks. A nick is a discontinuity in a double stranded DNA molecule where there is no phosphodiester bond between adjacent nucleotides of one strand typically through damage or enzyme action.
14. Enzymatic master mix should be prepared on ice. Ligase is unstable, even on ice and should be promptly placed back in −20 °C.
15. An ideal ChIP should have a good qPCR enrichment and the picogreen quantification for combined ChIP enrichment should be approximately 1 μg for constructing one multiplex ChIA-PET library. The percent input calculation is shown below and the profile of an ideal ChIP qPCR enrichment is illustrated below in Fig. 4. We find that the percent input calculation is more accurate in reflecting ChIP enrichments as opposed to other methods such as fold enrichment calculation.

Calculations for Percent Input

Step 1: Adjust input to 100%

Input adjustment to 100% = Ct of input − 6.644 (Note: For example, if the starting input fraction is 1%, then a dilution factor (DF) of 100 or 6.644 cycles (i.e., log 2 of 100) is subtracted from the Ct value of diluted input.)

Step 2: Percent input calculation

Take the triplicate average of the Cts of IP

Percent input of IP = 100 × 2^(Adjusted input − Average Ct (IP))

16. SpinX column will remove any residual beads from the eluate to ensure that there will not be any carryover of beads to the next step.
17. SpinX column will remove the gel slices from the eluate.
18. Ideal gel electrophoresis profiles of ChIA-PET QC and scale-up should give a library band size of 223 bp with minimal smear (Fig. 5c). When doing gel excision, cut as close to the band as possible, without including the ends of the band (Fig. 5d). The ideal Agilent Bioanalyser DNA 1000 profile should have a peak at the 223 bp position, without the presence of any other peaks (Fig. 5e)

5 Appendix

1. Half-linkers for ChIA-PET.
   Annealed linker A.
   3′ CAACCTATTCTA/iBiodT/AGCGCCGG 5′.
   5′ GTTGGATAAGAT A TCGC 3′.
   Annealed linker B.
   3′ CAACCTTACATA/iBiodT/AGCGCCGG 5′.
   5′ GTTGGAATGTAT A TCGC 3′.
   Annealed linker C.
   3′ CAACCTTCAATA/iBiodT/AGCGCCGG 5′.
   5′ GTTGGAAGTTAT A TCGC 3′.
   Annealed linker D.
   3′ CAACCTACTTTA/iBiodT/AGCGCCGG 5′.
   5′ GTTGGATGAAAT A TCGC 3′.
   Annealed linker E.
   3′ CAACCTTAACTA/iBiodT/AGCGCCGG 5′.
   5′ GTTGGAATTGAT A TCGC 3′.
   Annealed linker F.
   3′ CAACCTCTATTA/iBiodT/AGCGCCGG 5′.
   5′ GTTGGAGATAAT A TCGC 3′.
   Annealed non-biotinylated linker H.
   3′ CAACCTAGGCTATAGCGCCGG 5′.
   5′ GTTGGATCCGATATCGC 3′.
2. Adapters for ChIA-PET.

   MmeI Adapter A.
   5′ CCATCTCATCCCTGCGTGTCCCATCTGTTCCCTC CCTGTCTCAGNN 3′.
   3′ GGTAGAGTAGGGACGCACAGGGTAGACAAGGGA GGGACAG

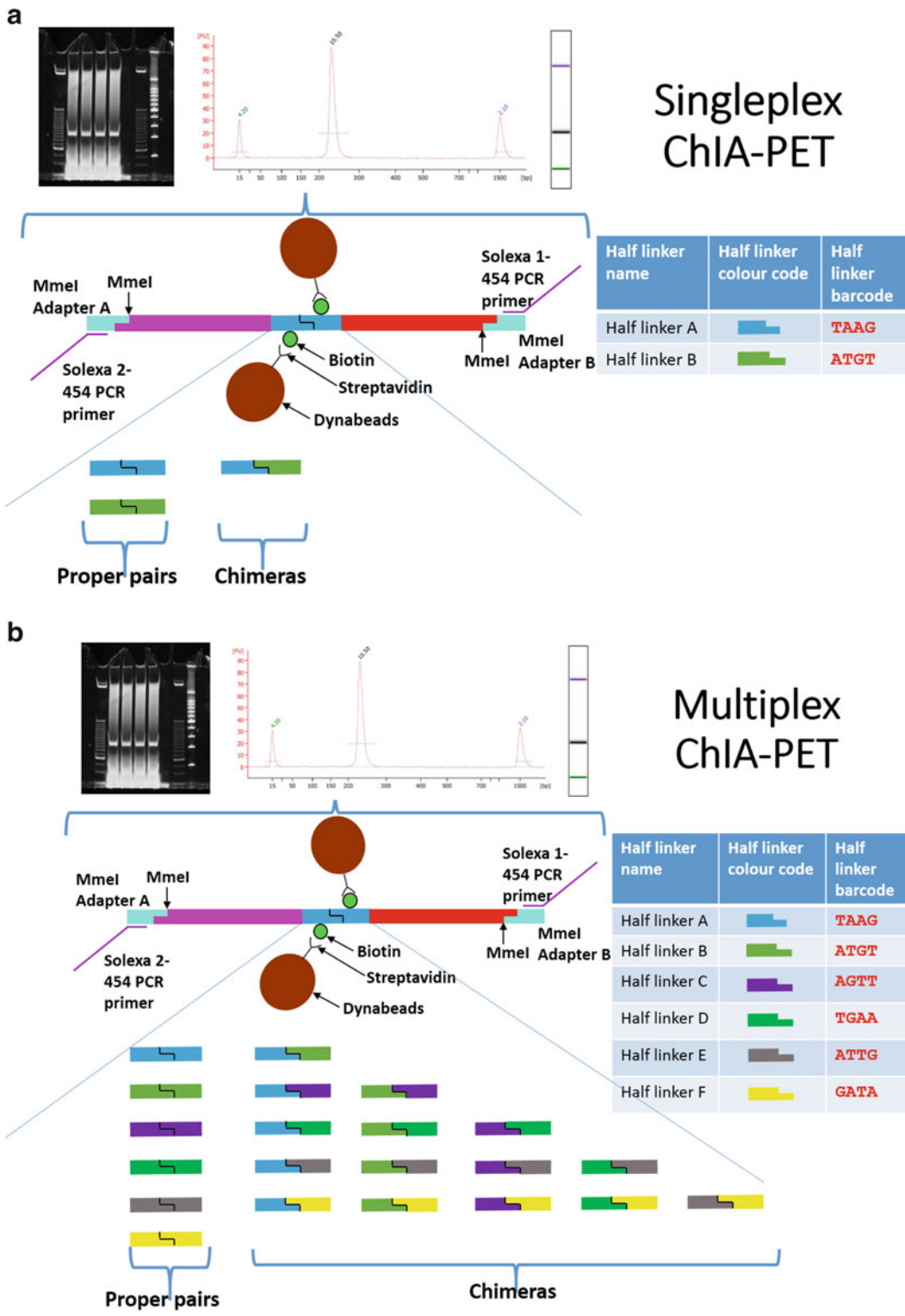
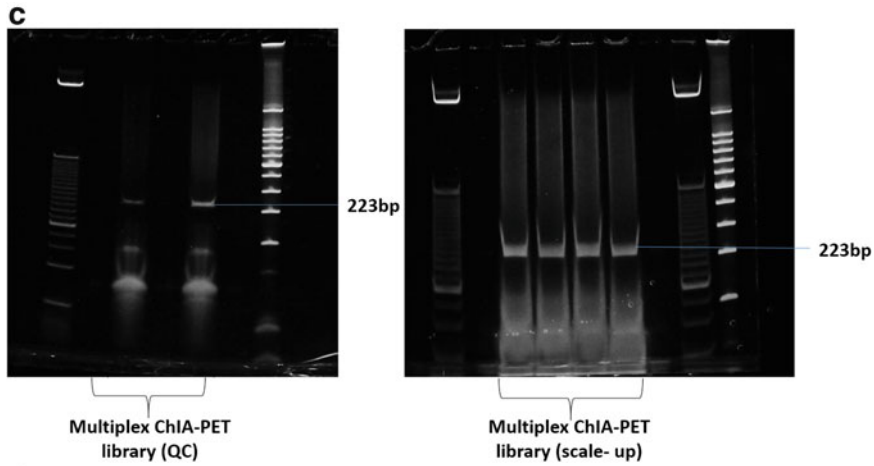
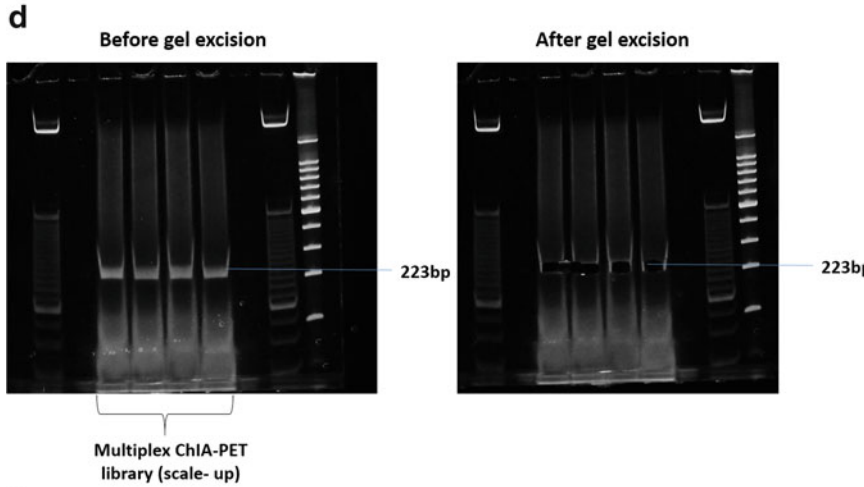


Fig. 5 (**a**) Singleplex ChIA-PET schematic. (**b**) Multiplex ChIA-PET schematic. Multiplex ChIA-PET involves the use of six pairs of barcoded half-linkers as compared to Singleplex ChIA-PET which uses only two pairs of barcoded half-linkers. (**c**) Ideal gel electrophoresis profiles of ChIA-PET QC (*left*) and ChIA-PET scale-up (*right*). (**d**) Gel electrophoresis profile before gel excision of library (*left*) and after gel excision of library (*right*). (**e**) Ideal Agilent Bioanalyser DNA 1000 profile of the Multiplex ChIA-PET library





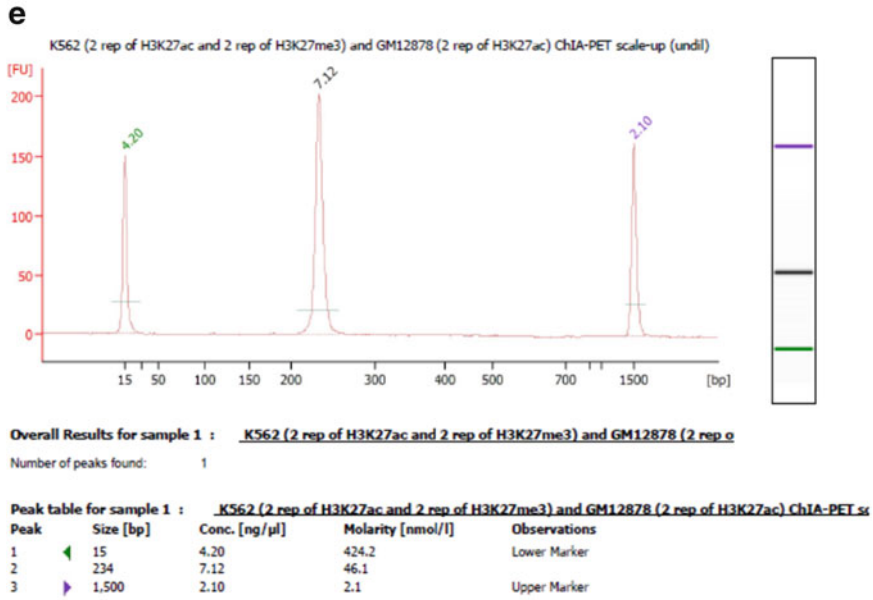


Fig. 5 (continued)

AGTC 5′.
MmeI Adapter B.
5′ CTGAGACACGCAACAGGGGATAGGCAAGGCACAC
AGGGGATAGG 3′.
3′ NNGACTCTGTGCGTTGTCCCCTATCC
GTTCCGTGTGTCCCCTATCC 5′.

3. ChIA-PET PCR primers.

Solexa 1-454.
5′ AATGATACGGCGACCACCGAGATCTACACCCTAT
CCCCTGTGTGCCTTG 3′.
Solexa 2-454:
5′ CAAGCAGAAGACGGCATACGAGATCGGTCCATCT
CATCCCTGCGTGTC 3′.

4. Sequencing primers.

Solexa 3-454 sequencing primer (Reverse primer).

5′-TGC GTG TCC CAT CTG TTC CCT CCC TGT CTC AG-3′.

Solexa 4-454 sequencing primer (Forward primer).

5′-GTG CCT TGC CTA TCC CCT GTT GCG TGT CTC AG-3′.

Acknowledgments

This work was supported by the National Research Foundation (NRF) Fellowship Grant to Melissa Jane Fullwood. The NRF grant number is R-713-000-143-281. In addition, this work is supported by NRF Funding and Ministry of Education funding to the Cancer Science Institute under the Research Centre of Excellence framework, as well as Yale-NUS and NTU start-up funds. This research is supported by the RNA Biology Center at the Cancer Science Institute of Singapore, NUS, as part of the funding under the Singapore Ministry of Education's Tier 3 grants.

References

1. Bartkuhn M, Renkawitz R (2008) Long range chromatin interactions involved in gene regulation. Biochim Biophys Acta 1783:2161–2166
2. Ma W, Ay F, Lee C et al (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Methods 12:71–78
3. Lai F, Orom UA, Cesaroni M et al (2013) Activating RNAs associate with mediator to enhance chromatin architecture and transcription. Nature 494:497–501
4. Xiang JF, Yin QF, Chen T et al (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res 24:513–531
5. Fullwood MJ, Liu MH, Pan YF et al (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462: 58–64
6. Zhang J, Poh HM, Peh SQ et al (2012) ChIA-PET analysis of transcriptional chromatin interactions. Methods 58:289–299
7. Li G, Fullwood MJ, Xu H et al (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol 11:R22

8. Zhang Y, Wong CH, Birnbaum RY et al (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504:306–310
9. Chen D, Fu LY, Zhang Z et al (2014) Dissecting the chromatin interactome of microRNA genes. Nucleic Acids Res 42: 3028–3043

Chapter 8

Identification of Transcribed Enhancers by Genome-Wide Chromatin Immunoprecipitation Sequencing

Steven Blinka, Michael H. Reimer Jr., Kirthi Pulakanti, Luca Pinello, Guo-Cheng Yuan, and Sridhar Rao

Abstract

Recent work has shown that RNA polymerase II-mediated transcription at distal *cis*-regulatory elements serves as a mark of highly active enhancers. Production of noncoding RNAs at enhancers, termed eRNAs, correlates with higher expression of genes that the enhancer interacts with; hence, eRNAs provide a new tool to model gene activity in normal and disease tissues. Moreover, this unique class of noncoding RNA has diverse roles in transcriptional regulation. Transcribed enhancers can be identified by a common signature of epigenetic marks by overlaying a series of genome-wide chromatin immunoprecipitation and RNA sequencing datasets. A computational approach to filter non-enhancer elements and other classes of noncoding RNAs is essential to not cloud downstream analysis. Here we present a protocol that combines wet and dry bench methods to accurately identify transcribed enhancers genome-wide as well as an experimental procedure to validate these datasets.

Key words eRNA, Chromatin immunoprecipitation sequencing, Global run on sequencing, Noncoding RNA, Transcribed enhancer, ENCODE

1 Introduction

Enhancers are distal *cis*-regulatory elements that, in contrast to promoters, activate gene expression independent of distance and orientation. Seminal work from several groups has described a series of epigenetic marks that define enhancer elements including a combination of histone marks that predict tissue-specific enhancers and their activity [1–5]. Histone H3 lysine 4 monomethylation (H3K4me1) is a hallmark for all enhancers, whereas the presence of histone H3 lysine 27 acetylation (H3K27Ac) further defines an active enhancer [4–6]. Consistent with these observations, the COMPASS complexes (which catalyze H3K4me1) and the histone acetyltransferase p300 (which catalyzes H3K27Ac) are commonly found at active enhancers in addition to promoters and gene bodies. Genome-wide locations of enhancer elements can be identified by

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_8, © Springer Science+Business Media New York 2017

profiling these histone marks, using chromatin immunoprecipitation coupled with next-generation sequencing (ChIP-Seq). While transcription factors, coactivators (Mediator), and low DNA methylation may be used to assist with identification enhancer elements, they are not required, thereby eliminating the need for additional cell type-specific datasets [7]. In addition, with the availability of publicly accessible databases, many of these epigenetic marks have been identified in a variety of cell types.

RNA Polymerase II (RNAPII) binds a subset of enhancers and produces a unique class of long noncoding RNAs termed eRNAs. eRNAs are bidirectionally transcribed and unspliced, making them distinct from other types of long noncoding RNAs (lncRNAs) and have been demonstrated by a number of groups to be a mark of highly active enhancers [8–14]. eRNAs have been shown to have diverse roles in regulating transcription in *cis* including stabilizing enhancer looping and regulating RNAPII phosphorylation state at gene promoters [15, 16]. Genes associated with eRNA producing enhancers are thought to be critical to controlling cell identity and lineage commitment [14]. Functionally, the enhancers described above are similar to super enhancers or stretch enhancers, which drive expression of genes critical to cell identity [17]. Identification of eRNAs is often achieved by overlaying ChIP-Seq datasets with genome-wide RNA sequencing datasets (e.g., global run on sequencing; GRO-Seq). Nonetheless, our own work demonstrates that the ChIP-Seq datasets alone can be used to identify highly active enhancers likely to produce eRNAs [14]. Rigorous analyses are essential as it is challenging to distinguish enhancer transcribed RNAs from other lncRNAs (e.g., long intergenic noncoding RNAs–lincRNAs).

Here, we describe in detail our procedure for accurate identification of eRNAs using a combination of wet and dry bench approaches. We use mouse embryonic stem cells (mESCs) as a model system because the transcriptional and chromatin regulatory networks controlling pluripotency have been well characterized on a genome-wide basis via integration of existing data sets. Specifically, we outline how to generate high quality ChIP DNA libraries for sequencing. An alternative starting point includes access to published datasets (e.g., ENCODE or GEO omnibus) that allow users to perform analyses *in silico*. Upon generation or download of ChIP-Seq datasets (H3K27Ac, H3K4me1, and RNAPII) we describe the dry bench analysis by which we: (1) define putative enhancers, (2) identify eRNA positive enhancers, and (3) exclude eRNA negative enhancers and other proximal *cis*-regulatory elements such as promoters. A dry bench strategy to eliminate non-enhancer elements (e.g., pseudogenes, microRNAs, and lncRNAs) that cloud analysis using computational approaches is essential. Lastly, we validate the ChIP-Seq data by wet bench approaches including ChIP-qPCR.

2 Materials

2.1 Solutions

1. Mouse ESC media: 500 mL Dulbecco's Modified Eagle's Medium (DMEM), 100 mL fetal bovine serum (FBS) Benchmark™, 12.5 mL Penicillin-Streptomycin Solution 100×, 6.25 mL l-glutamine, 100× liquid, 6.25 mL MEM non-essential amino acids, 6.25 mL EmbryoMax® Nucleosides (100×), 4.4 μL 100% 2-mercaptoethanol, 62.5 μL leukemia inhibitory factor (LIF). The final concentration of LIF is 10^3 μ/mL. Good for 3–4 weeks at 4 °C.
2. 1× DPBS: 5 mL 10× DPBS, 45 mL autoclaved reverse osmosis (RO) water.
3. 2.5 M glycine: 187 g glycine, 1 L RO water. Good for 1 year.
4. 70% Ethanol: 35 mL 100% ethanol, 15 mL RO water.
5. 10% sodium deoxycholate: 5 g sodium deoxycholate, 50 mL RO water.
6. SDS Lysis Buffer: 250 μL 20% SDS, 200 μL 0.5 M EDTA, 1.5 mL 5 M NaCl, 500 μL Triton X-100, 1 mL Tris–HCl pH 8.0, 46.6 mL RO water. Store at 4 °C. Good for 6 months when stored without protease inhibitors.
7. Low-Salt Wash Buffer II: 250 μL 20% SDS, 200 μL 0.5 M EDTA, 1.5 mL 5 M NaCl, 500 μL Triton X-100, 1 mL Tris–HCl pH 8.0, 46.6 mL RO water. Good for 6 months.
8. Wash Buffer III (LiCl): 2.5 mL 5 M LiCl, 2.5 mL 10% NP40, 2.5 mL 10% deoxycholate, 100 μL 0.5 M EDTA, 500 μL Tris–HCl pH 8.0, 41.9 mL RO water. Good for 6 months.
9. TE: 500 μL 1 M Tris–HCl pH 8.0, 100 μL 0.5 M EDTA, 49.4 mL RO water.
10. SDS Elution Buffer: 2.5 mL 20% SDS, 1 mL 0.5 M EDTA, 2.5 mL 1 M Tris–HCl, 44 mL RO water. Good for 6 months.

2.2 Lab Equipment

1. Phase lock gel tubes—heavy (5 PRIME—2302810).
2. Qubit® Fluorometer.
3. Dynal magnetic separation rack.
4. Qsonica Q125 Sonicator with 1/8″ in diameter tip or Diagenode Bioruptor® Pico.
5. 1.5 mL Bioruptor® microtubes (Diagenode C30010016) if using the Bioruptor® Pico.
6. Eppendorf Tubes® 5.0 mL if using the Qsonica Sonicator.
7. Bioanalyzer.
8. AMPure® XP Beads.

2.3 Chemicals

1. Mouse ESC media.
2. 1× DPBS.

3. 2.5 M glycine.
4. 70% Ethanol.
5. 10% Sodium deoxycholate.
6. SDS Lysis Buffer.
7. Low-Salt Wash Buffer II.
8. Wash Buffer III (LiCl).
9. TE.
10. SDS Elution Buffer.
11. 16% Methanol-free formaldehyde.
12. Protease inhibitor cocktail.
13. Phenylmethylsulfonyl fluoride (PMSF).
14. Protein A or G beads.
15. Phenol:Chlorofrom:Isoamyl Alcohol (25:24:1 v/v).
16. Agarose powder.
17. RNase A.
18. Proteinase K.
19. Glycogen.
20. 3 M Sodium acetate.

2.4 Kits

1. Qubit® dsDNA HS Assay Kit.
2. NEBNext® ChIP-Seq Library Prep Master Mix.
3. NEBNext® Singleplex or Multiplex Adapters. Adapter combinations will vary based on sample number and complexity of library. Refer to protocol for pooling and adapter ligation included with the ChIP-Seq library kit.

2.5 ENCODE Antibodies

1. H3K4me1 (Abcam ab8895).
2. H3K27Ac (Abcam ab4729).
3. RNA Polymerase II (Abcam 8WG16).
4. H3K36me3 (Abcam ab9050-optional).

3 Methods

3.1 Wet Lab Protocol for ChIP

3.1.1 Cell Preparation and Chromatin Immunoprecipitation

This protocol is designed to perform ChIP in mESCs for endogenous proteins and will need to be optimized for additional cell types. The total time from starting the procedure to having ChIP DNA ready for downstream processing (such as quantitative PCR or ChIP-Seq library generation) is 4 days (Fig. 1). This does not include the preparation/splitting of cells [18–22].

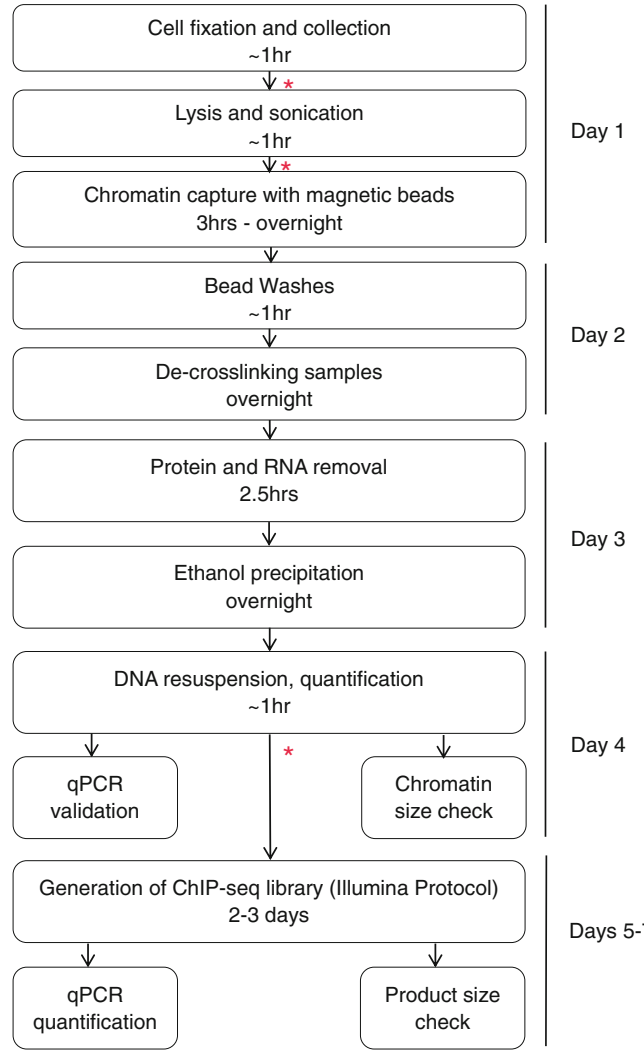


Fig. 1 Flowchart of wet bench protocol to generate ChIP-Seq library. * Indicates a safe stopping point in the protocol, overnight or a couple days

Prior to Day 1

Prepare mESCs on gelatin-adapted plates. This protocol is written for two 15 cm plates that are 30–60% confluent, which would yield approximately 100 million cells total (*see* **Note 1**).

Day 1

1. Add 1250 μL of 16% formaldehyde to 20 mL media and cells (final concentration 1.0%). Incubate at room temperature for 5 min with gentle rocking to mix (*see* **Note 2**).
2. Quench formaldehyde with 1.0 mL of 2.5 M Glycine (final concentration 125 mM). Incubate at room temperature for 5 min with gentle rocking to mix.
3. Rinse plate with 20 mL ice-cold Dulbecco's phosphate-buffered saline (DPBS) (without magnesium and calcium)

containing 1:1000 protease inhibitors (PI) and 1:200 phenylmethylsulfonyl fluoride (PMSF). Rinse fixed cells 3× total. Keep plates on ice while rinsing fixed cells.

4. Add 15 mL DPBS containing inhibitor 1:100 PI and 1:200 PMSF and scrape fixed cells into a 50 mL conical tube on ice. Rinse plate two more times with 15 mL DPBS and collect with initial scraping.
5. Centrifuge at 750 × *g* for 10 min at 4 °C to and aspirate supernatant. Transfer cell pellet to a 1.7 mL microcentrifuge tube and flash freeze the sheared chromatin pellet on Dry Ice. Store at −80 °C for up to several months.
6. Lyse cells with 1 mL SDS Lysis Buffer containing inhibitors (1:100 PI and 1:200 PMSF) for each 15 cm plate that was approximately 30–60% confluent to start. Pipette up and down to break apart aggregates of fixed cells.
7. Transfer to a 5 mL Eppendorf tube and incubate 10 min on ice. A large 5 mL tube allows the Qsonica microtip to be inserted without touching the sides of the tube, yet still come very close to the bottom of the conical (*see* **Note 3**).
8. Proceed with sonication using a microtip. Each sample should receive three cycles at Amplitude = 5 in ice water. Each cycle should consist of a burst of 1 s on and 4 s off, for a total of 30 s on. There should be a 3 min pause between each cycle (*see* **Notes 4** and **5**).
9. Pellet insoluble fraction by spinning at maximum speed for 10 min at 4 °C. Transfer supernatant to a new 1.7 mL microcentrifuge tube.
10. Remove a small aliquot (100 µL) to be saved as Input/genomic DNA in a screw cap microcentrifuge tube. Store at −80 °C. If needed, the samples can be frozen at −80 °C for months.
11. Boil 50 µL of each sample for 15 min. Spin at max speed in a microcentrifuge for 5 min at room temperature. Run 10–20 µL on a 1% agarose gel. The bulk of the decross-linked DNA should be 200–500 base pairs (bp) (*see* **Note 6**) (Fig. 2).
12. Add 4–8 micrograms of antibody to chromatin and place at 4 °C overnight (*see* **Note 7**).

Day 2

1. Pipette 50–100 µL of Protein A or G Dynabeads into a fresh 1.7 mL microcentrifuge tube and place into magnetic separation rack for 2 min (*see* **Note 8**).
2. Remove liquid using a 1 mL micropipette. Resuspend in 1 mL ChIP Lysis Buffer with 1:1000 PI and 1:200 PMSF and rotate for 5 min at 4 °C.
3. Quick spin samples to pull down liquid from cap and place tubes into magnetic separation rack for 2 min and remove liquid. Wash beads 3× total in ChIP Lysis Buffer with inhibitors.

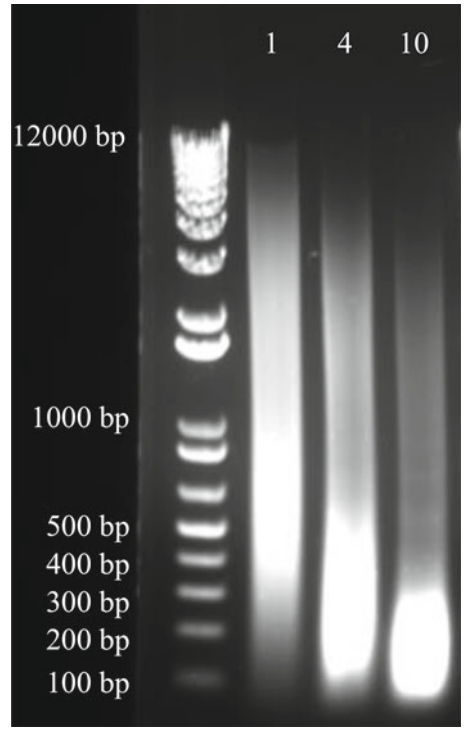


Fig. 2 Approximately 20 million mouse embryonic stem cells were fixed for 5 min with 1 % formaldehyde. 300 μL of each sample was sheared in 0.1 % SDS Lysis Buffer using the Diagenode Bioruptor® Pico. From left to right on the gel, samples were subjected to 1, 4, and 10 cycles of sonication. One microgram of decross-linked and RNase/proteinase K sample was separated by electrophoresis on a 1 % agarose gel that was stained with ethidium bromide. Optimally, sheared chromatin will yield a smear between 200 and 500 bp (as seen with 4 cycles above). One cycle yields under-sheared chromatin (400–1000+ bp) and ten cycles produces over-sheared chromatin (100–300 bp)

4. Transfer supernatant containing sheared chromatin and antibody to tubes with washed Dynabeads.
5. Rotate for at least 3 h at 4 °C.
6. Quick spin the samples to bring down the liquid and place in magnetic rack for 2 min.
7. Remove liquid with a 1 mL micropipette to avoid disturbing beads.
8. Washes can be performed at room temperature and should be quick to prevent the beads from drying out.
9. Wash the tubes using the following procedure: add 1 mL of wash buffer and resuspend by pipetting, place in tube rotator at 4 °C for 10 min, quick spin to bring down the liquid, place in Magnetic Rack for 2–3 min at room temperature, carefully

pipette all liquid without disturbing beads, remove sample from rack, and proceed with next wash buffer. Gently pipette samples up and down to ensure aggregates of beads are broken up. You may use fresh tubes for each wash, to ensure there is no carryover.

10. Wash beads with 1 mL Buffer in the following order: ChIP Lysis Buffer (1×), Low-Salt Wash Buffer II (1×), Wash Buffer III (LiCl) (1×), and TE (1×) (*see* **Note 9**).
11. After final wash, remove all traces of TE with another spin and resuspend beads in 150 µL SDS Elution Buffer.
12. Transfer all samples to screw cap microcentrifuge tubes to minimize evaporation.
13. Incubate at 65 °C overnight (preferably in a water bath to minimize evaporation). Remove the saved Input sample and begin to process in parallel. This performs both the decross-linking and the elution in a single step.

Day 3

1. Quick spin the samples and place into magnetic rack for 3 min. Input sample should be spun at maximum speed for 10 min at room temperature. Transfer supernatant to a new microcentrifuge tube and bring volume to 200 µL with TE.
2. Place new microcentrifuge tube with supernatant into magnetic rack for another 3 min to ensure all beads are removed.
3. Add 2 µL of RNase A to each sample (including Input) and incubate at 37 °C for 30 min.
4. Add 2 µL of Glycogen and 4 µL of Proteinase K to each sample and incubate at 37 °C for 2 h. Glycogen is added as a DNA carrier.
5. Pre-spin 2 mL phase lock tubes for 2 min at maximum speed to pellet resin.
6. Transfer sample to 2 mL phase lock tube. Add 1 volume (200 µL) of Phenol:Chloroform:Isoamyl Alcohol. Mix well by inverting at least 10× and spin at maximum speed for 5 min at room temperature.
7. Add 1/10th volume (20 µL) 3 M Sodium Acetate and 2.5 volumes (0.5 mL) 100% Ethanol to the tubes. Place samples on Dry Ice until they freeze completely. Place at –20 °C overnight to maximize DNA yield.

Day 4

1. Spin at maximum speed for 15 min at 4 °C. Carefully use a 1 mL micropipette and remove supernatant, preserving the small white pellet. Quick spin a second time to ensure all liquid is at bottom of tube and remove remaining liquid.
2. Air-dry the sample for 3–5 min. Do not over-dry the DNA pellet.
3. Resuspend in 25–50 µL of water. Quantitate DNA by Qubit DNA High Sensitivity at a 1:40 dilution. Sample can now be

stored at −80 °C indefinitely for downstream applications. Aliquot samples to avoid freeze thaw (*see* **Note 10**).

4. Run a small amount of precipitated DNA from samples and Input (if you have excess DNA) in a 1% agarose gel to ensure proper sonication (Fig. 2). Input sample is preferred for ChIP-Seq to assess enrichment of proteins.
5. ChIP DNA is quality controlled using an Agilent Bioanalyzer. The Bioanalyzer validates the size of DNA fragments (200–500 bp) and determines the concentration and purity of the sample.

3.1.2 ChIP-Seq Library Generation and Validation

1. ChIP-Seq libraries are generated using the NEBNext® ChIP-Seq Library Prep Master Mix Set for Illumina according to manufacturer's instructions (*see* **Note 11**).
2. ChIP-Seq libraries are quality controlled using an Agilent Bioanalyzer prior to sequencing. Typically, the tracing will show a narrow range of products between 150 and 500 bp depending on the size of the original ChIP DNA. Details are provided in the library generation kit to facilitate decision about whether the library is of sufficient quality to provide good quality sequencing results.
3. It is critical to validate that the ChIP-Seq library is representative of the precipitated ChIP DNA in Subheading 3.1.2.3, **step 38**. Test for enrichment of protein at active enhancers (positive control) and inactive enhancers (negative control) by ChIP-qPCR prior to sequencing. As little as 0.1 ng DNA can be used for each reaction. Perform ChIP-qPCR on the Input and include a negative control antibody (e.g., IgG sample) (*see* **Notes 12** and **13**) (Fig. 3).
4. After protein enrichment is confirmed, sequence on an Illumina HiSeq and obtain a minimum of 20–40 million reads for H3K4me1 or H3K27Ac and 10–20 million reads for RNAPII. Higher reads are used for histone marks because they typically bind larger chromatin regions rather than a specific DNA element. Paired-end sequencing can be performed, but typically does not provide additional information. Indexing will depend on the run type and number of samples.

3.2 Dry Bench Analysis of ChIP-Seq to Identify Putative Enhancers

Discriminative filters and thresholds are used to specifically identify enhancers and not other *cis*-regulatory elements that may act as distal or alternative promoters. Moreover, non-enhancer elements that produce other classes of long noncoding RNAs need to be eliminated so they do not cloud analyses and computational approaches. Additionally, intragenic enhancers must be filtered to eliminate coding strand transcripts. For optimum computational performance, a minimum system requirement of 4 cores and 16GB RAM or more is recommended. Most of the ChIP-Seq computational analyses are done in Unix-like operating systems given the availability of several

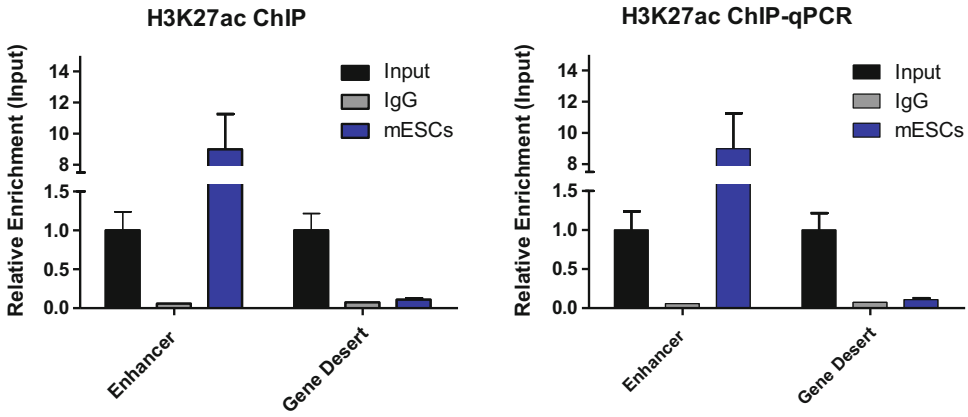


Fig. 3 ChIP-qPCR showing enrichment of the activating histone mark H3K27Ac at a pluripotency associated enhancer. Primers within a gene desert on chromosome 6 were used as a negative control. Rabbit IgG was used as mock control. Values were normalized to primers within the promoter of GAPDH

methods targeting these systems and that often provide a command line interface for their execution. R and Python are used to perform statistical analysis and to automatize analysis of the genomic data. In this section, we describe the tools and data formats used to analyze ChIP-Seq datasets to define putative enhancers (Fig. 4). The dry bench datasets generate a variety of different types. For a brief overview of file types and the data they contain, please see http://www.broadinstitute.org/software/igv/FileFormats

3.2.1 Data Mining and Retrieval

Public ChIP-Seq Data Files

1. If ChIP-Seq datasets are published for your tissue of interest, they can be downloaded from a freely available online repository such as GEO Omnibus (http://www.ncbi.nlm.nih.gov/geo/), RIKEN-FANTOM (http://fantom.gsc.riken.jp/data/), or EMBL-EBI (http://www.ebi.ac.uk/ena) [23].
2. Use the SRA toolkit that has a set of data-dump utilities, which will allow reformatting from SRA to FASTA, FASTQ, or SAM. The SRA toolkit can be downloaded from NCBI (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) and is available for Mac, Linux, and Windows operating systems.
3. Use "fastq-dump" utility to convert .sra to .fastq file format to generate a FASTQ file from SRA file(s). Each read/sequence in FASTQ file consists of 4 lines. The first line starting with "@" indicates the read identifier. The second line is the actual DNA sequence. The third line starting with "+" is an optional title line. The fourth line is the quality score symbol for each base in the sequence which is encoded in ASCII character code following usually the PHRED33 convention (other quality encodings may be used depending on the Illumina software, for more information refer to: https://en.wikipedia.org/wiki/FASTQ_format#Quality).

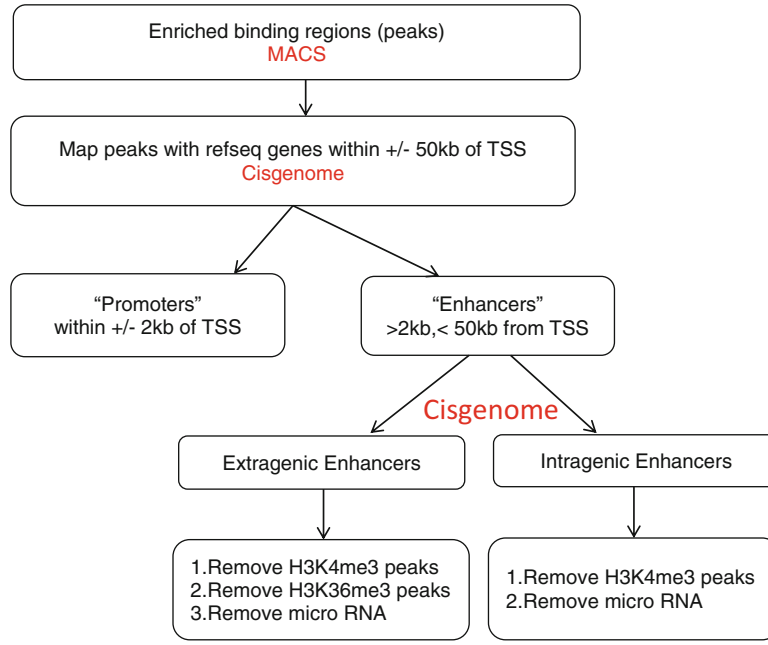


Fig. 4 Schematic representation of the workflow of enhancer detection. Peaks are called by MACS. Mapping/annotation to nearest gene is executed by Cisgenome. Distance based parsing is performed to categorize the peaks to enhancers and promoters. Enhancers are divided into extragenic and intragenic to prevent clouding of downstream analysis. Further filtration is performed to remove promoters, unannotated genes, and noncoding RNAs

Private ChIP-Seq Data from Server

1. Download raw data (typically a FASTQ file) from your sequencing instrument.
2. If data is provided in a SRA format, use the NCBI SRA toolkit to convert data to achieve a data format in order to run the alignment (*see* **Note 14**).

Quality checking FASTQ files

1. Perform quality control checks using FastQC (developed at Babraham Institute) to ensure that raw data from single-end reads is free of biases (originating from sequencing or library preparation) [24]. Decoding quality scores in a FASTQ file depends on the type of platform used. Sanger, Illumina, Solexa, and PHRED software reads the DNA sequences, calls bases, and assigns a quality value for each base called. PHRED33 quality score is the most common quality metric adopted. For Illumina the quality ranges from 0 to 62 and base pair quality score of 20 is minimally required to trust the DNA nucleotide identified. Some modules in FASTQC that are helpful to judge your sequence are Per Base Sequence Quality Report which can help you decide if sequence trimming is needed before alignment. The Sequence Duplication Level Report is informative for library enrichment. The Overrepresented Sequence Report assesses for adapter contamination.

2. Check for barcodes after downloading the FASTQ file. Any adapter sequences that are used in sequencing library construction should be trimmed, for example, using the Trimmomatic utility (http://www.usadellab.org/cms/?page=trimmomatic) [25].
3. Trim low quality sequences. Low quality reads could have high sequencing error, resulting in misalignment to the reference genome. Other preprocessing tasks for FASTQ files such as filtering sequences based on quality, formatting the width of sequences, converting the FASTA sequence to RNA/DNA, etc. can be done using FASTX-toolkit [26].

Mapping Reads to the Reference Genome

4. ChIP sequencing is most often performed with single-end reads. Use Bowtie 1.1.2 algorithm to align single-end reads to mouse genome build mm9 using parameters (–p 6 –n 2 –l 49 –e 70 –m 1 – – best for unique mapping), which allows a maximum of 2 mismatches (n) in the 49 bases (l) and 1 unique alignment per read (m) and uses 6 cores (p). If your machine has more than 6 cores you should adjust this parameter (*see* **Note 15**) [27].
5. The output is a TAB-delimited Sequence Alignment/MAP (SAM) file describing mapped alignments (now known as "tags") of sequencing reads to a reference sequence.
6. Convert SAM to BAM format using Samtools [28, 29]. BAM is a compressed and binary equivalent of SAM.
7. Use SAMtools which has a set of utilities to manipulate the alignments in BAM format for further downstream analysis. SAMtools can be adopted for merging, sorting, and indexing the BAM files.

3.2.2 Peak Calling

1. Peak calling is done to identify the binding sites for RNAPII or histone modifications. MACS2 2.1.0 (model-based analysis of ChIP-Seq) is used to identify significantly enriched regions (sites of DNA-protein binding peaks) over background (the Input samples used to estimate the per base pair noise levels) [30].
2. MACS reports all binding sites with p-values below a defined threshold (default 10^{-5}) in a BED format. Set a p-value threshold of enrichment of 10^{-5} and option – – broad to identify H3K4me1 and H3K27Ac ChIP-Seq data since distribution of histone reads have a continuous property and peaks are broad. Only H3K4me1 and H3K27Ac peaks greater than 1 kb in length are considered in this analysis. This assists with eliminating spurious genomic regions that are less likely to possess enhancer function. In addition, given that eRNAs are a lncRNA, this size discrimination assists in eliminating other elements that may produce small noncoding RNAs. A p-value of 10^{-6} is used to detect narrow well-defined (non-broad) RNAPII ChIP-Seq data (*see* **Note 16**).

3.2.3 Putative Enhancer Detection

Enhancers are noncoding DNA elements that act independent of distance and orientation to regulate gene transcription. However, many of the marks described above are not exclusive to enhancers. As a result, genomic elements that mimic transcribed enhancers (e.g., pseudogenes and microRNAs) must be removed to allow for more accurate analysis of eRNA producing enhancers. However, more sophisticated analyses require generation or availability of additional histone ChIP-Seq datasets.

1. Map the called peaks in the previous step with the nearest genes using UCSC RefFlat annotations. Use Cisgenome tool to map all ChIP-Seq tag peaks to annotated genes that are ±50 kb of TSS [31]. Specifically, use the feature refgene_getnearestgene with options –r 1 –up 50,000 –down 50,000.
2. Remove any peaks located in promoter regions. For this analysis, promoters are defined as regions 2 kb upstream and downstream of the TSS (4 kb total) (*see* **Note 17**).
3. The resulting peak list should have all the enhancer regions between 2 and 50 kb of the nearest neighbor gene TSS. Enhancers >50 kb from the TSS of a gene can be saved by altering the options in 3.2.3.1, if desired.
4. To determine whether enhancers are located within actively transcribed genes, use Cisgenome (refgene_getlocationsummary) to classify the enhancer as intragenic versus extragenic.
5. This step requires additional histone modification datasets. Use BEDTools to eliminate extragenic and intragenic enhancers that overlap with a region of H3K4me3 to remove any unannotated gene or other classes of ncRNAs. Extragenic enhancers that overlap with H3K36me3 regions should be eliminated for the same reason. This cannot be used for intragenic peaks since many intronic and exonic enhancers may show some degree of H3K36me3 enrichment (*see* **Notes 18** and **19**) [32].

3.3 Validating ChIP-Seq Data

3.3.1 Detection of Transcribed Enhancers by GRO-Seq Overlay

There is rapidly growing evidence that eRNA production is a mark of a highly active enhancer and that eRNAs have diverse roles transcriptional regulation. eRNA producing enhancers can be identified by overlapping RNAPII bound enhancers with GRO-Seq datasets. Not surprisingly, enhancers bound by RNAPII show higher eRNA production rates than unbound sites (*see* **Note 20**).

1. Use BedTools (intersectBed with –f 0.5 –r) to identify enhancers that overlap with RNA Pol II (50% minimum overlap). We have found that enhancers occupied by RNAPII are highly enriched for eRNA production (*see* **Note 21**) [14].
2. To estimate expression levels for enhancers that are bound by RNAPII, processed GRO-Seq data available on GEO omnibus (GSE27037) was downloaded.

3. For extragenic enhancers, use BEDTools suite (coverageBed) to count RNA reads from both strands.
4. For intragenic enhancers, use BEDTools suite (coverageBed) to count RNA from only the antisense strand to prevent counting reads from sense-strand gene transcription. Since, the sense strand is the coding strand, there should be approximately half as many reads. Accordingly, transcribed intragenic enhancers cannot be directly compared with transcribed extragenic enhancers. Genes need to be separated by coding strand and counted separately. Using this approach, intragenic enhancers that produce eRNAs can be identified, with approximately half the number of transcripts of extragenic enhancers [14].
5. Compute RPKM (reads per kilobase of genomic region per million mapped reads) for each enhancer that is associated with the nearest gene (*see* **Note 22**).

3.3.2 Wet Bench Approach to Validate Presence of H3K4me1, H3K27Ac, RNAPII, and Tissue Specific eRNAs

1. Confirm enrichment of protein at an enhancer by qPCR with ChIP DNA as described in Subheading 3.1.2.
2. Validate the presence of cell type-specific eRNA production by RT-qPCR (*see* **Note 23**).

4 Notes

1. Do not perform cross-linking on plates with a large number of dead cells. Change media the morning of cross-linking to remove dead cells. Let cells incubate for 2–3 h to ensure they equilibrate. If combining more than one plate be sure to scale up volumes.
2. Formaldehyde mediated cross-linking is one of the key aspects to both data quality and reproducibility from ChIP. Ideally, a short enough incubation time is used to cross-link DNA and proteins within close physical proximity, without causing distal interacting sites/proteins to cross-link. Fixing cells for too long may reduce the number of available epitopes and make it more difficult to lyse and shear the chromatin, thus reducing DNA yields. It may also make reverse cross-linking more difficult which will interfere with downstream steps. For the vast majority of cells, cross-linking is between 5 and 7 min at room temperature, and rarely requires more than 10 min. Use fresh formaldehyde as air and light exposure can change the contents. Methanol-free formaldehyde is preferred as methanol can disrupt cell membranes and effect lysis. We find individual ampules of methanol free formaldehyde reduces the variability from assay to assay significantly.
3. Sonication is arguably the most important step of a ChIP assay. The sonication microtip should be consistently placed as close to the bottom of the 5 mL conical tube as possible for all samples. This prevents foaming and ensures similar sonication

between samples. If there is significant frothing/foaming, pause, remove sample and spin down quickly in a microcentrifuge to remove foam, and restart. The most likely cause of frothing is because the tip is not close enough to the bottom of the tube. Make sure the microtip does not contact the tube (bottom or sides). If you see precipitate, you may want to discard the sample and fix new cells if available.

4. Each cell type requires different sonication conditions and SDS Lysis Buffer. If SDS Lysis Buffer requires a SDS concentration greater than 0.1%, samples must be diluted (final concentration of 0.1% or less) prior to adding antibody. SDS interferes with the antibody epitope interaction. It may also affect downstream PCR. Moreover detergents (e.g., SDS) can precipitate out of solution at temperatures lower than 15 °C when stored too long. Prepare fresh lysis buffer for each experiment. If using a different number of cells (by greater than a factor of 2), type of cells, tube, or sample volume, you will need to reoptimize sonication conditions to ensure adequate fragmentation in the minimal number of cycles. Optimal size fragments are in the 200–500 bp range (Fig. 2). Fragments greater than 500 bp do not pull down as well and may result in an increase in nonspecific binding in the ChIP assay. Over shearing chromatin (100 bp or less) can be detrimental to downstream applications such as ChIP-qPCR. Over shearing may also damage proteins and alter epitopes.
5. As an alternative to using a microtip, many ChIP-Seq data sets are created using a Diagenode Bioruptor® for sonication. A Bioruptor® allows you to shear multiple samples at one time and eliminates variation due to microtip placement. Moreover, problems noted above including frothing/foaming are eliminated. For mESCs we use sonication conditions of 30 s On, 30 s Off for 4 cycles. 1.5 mL Diagenode Bioruptor® microtubes containing 300 μL ChIP Lysis Buffer plus inhibitors with approximately 15 million cells are used for each sample.
6. Gel electrophoresis of boiled and sheared chromatin on Day 1 is a quick method to check sonication efficiency. However, to be safe, a small amount of precipitated Input/genomic DNA should be run out to confirm that the sonication was optimal. This is representative of the ChIP DNA pulled down after RNase A and Proteinase K treatment. The band range of precipitated DNA may differ from the boiled and sheared chromatin (Fig. 2).
7. When possible, use ChIP-Seq grade antibodies that are published and preferably used to create a ENCODE dataset. Using more than the indicated amount of antibody does not result in greater DNA yield and may lead to more nonspecific binding, thereby interfering with downstream analysis. Antibodies for common histone marks such as H3K4me1 result in a high yield of DNA; therefore, less sheared chromatin may be used. For more information on how to test and validate antibodies see [33].

8. To ensure magnetic bead/antibody interaction, you can use a 50:50 mixture of protein A and protein G beads.
9. For this protocol, the ChIP Lysis Buffer and Low-Salt Wash Buffer II are the same because we lyse mESCs in 0.1 % SDS. For other cell types you may have to increase the percentage of SDS in the ChIP lysis Buffer, but Low-Salt Wash Buffer II should stay at 0.1 %.
10. A fluorometry based approach is necessary to quantify ChIP DNA. Spectrometry-based methods do not distinguish between RNA, double stranded DNA, single stranded DNA, and free nucleotides. QuBit 2.0 Fluorometer is more sensitive and accurate than spectrometry-based methods because it uses a fluorescent dye that specifically intercalates into double stranded DNA. This allows quantification of very low amounts of DNA (as low as 10 pg/μL) without interference due to other nucleotide species.
11. The ChIP-Seq library preparation kit can be purchased for any platform (although Illumina HiSeq is the most common). Alternatively it is often more efficient and cost effective to have the company performing the sequencing make the library.
12. To ensure quality of the precipitated DNA always include a negative control. A good antibody for negative control in mESCs is IgG. Verify by ChIP-qPCR (Fig. 3).
13. Negative control primers can be used for all samples if designed in gene deserts. Positive control primers for RNAPII may be designed at a known active promoter or enhancer in tissue of interest. Alternatively primers can be designed after downstream ChIP-Seq analysis based on the presence of a ChIP-Seq tag peak. ChIP-Seq tag peaks can be viewed by uploading files to Integrated Genome Viewer (IGV). ChIP-Seq tag peaks correspond to enriched presence or binding of the target protein. Be sure to run a melting curve when using new primers to ensure that you are amplifying a single PCR product.
14. There are other files to browse to look for run settings, quality metrics, etc., from your sequencer report. The raw FASTQ files are necessary to publish data. Create a backup as soon as you download your raw files.
15. Bowtie2 is generally faster and more sensitive than Bowtie1 for reads longer than 50 bp. Set seed length (l) to length of the read for each data file. Specifying the number of parallel search threads (p) increases alignment throughput. Option –m and –– best in Bowtie results in fewer unique alignments than just specifying –m. For paired-ends, the alignment can be time consuming. Option –I and –X in Bowtie are critical to get fair percentage of aligned reads. Other popular short read aligner algorithms (ELAND) could be used depending upon the type

of data. BWA is used for exome sequence reads, whereas TopHat and STAR are for RNA-Seq data. A minimum of 4–10 GB of RAM is required to run Bowtie. Bowtie can utilize all cores on a node. For a single alignment run job, you can specify the number of cores to use with the –p option. The index files can be downloaded from Bowtie website for the most common assembly (mm8, mm9, mm10).

16. MACS uses control samples to minimize bias and calculates an empirical false discovery rate (FDR). p-Values vary for different datasets depending on strength of enrichment. A good way to choose the best p-value is to visualize the signal files (wiggle files) in the genome browser and to look for peaks called by MACS. To determine if the ChIP experiment worked, sort FDR from lowest to highest and then sort fold enrichment from highest to lowest and look for the number of peaks. There should be one to several thousand peaks.
17. The size of a promoter can be around 3 to 5 kb. Simple distance based calculations in Microsoft Excel were used to identify promoters in the output Cisgenome yielded. Promoters can then be removed using Microsoft Excel software.
18. Extragenic and intragenic peaks that overlap with a region of H3K4me3 (a mark of promoters) may be eliminated to remove any unannotated gene or other classes of noncoding RNAs. However, many eRNA producing enhancers have higher levels of H3K4me3; thus, this stringent filter will remove some transcribed enhancers prior to downstream analysis. Extragenic peaks that overlap with H3K36me3 (an epigenetic mark found in gene bodies and long intergenic noncoding RNAs) may be eliminated for the same reason. Intragenic peaks may not be removed since many intronic and exonic enhancers may show some degree of H3K36me3 enrichment. Elimination of these peaks can be done using BEDTools (intersectBed) [32]. The same filtering methods described in Subheading 3 can be used to identify RNAPII, transcription factors, or coactivators at enhancers.
19. IntersectBed from BedTools was used to get the overlapping regions and nonoverlapping regions.
20. GRO-Seq is more sensitive than RNA-Seq at capturing nascent RNAs, which have properties more similar to eRNAs.
21. In this protocol we describe a direct way to identify eRNA producing enhancers using RNAPII ChIP-Seq and GRO-Seq. However, transcribed enhancers can be indirectly identified by very high levels of H3K27Ac and H3K4me3.
22. Use TopHat to align GRO-Seq or RNA-Seq data to the genome and then use Cufflinks to quantify the abundance of transcript.
23. eRNAs are expressed at levels 1:100 to 1:1000 of the mRNA of the gene they are associated with. Thus, it is important to confirm detection of tissue specific expression of eRNAs and not

background/noise. By RT-qPCR we compare cDNA from pluripotent ESCs to cells that were treated with 5uM retinoic acid for 6 days which induces complete differentiation. Given that eRNAs are unspliced, it is essential to have DNA free RNA to make sure you are only amplifying cDNA (run a no reverse transcriptase control). Moreover, not all eRNAs can be converted to cDNA during reverse transcriptase using either oligo dT or random hexamers. We use BioRad iScript cDNA Synthesis Kit because it combines both oligo dT and random hexamers.

References

1. 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
2. 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
3. Visel A, Blow MJ, Li Z et al (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854–858
4. Zentner GE, Tesar PJ, Scacheri PC (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res 21:1273–1283
5. Rada-Iglesias A, Bajpai R, Swigut T et al (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470:279–283
6. Creyghton MP, Cheng AW, Welstead GG et al (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107:21931–21936
7. Stadler MB, Murr R, Burger L et al (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480:490–495
8. Kim T-K, Hemberg M, Gray JM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
9. De Santa F, Barozzi I, Mietton F et al (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8, e1000384
10. Koch F, Fenouil R, Gut M et al (2011) Transcription initiation platforms and GTF recruitment at tissue specific enhancers and promoters. Nat Struct Mol Biol 18:956–963
11. Wang D, Garcia-Bassets I, Benner C et al (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474:390–394
12. Orom UA, Derrien T, Beringer M et al (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58
13. Lai F, Orom UA, Cesaroni M et al (2013) Activating RNAs associate with mediator to enhance chromatin architecture and transcription. Nature 494:497–501
14. Pulakanti K, Pienello L, Stelloh C et al (2013) Enhancer transcribed RNAs are produced from hypomethylated genomic regions in a Tet-dependent manner. Epigenetics 8:1303–1320
15. Schaukowitch K, Joo JY, Liu X et al (2014) Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell 56:29–42
16. Maruyama A, Mimura J, Itoh K (2014) Noncoding RNA derived from the region adjacent to the human HO-1 E2 enhancer selectively regulates HO-1 gene induction by modulating Pol II binding. Nucleic Acids Res 42:13599–13614
17. Whyte WA, Orlando DA, Hnisz D et al (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153:307–319
18. Rao S, Zhen S, Roumiantsev S et al (2010) Differential roles of Sall4 isoforms in embryonic stem cell pluripotency. Mol Cell Biol 30:5364–5380
19. Kim J, Cantor AB, Orkin SH, Wang J (2009) Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells. Nat Protoc 4:506–517
20. Nelson JD, Denisenko O, Bomsztyk K (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat Protoc 1:179–185
21. Broad (2010) Broad ChIP protocol for full REMC (6 marks). http://www.roadmapepigenomics.org/protocols/type/experimental/. Accessed 19 July 2015

22. Das PP, Shao Z, Beyaz S et al (2014) Distinct and combinatorial functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in mouse embryonic stem cell identity. Mol Cell 53:32–48
23. Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995
24. Babraham Bioinformatics (2015) FastQC. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc. Accessed 19 July 2015
25. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
26. Hannon Lab Cold Spring Harbor (2015) FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed 19 July 2015
27. 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:R25
28. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079
29. Li H (2011) Statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993
30. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
31. Ji H, Jiang H, Ma W et al (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26:1293–1300
32. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
33. Landt S, Marinov GK, Kundaje A et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831

Chapter 9

Global Run-On Sequencing (GRO-Seq)

Alessandro Gardini

Abstract

Transcription occurring at gene loci results in accumulation of mature RNA molecules (i.e., mRNAs) that can be easily assayed by RT-PCR or RNA sequencing. However, the steady-state level of RNA does not accurately mirror transcriptional activity per se. In fact, RNA stability plays a major role in determining the relative abundance of any given RNA molecule. Here, I describe a protocol of Nuclear Run-On assay coupled to deep sequencing to assess real-time transcription from engaged RNA polymerase. Mapping nascent transcripts at the genome-wide scale provides a reliable measure of transcriptional activity in mammalian cells and delivers a high-resolution map of coding and noncoding transcripts that is especially useful for annotation and quantification of short-lived RNA molecules.

Key words Nuclear run-on, Nascent RNA, RNA polymerase, Transcription initiation, Transcription elongation, Noncoding RNA, Enhancer RNA, Deep sequencing

1 Introduction

The Nuclear Run-On Assay was introduced over 40 years ago as a method to assess RNA polymerase that is transcriptionally engaged [1, 2]. Nuclei from mammalian cells are isolated, washed to remove free nucleotides, and kept at ice-cold temperature to arrest ongoing transcription. Transcription is resumed in vitro when nuclei are incubated at 30 °C in the presence of radiolabeled nucleotides and the anionic detergent sarkosyl, which prevents *de novo* assembly of the pre-initiation complex and avoids re-initiation. Transcripts that were initiated at the time of nuclei isolation (commonly referred to as *nascent RNAs*) will be further elongated by engaged RNA polymerase, to allow incorporation of radioactive nucleotides. Traditionally, radiolabeled RNA is hybridized to an array of specific DNA probes (representing different genes or different portions of a given gene) that are cross-linked to a nylon membrane using a dot blot system. The extent of nascent transcription is ultimately revealed by autoradiography. Nuclear Run-On has been successfully

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_9, © Springer Science+Business Media New York 2017

performed on a variety of cell types and organisms, including plants [3], D. Melanogaster [4], and fission yeast [5].

Recently, Lis et al. developed a modified Run-On protocol to isolate nascent RNA that can be ultimately converted into a DNA library suitable for deep sequencing [6]. Such high-throughput evolution of the Run-On assay has been named GRO-seq (Global Run-On sequencing) and allows unbiased mapping of nascent transcripts genome-wide. The main alteration to the original Run-On assay lies in the use of brominated nucleotides instead of radioactive analogs. RNA molecules that have incorporated Br-UTP can be affinity purified by means of commonly used antibodies against bromodeoxyuridine (anti-BrdU). Such immunoprecipitation step is fundamental to ensure proper enrichment of nascent RNA before initiating library preparation. Illumina-compatible DNA libraries are prepared similar to conventional protocols for directional sequencing of total RNA.

Major limitations of GRO-seq are the laboriousness of the technique and the amount of starting material (the number of cells that are required lies in the 10^7 range). Nonetheless, GRO-seq is an exceptionally sensitive method to estimate transcriptional activity throughout the entire genome and has generated crucial information on RNA polymerase II (RNAPII) density at different classes of protein coding genes [6]. GRO-seq has shown unprecedented accuracy to ascertain defects in RNAPII elongation and pause release [7–9] as well as termination [10]. Additionally, GRO-seq has revealed that RNAPII fires bidirectionally at most mammalian promoters [6], initiating noncoding RNAs that are transcribed antisense with respect to the messenger RNA. Owing to their instability, these transcripts do not accumulate in the nucleus and elude most RNA detection protocols. Due to its sensitivity, GRO-seq is suitable to assess low-abundant long noncoding RNAs, such as the recently characterized enhancer-associated RNAs (eRNAs). Bidirectional eRNAs are hard to detect by conventional sequencing methods that gauge steady-state transcription. GRO-seq has been employed to reveal the full extent of eRNA transcription in response to stimuli such as estrogen, LPS, and Epidermal Growth Factor [10–12].

2 Materials

2.1 Nuclei Isolation

1. Swelling Buffer (500 ml): 492.5 ml MilliQ water, 5 ml 1 M Tris–HCl pH 7.5 (final 10 mM), 1 ml 1 M $MgCl_2$ (final 2 mM), 1.5 ml 1 M $CaCl_2$ (final 3 mM).
2. Swelling buffer with glycerol: 90 ml of swelling buffer, 10 ml of pure glycerol.
3. Lysis buffer: 99 ml of swelling buffer with glycerol, 1 ml of Igepal (NP-40).

4. Freezing buffer: 27.5 ml MilliQ water, 20 ml glycerol, 2.5 ml Tris–HCl pH 8, 250 μl of 1 M $MgCl_2$, 10 μl of 0.5 M EDTA.
5. Ice-cold PBS.
6. SUPERase In RNase Inhibitor (20 U/μl), ThermoFisher Scientific (AM2696).

2.2 Nuclear Run-On

1. 2× Nuclear Run-On (NRO) buffer: 10 mM Tris–HCl pH 8, 5 mM $MgCl_2$, 300 mM KCl, 1 mM DTT, 500 μM ATP, 500 μM GTP, 500 μM Br-UTP, 2uM CTP, 200 U/ml Superase In, 1% *N*-Laurylsarcosine (sodium salt solution), nuclease-free water. Prepare 100 μl per sample.
2. TRIzol LS reagent.
3. Chloroform, molecular biology grade.
4. 100% Ethanol and 75% ethanol in nuclease-free water.
5. Sodium chloride solution, 5 M.
6. GlycoBlue coprecipitant.

2.3 RNA Immunoprecipitation

1. RNAse-free DNAse.
2. RNA fragmentation reagents.
3. Micro Bio-Spin P-30 Gel Columns.
4. T4 Polynucleotide Kinase (PNK), 10× PNK buffer.
5. Binding buffer: 0.25× SSPE, 0.05% Tween 20, 37.5 mM NaCl, 1 mM EDTA in nuclease-free water.
6. Blocking buffer: 1× binding buffer with the addition of 0.1% polyvinylpyrrolidone and 0.1% BSA (use Ultrapure BSA in solution).
7. Low-salt wash buffer: 0.2× SSPE, 0.05% Tween 20 (50 μl 10% tween), 1 mM EDTA in nuclease-free water.
8. High-salt wash buffer: 0.2× SSPE, 137.5 mM NaCl, 0.05% Tween 20, 1 mM EDTA in nuclease-free buffer.
9. Elution buffer: 50 mM Tris pH 7.5, 150 mM NaCl, 0.1% SDS, 20 mM DTT, 1 mM EDTA in nuclease-free water.
10. TE + Tween buffer: TE (0.01 M Tris, 0.001 M EDTA, pH 7.4) + 0.05% Tween 20.
11. Anti-BrdU-conjugated agarose beads (Santa Cruz Biotechnologies).
12. 100% Ethanol and 75% ethanol in nuclease-free water.
13. Glycogen solution (20 mg/ml), molecular biology grade.

2.4 Library Preparation

1. *E. Coli* Poly(A) Polymerase (New England Biolabs), which contains the enzyme, 10× reaction buffer and 10 mM Adenosine-5′-Triphosphate (ATP). Prepare a 2.5× dilution of ATP in nuclease-free water.

2. Superscript III Reverse Transcriptase kit (ThermoFisher Scientific).
3. Oligonucleotides: NTI223—/5Phos/GAT CGT CGG ACT GTA GAA CTC T/idSp/CA AGC AGA AGA CGG CAT ACG ATT TTT TTT TTT TTT TTT TTT VN where *5Phos* indicates 5′ phosphorylation, *idSp* indicates the 1′,2′-Dideoxyribose modification that introduces a stable abasic site, and *VN* indicates degenerate nucleotides.

   NTI200—CAA GCA GAA GAC GGC ATA.

   NTI201—AAT GAT ACG GCG ACC ACC GAC AGG TTC AGA GTT CTA CAG TCC GAC G.

   NTI202—CGACAGGTTCAGAGTTCTACAGT CCGACGATC.
4. Exonuclease I.
5. NaOH 1 M and HCl 2 M.
6. 10% Polyacrylamide pre-cast gel (TBE-Urea for denaturing ssDNA PAGE and TBE for size selection of dsDNA library).
7. Nondenaturing DNA sample buffer (2× concentrated).
8. Nonmutagenic DNA stain (i.e., SYBR Gold, ThermoFisher Scientific).
9. Blue-light transilluminator.
10. Elution Buffer I: TE (0.01 M Tris, 0.001 M EDTA, pH 7.4) with 0.1% Tween 20.
11. 100% Ethanol and 75% ethanol in nuclease-free water.
12. Ultrafree MC-HV columns.
13. Betaine 5 M.
14. Circular Ligase kit.
15. Human apurinic/apyrimidinic (AP) endonuclease, APE 1.
16. 4× Relinearization mix: 100 mM KCl, 2 mM DTT in nuclease-free water.
17. Proofreading DNA polymerase (i.e., Phusion Polymerase with High-Fidelity 5× buffer).
18. dNTP mix (dATP, dCTG, dTTP, dGTP, 10 mM each).
19. 6× Gel Loading buffer for DNA.
20. PCR purification columns for microcentrifuge.
21. Corning Costar Spin-X centrifuge tube filters (Sigma-Aldrich).
22. Elution Buffer II: TE (0.01 M Tris, 0.001 M EDTA, pH 7.4) with 0.1% Tween 20 and 150 mM NaCl.
23. 100% Ethanol and 75% ethanol in nuclease-free water.
24. Glycogen solution (20 mg/ml), molecular biology grade.

3 Methods

3.1 Nuclei Isolation

1. Collect between 10^7 and 10^8 cells (accurate cell count to be performed during **step 8**) in a 50 ml polypropylene tube and immediately put on ice (*see* **Note 1**). Centrifuge at 4 °C at 1,000 RPM.
2. Wash cell pellet twice with ice-cold PBS.
3. Resuspend in 10 ml of ice cols swelling buffer, incubate on ice for 5 min.
4. Centrifuge at 400 × *g* for 10 min.
5. Discard supernatant and gently resuspend cells in 10 ml of swelling buffer with glycerol. Add Superase In to the buffer (2 U/ml).
6. Slowly add 10 ml of lysis buffer while agitating the tube (*see* **Note 2**), incubate on ice for 5 min. Add Superase In to the buffer (2 U/ml).
7. Add 25 ml of lysis buffer and centrifuge at 600 × *g* for 5 min.
8. Discard supernatant and resuspend nuclei in 10 ml of freezing buffer (add Superase In), transfer to a 15 ml tube. Use 5–10 µl of cell suspension to count nuclei with a Neubauer chamber (no intact cells should be visible at this stage).
9. Centrifuge at 900 × *g* for 6 min, discard supernatant, and resuspend in an appropriate amount of freezing buffer (10 µl per 1×10^6 of nuclei). Proceed to the Nuclear Run-On reaction or store nuclei at −80 °C. If properly stored, frozen nuclei can be used several months after collection.

3.2 Nuclear Run-On

1. Use 100 µl of frozen nuclei preparation that correspond to approximately 1×10^7 nuclei (*see* **Note 3**).
2. Prepare individual 1.5 ml tubes with 100 µl of 2× NRO buffer, add 100 µl of nuclei with immediate mixing. Pipette gently several times and incubate at 30 °C for 7 min. During this step, brominated UTP is incorporated in the nascent RNA.
3. Block the reaction by adding 600 µl of TRIzol LS reagent. Vortex thoroughly until nuclei dissolves. Incubate at room temperature for 5 min.
4. Add 160 µl of chloroform, shake vigorously for 20 s. Incubate at room temperature for 3 min. Centrifuge at 4 °C (12,000 × *g*) for 30 min.
5. Transfer the aqueous phase (upper layer) to a clean 1.5 ml tube. Add NaCl (up to 300 mM), 1 µl of GlycoBlue and 1 ml of cold 100% ethanol. Incubate at −20 °C for at least 2 h (can be done overnight).
6. Precipitate RNA at 4 °C in a microcentrifuge (maximum speed) for 30 min.

7. Discard supernatant and wash with 75 % ethanol, centrifuge for 5 min at maximum speed.
8. Carefully remove supernatant and air-dry the RNA pellet (do not let over-dry, otherwise the pellet may become difficult to resuspend).
9. Resuspend RNA in 20 μl of DNAse- and RNAse-free water supplemented with Superase In (1 U/μl).

3.3 RNA Immunoprecipitation

Part A: DNAse Treatment

1. Incubate resuspended RNA at 60 °C for 10 min.
2. Add 2.8 μl 10× DNAse buffer and 3 μl of DNase (from Turbo DNA-free kit). Incubate at 37 °C for 30 min.
3. Mix well the DNAse inactivation reagent (from Turbo DNA-free kit) and add 2 μl to the reaction. Incubate at room temperature for 5 min. Mix 2–3 times during incubation by flicking the tube.
4. Centrifuge at 10,000 × *g* for 2 min and carefully transfer the supernatant to a clean tube. Repeat DNAse treatment. Add 3 μl of DNAse and incubate 20–30 min at 37 °C.
5. Clean up with 2 μl of DNAse inactivation reagent, perform centrifugation as before, and transfer the supernatant to a clean 1.5 ml tube. Proceed to fragmentation.

Part B: RNA Fragmentation and PNK Treatment

6. For 30 μl of DNAse-free RNA solution, add 24 μl of nuclease-free water.
7. Add 6 μl fragmentation reagent (from Ambion RNA fragmentation kit) to each sample. Split the reaction into 0.2 ml PCR tubes (10 μl per tube).
8. Incubate at 70 °C for 8 min. Add 1 μl of Stop solution (from Ambion RNA fragmentation kit) to each 10 μl reaction and keep on ice.
9. Pool reactions (approximate total volume: 60 μl) and purify using Micro Bio-Spin P-30 Gel Columns. Prepare columns as per manufacturer's instruction, apply sample on top of the resin, and centrifuge at 4000 × *g* for 4 min, and then proceed to the polynucleotide kinase reaction.
10. Add fresh Superase In to the sample (2 μl of inhibitor for 50–55 μl of sample). Set up a reaction with 30 U of T4 Polynucleotide Kinase (PNK) and 1× PNK buffer. Incubate at 37 °C for 60 min.
11. Add 20 U of PNK and EDTA (final 10 mM) and incubate for additional 60 min at 37 °C. Inactivate the enzyme at 75 °C for 5 min.

Part C: Immunoprecipitation

12. Bring samples to 200 µl with binding buffer (*see* **Note 4**).
13. Prepare beads for the RNA immunoprecipitation. Use 50 µl of bead slurry per sample. During the washing and blocking procedure, beads can be pooled (up to 5 samples or 250 µl of bead slurry). Precipitate beads by centrifugation at 900 × *g* for 3 min, wash beads twice with blocking buffer, and resuspend them in 5 volumes of blocking buffer (1 volume refers to the dry bead pellet). Rotate at room temperature for 60 min.
14. Wash beads twice with 5 volumes of binding buffer, aliquot beads in individual tubes for immunoprecipitation. Beads are finally resuspended in 500 µl of binding buffer (for each 50 µl of original bead slurry).
15. Add 200 µl of RNA sample to 500 µl of resuspended beads and incubate at room temperature for 60 min in a rotisserie-style tube rotator.
16. Precipitate beads at 900 g for 3 min, discard the supernatant, and perform the following washes: 2× with 500 µl of binding buffer, 2× with 500 µl of low salt buffer, 1× with 500 µl of high salt buffer, 2× with 500 µl of TE + tween buffer. All washes are performed at room temperature, with rotation, for 2–3 min followed by centrifugation at 900 × *g* for 2 min.
17. Elute brominated nascent RNA from beads using 100 µl of elution buffer during 10 min incubation (*see* **Note 5**). Repeat the elution three additional times and collect the resulting 400 µl eluate in a clean 1.5 ml tube.
18. Purified nascent RNA by ethanol precipitation: add 1 µl glycogen, 300 mM NaCl and 1 ml of cold 100% ethanol. Incubate at −20 °C for 2 h or overnight. Centrifuge tubes at maximum speed for 30 min wash once with 75% ethanol, air-dry, and resuspend in 5 µl of nuclease-free water (containing 0.05% Tween 20 and 1 U/µl of Superase In).

3.4 Library Preparation

1. Before cDNA synthesis and adapter ligation, immunopurified RNA fragments are subjected to a poly-adenosine tailing reaction. Addition of a 3′ poly-A stretch allows first-strand cDNA synthesis to be performed with the NTI223 library adaptor, which contains a poly-dT stretch for priming. To each RNA sample, add the following: 0.7 µl of poly(A) polymerase buffer, 0.25 µl of diluted ATP, 0.7 µl of poly(A) polymerase (3.75 U). Incubate reactions at 37 °C for 30 min.
2. Prepare the reverse transcriptase reaction by adding 0.9 µl of dNTP mix (10 mM each, included in the RT kit) and 0.9 µl of NTI223 oligo. Heat the sample at 75 °C for 3 min, chill on ice. Add 1.7 µl of 10× RT buffer (provided with the kit), 3 µl of 25 mM $MgCl_2$, 1.7 µl of 0.1 M DTT, 0.5 µl of Superase In, 1 µl of SuperScript III. Incubate at 48 °C for 40 min in a thermal cycler.

3. The excess of NTI223 oligonucleotide is removed by digestion with 3 μl of exonuclease I (20 U/μl). Samples are incubated at 37 °C for 15 min.
4. Add 2 μl of NaOH 1 M and incubate at 98 °C for 20 min to allow enzyme inactivation and degradation of the RNA strand. Neutralize with 1 μl of HCl 2 M.
5. Run samples on a denaturing 10% TBE-Urea polyacrylamide gel. Prior to electrophoresis, samples should be denatured at 70 °C for 3 min, chilled on ice, and then resuspended in an equal volume of nondenaturing sample buffer. Load samples along with an appropriate ladder (*see* **Note 6**) and run at 180 V for approximately 1 h. Stain gel with a nonmutagenic dye and visualize using a blue-light transilluminator. Excise DNA fragments from 100 to 400 bp and elute in Elution buffer I (TE with 0.1% Tween 20). Perform elution at room temperature in a rotisserie-style tube rotator (*see* **Note 7**). Use a Millipore Ultrafree MC-HV column (as per manufacturer's instructions) to discard gel debris and recover the eluate, which contains single-stranded cDNA.
6. Precipitate DNA by adding 1 μl glycogen solution, 30 μl of NaCl 5 M, and 1 ml of cold 100% ethanol. Incubate at –20 °C for 2 h or overnight. Centrifuge at maximum speed for 30 min, discard the supernatant, and wash pellet once with cold 75% ethanol. Air-dry pellet and resuspend in 7.5 μl of nuclease-free water.
7. The NTI223 adaptor bears both 5′ and 3′ library adaptors, separated by a cleavable spacer. The adaptor is ligated to the free 3′ end of ssDNA with a circularization reaction, followed by cleavage of the abasic spacer in NTI223 oligo to relinearize cDNA. Circularization is performed by adding the following reagents to the 7.5 μl sample: 1 μl of CircLigase 10× reaction buffer, 0.5 μl of 1 mM ATP, 0.5 μl of 50 mM $MnCl_2$, 0.4 μl of CircLigase (100 U/μl). Incubate at 60 °C for 60 min, followed by 10 min at 80 °C to inactivate the enzyme.
8. Relinearization of cDNA is achieved by adding 3.3 μl of 4× relinearization mix and 1 μl of APE1 phosphodiesterase (15 U) followed by incubation at 37 °C for 45 min. Repeat the reaction by adding 1 μl of fresh APE1 and incubate at 37 °C for additional 45 min. Inactivate the enzyme at 65 °C for 20 min. During this reaction, the spacer in NTI223 is cleaved to generate linear cDNA inserts with 5′ and 3′ adapters that will be used to amplify the library by conventional PCR.
9. Dilute samples 2× with nuclease-free water (approximate final volume of cDNA: 30 μl) and perform amplification in a thermal cycler using proofreading DNA polymerase and betaine to prevent formation of secondary structures. Add the following reagents to the PCR reaction: 500 nm NTI200, 500 nm

NTI201, 10 μl of 5× High-Fidelity buffer, 200 μM of dNTPs, 5 μl of betaine, 1 U of Phusion DNA polymerase. Perform reactions in a final volume of 50 μl using no more than 5 μl of cDNA as template (set up multiple reactions per sample). Use the following PCR routine: 30 s at 98 °C for initial denaturation, 20 cycles of (10 s at 98 °C, 30 s at 57 °C, 15 s at 72 °C), 5 min at 72 °C as final elongation step.

10. Pool PCR reactions from the same sample and concentrate using a PCR purification column (alternatively, perform ethanol precipitation). Elute in 25–30 μl of nuclease-free water (or a proprietary elution buffer), mix with 6× Gel Loading Buffer.
11. Pre-run a 10% polyacrylamide TBE gel for 15 min. Run samples for 2 h at 120 V, along with an appropriate low molecular weight DNA ladder. Excise fragments comprised between 150 and 300 bp.
12. Elute the library using 400 μl of Elution buffer II (TE with 0.1% Tween 20 and 150 mM NaCl), rotate at room temperature for 4 h. Transfer eluate and gel debris to a Spin-X filter column and centrifuge at 14,000 × *g* for 2 min. Add 300 mM NaCl, 1 μl of glycogen and ethanol to precipitate DNA. Incubate overnight at −20 °C. Centrifuge at maximum speed for 30 min, wash once with 75% ethanol, and resuspend in 10–20 μl of TE (*see* **Note 8**).
13. Libraries can be clustered using Illumina platforms and sequenced with oligo NTI202. Before sequencing, samples should be evaluated on a BioAnalyzer (Agilent) using the High Sensitivity DNA analysis kit. Optimal size range for GRO-seq libraries is 200–250 bp, which is shorter than conventional RNA-seq libraries. Use KAPA library quantification kit to measure the exact concentration (expected range: 100–200 nM).

4 Notes

1. Adherent cells should be trypsinized as fast as possible and collected using ice-cold medium to ensure that transcription is arrested. Alternatively, culture dishes can be placed over an ice bed to collect cells with a disposable scraper.
2. Gentle agitation is needed while the detergent solution (lysis buffer) is added to the swollen cell preparation to disrupt the cytoplasmic membrane. Avoid excessive foaming that would cause the nuclear membrane to break. This step can be performed with the aid of a vortex benchtop mixer using mild rotation settings.
3. Place frozen nuclei on ice for approximately 5 min before use. Unused nuclei can be frozen again and stored at −80 °C.
4. All buffers used during immunoprecipitation should be supplemented with RNAse inhibitor right before use (add Superase

In to a final concentration of 4 U/ml). Buffers should be kept cold to preserve the inhibitor, however, all binding and washing steps can be performed at room temperature.

5. Elution is best performed at 37 °C with continuous shaking to keep beads in suspension. Better results are achieved using a tube mixer/incubator such as the Thermomixer (Eppendorf).
6. An appropriate ladder for single-stranded nucleic acids should be loaded on the denaturing gel. Use of single stranded RNA marker with a low molecular range (i.e., 1000–100 bp) is recommended. Denature RNA ladder as per manufacturer's instructions.
7. To maximize recovery of ssDNA from the elution, cut gel bands into small pieces using a sharp blade.
8. Library yield after precipitation can be evaluated by NanoDrop (Thermo Scientific). Concentration should lie between 30 and 100 ng/μl.

References

1. Gariglio P, Buss J, Green MH (1974) Sarkosyl activation of RNA polymerase activity in mitotic mouse cells. FEBS Lett 44:330–333
2. Gariglio P, Bellard M, Chambon P (1981) Clustering of RNA polymerase B molecules in the 5′ moiety of the adult beta-globin gene of hen erythrocytes. Nucleic Acids Res 9:2589–2598
3. Gatehouse J, Thompson AJ (1995) Nuclear "run-on" transcription assays. Methods Mol Biol 49:229–238
4. Rougvie AE, Lis JT (1988) The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell 54:795–804
5. Birse CE, Lee BA, Hansen K, Proudfoot NJ (1997) Transcriptional termination signals for RNA polymerase II in fission yeast. EMBO J 16:3633–3643
6. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848
7. Gardini A, Baillat D, Cesaroni M et al (2014) Integrator regulates transcriptional initiation and pause release following activation. Mol Cell 56:128–139
8. Chen FX, Woodfin AR, Gardini A et al (2015) PAF1, a molecular regulator of promoter-proximal pausing by RNA polymerase II. Cell 162:1003–1015
9. Saponaro M, Kantidakis T, Mitter R et al (2014) RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. Cell 157:1037–1049
10. Lai F, Gardini A, Zhang A, Shiekhattar R (2015) Integrator mediates the biogenesis of enhancer RNAs. Nature 525:399–403
11. Lam MT, Cho H, Lesch HP et al (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature 498:511–515
12. Li W, Notani D, Ma Q et al (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498:516–520

Chapter 10

Computational Approaches for Mining GRO-Seq Data to Identify and Characterize Active Enhancers

Anusha Nagari, Shino Murakami, Venkat S. Malladi, and W. Lee Kraus

Abstract

Transcriptional enhancers are DNA regulatory elements that are bound by transcription factors and act to positively regulate the expression of nearby or distally located target genes. Enhancers have many features that have been discovered using genomic analyses. Recent studies have shown that active enhancers recruit RNA polymerase II (Pol II) and are transcribed, producing enhancer RNAs (eRNAs). GRO-seq, a method for identifying the location and orientation of all actively transcribing RNA polymerases across the genome, is a powerful approach for monitoring nascent enhancer transcription. Furthermore, the unique pattern of enhancer transcription can be used to identify enhancers in the absence of any information about the underlying transcription factors. Here, we describe the computational approaches required to identify and analyze active enhancers using GRO-seq data, including data pre-processing, alignment, and transcript calling. In addition, we describe protocols and computational pipelines for mining GRO-seq data to identify active enhancers, as well as known transcription factor binding sites that are transcribed. Furthermore, we discuss approaches for integrating GRO-seq-based enhancer data with other genomic data, including target gene expression and function. Finally, we describe molecular biology assays that can be used to confirm and explore further the function of enhancers that have been identified using genomic assays. Together, these approaches should allow the user to identify and explore the features and biological functions of new cell type-specific enhancers.

Keys words GRO-seq, groHMM, Enhancer, Enhancer RNAs (eRNAs), Enhancer prediction, Gene regulation, Looping, Motif, Motif search, Promoter, Response element, Transcription, Transcription factor, Transcription unit

1 Introduction

1.1 Transcriptional Enhancers Function as Genomic Regulatory Elements

Transcriptional enhancers (enhancers) are DNA regulatory elements that are bound by transcription factors (TFs) and act to positively regulate the expression of nearby or distally located target genes [1, 2]. Enhancers are located throughout the genome, including promoters, gene bodies, and intergenic regions, and they function independent of their orientation and location with respect to their target genes [3–5]. They also function in a cell type-specific manner; an enhancer that is active in one cell type

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_10, © Springer Science+Business Media New York 2017

might not be in another [1, 6]. By controlling unique patterns of gene expression in different cell types, enhancers drive the unique biology of those cells types. Thus, identifying the repertoire of enhancers that are active in a given cell type, the set of target genes regulated by those enhancers, and the molecular mechanisms controlling enhancer function provide important clues for understanding biological outcomes.

1.2 Properties and Features of Active Enhancers

TF binding to a specific locus in the genome does not necessarily lead to the formation of an "active" enhancer (i.e., an enhancer that can drive the transcription of a target gene by RNA polymerase II, Pol II). In fact, TF binding events that fail to promote the formation of an active enhancer have been observed for a variety of transcription factors [7–9]. Active enhancers exhibit unique properties and features, many of which have been defined using deep sequencing-based genomic assays. These assays include:

Chromatin immunoprecipitation-sequencing (ChIP-seq), which determines the enrichment of TFs, chromatin- and transcription-related factors, and posttranslational modifications of histones across the genome [10].

Deoxyribonuclease digestion-sequencing (DNase-seq) and assay for transposase-accessible chromatin-sequencing (ATAC-seq), which determine the "openness" or accessibility of chromatin at specific loci across the genome [11–13].

Deep sequencing-based chromosome conformation capture (3C)-related assays (e.g., Hi-C), which monitor the formation of chromatin loops across the genome [14–16].

Global run-on-sequencing (GRO-seq) and related assays, which detect the location of active RNA polymerases and the production of nascent transcripts across the genome [17, 18]. These assays have been used to identify common features shared by active enhancers (Fig. 1).

Properties and features of active enhancers include (1) binding of one or more TFs to DNA sequence motifs specific for those TFs, (2) enhanced chromatin accessibility, (3) enrichment of specific histone modifications, including histone H3 lysine 4 mono/dimethylation (H3K4me1/me2) and H3 lysine 27 acetylation (H3K27ac), (4) binding of transcriptional coactivators, histone-modifying enzymes, and chromatin-modulating enzymes (e.g., the protein acetyltransferases p300 and CBP; the multipolypeptide Mediator complex), (5) recruitment of Pol II and active transcription of nascent enhancer RNAs (eRNAs) [19, 20], and (6) looping to target gene promoters [15, 21] (Fig. 1). While some of the features noted above are also shared with promoters, such as enrichment of coregulators and Pol II, others are more enriched at enhancers than promoters (e.g., H3K4me1/me2) [1, 3, 4, 22]. Although these enhancer features have been known for some time, how they contribute to the regulation and function of enhancers remains to be determined.

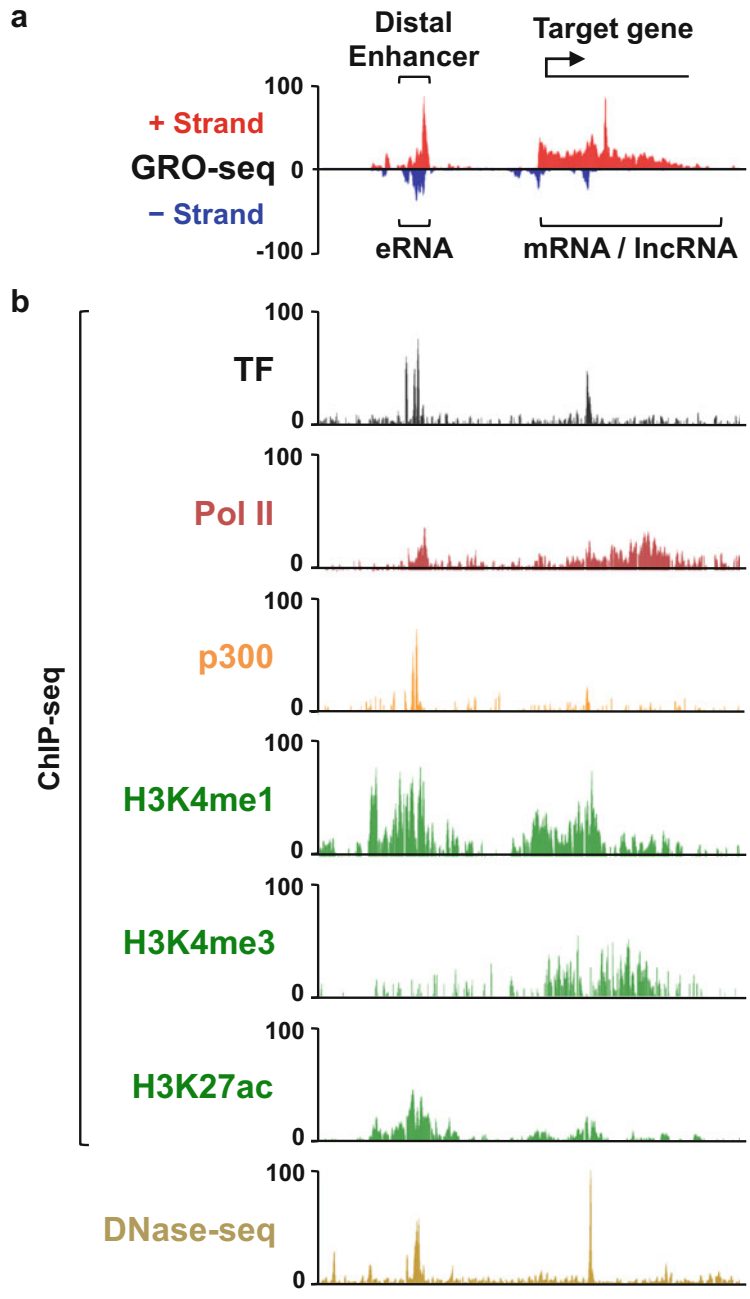


Fig. 1 Genomic features of active enhancers and promoters. Genome browser tracks showing (**a**) GRO-seq and (**b**) ChIP-seq and DNase-seq data at a representative locus of the human genome. Bidirectional transcription at the enhancer is evident, as is TF and p300 binding, recruitment of Pol II, and enrichment of histone modifications

1.3 Identifying and Characterizing Enhancer Transcripts

Active transcription at enhancers was first observed over a decade ago in locus-specific molecular biology experiments [23–25]. These observations were extended by the initial observation using ChIP-seq that Pol II is recruited to enhancers across the genome [22]. Subsequent studies using total RNA-seq in neurons and macrophages demonstrated that the Pol II bound at enhancers is indeed engaged in active transcription, producing short, bidirectional, noncoding transcripts called enhancer RNAs (eRNAs) [19, 20]. These studies also showed that the production of eRNAs correlates with the recruitment of transcription factors in response to neuron and macrophage activation [19, 20]. The genome-wide identification of transcription start sites in intergenic regions using TSS-seq and CAGE technology added further support for enhancer transcription [19, 26]. Taken together, these studies provide strong evidence for enhancer transcription as a general biological event.

Additional studies aimed at understanding signal-dependent transcriptional responses have used GRO-seq, a method for identifying the location and orientation of actively transcribing Pol II (and Pol I and Pol III) across the genome, to characterize signal-dependent transcription at enhancers [7, 8, 18, 27–29]. GRO-seq has been used to distinguish between TF binding sites (e.g., for estrogen receptor alpha, ERα, and NF-kB) that produce transcripts and those that do not [7, 8]. Only the former (i.e., TF binding sites that are transcribed) are enriched for genomic features associated with active enhancers (e.g., H3K4me1, DNaseI accessibility, p300/CBP binding) [7, 8]. In more recent studies, derivatives of GRO-seq (i.e., GRO-cap or 5′ GRO-seq), which enrich for 5′-capped nascent transcripts, have been used to study enhancer transcription [27, 28]. Collectively, these studies have shown that GRO-seq is an effective means to identify, characterize, and understand the regulation of enhancer transcription. Furthermore, these studies have shown that enhancer transcription is an early event in enhancer activation after TFs binding (which, of course, may require the prior binding of pioneer factors and chromatin remodeling). As such, enhancer transcription, as detected by GRO-seq, is a highly reliable mark of active enhancers, which can be exploited to identify and study these enhancers. In fact, it may be the most robust indicator of enhancer activity, even more so than the histone modifications typically enriched at enhancers [7, 20].

1.4 Using GRO-Seq and Related Approaches to Identify and Study Active Enhancers

GRO-seq and related approaches, such as PRO-seq [30], GRO-cap [27], and 5′ GRO-seq [28], are powerful techniques to identify actively transcribed regions of the genome, whether or not those regions have been annotated previously. As we describe below, GRO-seq data can be mined to identify active enhancers in an unbiased way in the absence of any prior information about the initiating TF. In addition, once enhancers are identified, they can be mined using bioinformatic approaches to identify putative

underlying TF motifs. In addition, the GRO-seq data can be integrated with other types of genomic data relating to enhancer function (e.g., ChIP-seq for TFs and histone modifications, DNase-seq, looping data; see for example [7, 31].

Recently, software has been developed to analyze GRO-seq (and related) data to search for enhancers and other regulatory elements. For example, groHMM, a software package in the R programing language that is available in Bioconductor [32], uses a two-state Hidden Markov Model to define the boundaries of transcription units. Using groHMM, one can identify actively transcribed regions of the genome from GRO-seq data. Furthermore, dREG (discriminative regulatory-element detection from GRO-seq), a computer program that uses read counts to employ support vector regression, can be used to identify active transcriptional regulatory elements from GRO-seq or PRO-seq data [33].

2 Materials: Computer, Data, and Software

Herein, we describe the use of computational tools, approaches, and pipelines to identify and characterize cell type-specific enhancers using GRO-seq and other genomic data. For executing these analyses, you will need a source of GRO-seq data, a suitable computer, and a variety of software.

1. A high-capacity computer suitable for analyzing high content, high complexity data sets.
2. GRO-seq data from a cell or tissue type of interest.
3. Additional genomic data for integration and comparison, as desired.
4. R, a programming language and software environment for statistical computing and graphics (www.r-project.org/).
5. Perl, a high-level, general-purpose, interpreted, dynamic programming language (https://www.perl.org).
6. Cutadapt, a Python module used to remove adapter sequences from high-throughput sequencing data (http://cutadapt.readthedocs.org/en/stable/index.html) [34], used here to trim the polyA tail and adapter sequences from GRO-seq reads.
7. Burrows-Wheeler aligner (BWA), a software package for mapping low-divergent sequences against a large reference genome (http://bio-bwa.sourceforge.net) [35].
8. groHMM, an R package from Bioconductor for analyzing GRO-seq data (http://www.bioconductor.org/packages/release/bioc/html/groHMM.html) [32].
9. Bedtools, a suite of computational tools for a wide-range of genomic analysis tasks (http://bedtools.readthedocs.org/en/latest/) [36].

10. Python, a general-purpose, high-level programming language (https://www.python.org/).
11. SAMtools, a set of utilities that manipulate alignments in the BAM format (http://samtools.sourceforge.net/) [37].

3 Methods

3.1 Processing and Aligning GRO-Seq Data

The following are a standard set of computational approaches that can be used to process GRO-seq data. The analytical steps involved include: (1) quality control analysis of the GRO-seq data, (2) pre-processing of the GRO-seq data depending on the information from the quality control analysis to improve the usability of the dataset, and (3) aligning the processed GRO-seq reads to a reference genome (mapping) to associate the signals with specific genomic locations. These steps are performed using a variety of open-source software, some of which have user-friendly graphical user interfaces, while others require the use of command lines. Below, we have provided commands that can be cut and pasted into the command line versions of the software noted.

3.1.1 Quality Control and Trimming the Adapter and polyA Sequences from the GRO-Seq Reads

Quality control is an important first step in processing high-throughput sequencing data, including GRO-seq. The GRO-seq data should be checked for contamination from the sequencing adapters or the polyA addition (pre-processing). Quality control analysis can be performed using tools like FastQC, a quality control tool for raw high-throughput sequencing data [38] (Fig. 2). In order to improve the alignment of reads to the reference genome for the species in which you are working, adapter and polyA trimming should be performed (Fig. 2). The adapter and polyA sequences should be trimmed from the GRO-seq reads to increase the fraction of reads that can be aligned to the reference genome. This can be done using various publicly available trimming tools, such as Cutadapt and Trimmomatic [39].

Here we show how adapter and polyA sequences can be trimmed using Cutadapt. Only reads which are >32 bp in length (--minimum-length) after adapter trimming are retained for further analysis. A default maximum error rate (-e) of 0.1 is used. In order to comply with the input format necessary for futher steps, all negative quality values are changed to zero (-z). The statistics regarding the reads that are trimmed in this step are redirected to an output statistics file.

The following example can be executed in the command line version of Cutadapt to trim adapter and polyA sequence contamination resulting from the GRO-seq protocol. An implementation of the commands in Bash scripts is available through the GitHub repository (see below). Trimming of the adapter sequence (1, below) should be sequentially followed by the execution of trimming polyA tail (2, below).

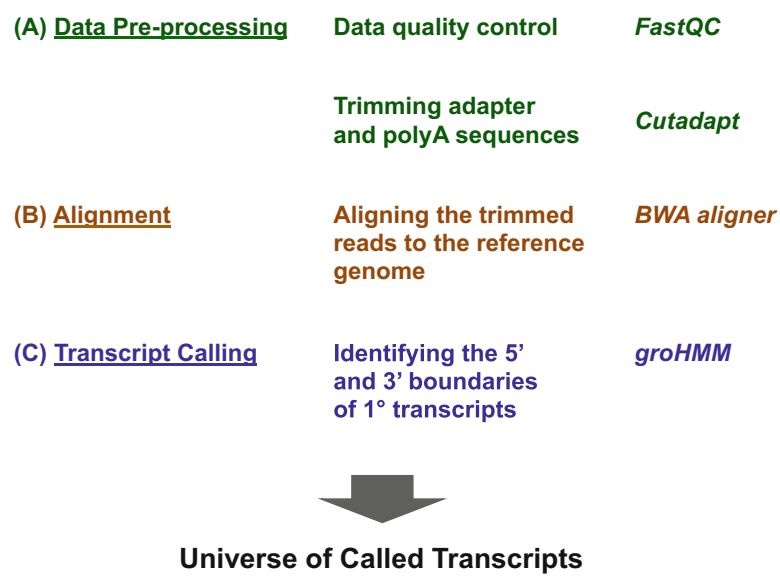


Fig. 2 Preprocessing, alignment, and transcript calling for GRO-seq data. Overview of GRO-seq data analysis, as well as software that can be used for the key steps in the analysis

(1) Trimming adapter sequence: GRO-seq data in the fastq format is provided as input for this step.

```
$ cutadapt -a <adapter sequence> -z -e 0.10
--minimum-length=32 --output=filename.noA-
dapt.fastq.gz inputfile.fastq.gz 2>&1 >>
RunCutadapt.out
```

(2) Trimming polyA tail: After trimming the adapter sequence, the output file from the above step (reads trimmed for adapter sequence) is now processed in this step to trim the polyA contamination.

```
$ cutadapt -a AAAAAAAAAAAAAAAAAAAA -z -e 0.10
--minimum-length=32 --output= filename.noPolyA.
noAdapt.fastq.gz filename.noAdapt.fastq.gz 2>&1
>> RunCutadapt.out
```

3.1.2 Aligning the Trimmed GRO-Seq Reads to the Reference Genome

After trimming the sequencing reads, the data should be aligned to the appropriate reference genome to provide the map of the sites of active transcription across the genome. The alignment can be accomplished using publicly available software, such as BWA [35] and SOAP [40] (Fig. 2).

Here we show the alignment of trimmed reads using the BWA aligner. We find that it works better for handling the unequal read lengths that are produced after the pre-processing step. A maximum of two mismatches (-n) and a subsequence seed length of

32 bp (-l) are used as parameters for alignment in this step. The "samse" command will produce an output with a maximum of one alignment per read (-n). After alignment the files containing the aligned reads will have to be in a specific format (i.e., bam, -b) to perform subsequent transcript calling and tuning using the groHMM package.

The following examples can be executed in the command line version of the BWA aligner, followed by conversion to the bam format using "SAMtools [37]." An implementation of the commands in a single Bash script is available from the GitHub repository (see below).

Aligning to the reference genome index: The output from Cutadapt after adapter and polyA trimming ('filename.noPolyA.noAdapt.fastq.gz') is provided as input to the BWA aligner. The final reads passing these criteria are aligned to the reference genome and are written to the 'alignedFile.sam' file.

```
$ bwa aln -n 2 -l 32 -t 8 Genome_INDEX.fa filename.noPolyA.noAdapt.fastq.gz > alignedFile.sai

$ bwa samse Genome_INDEX.fa -n 1 alignedFile.sai inputfile.fastq.gz > alignedFile.sam
```

Converting aligned files from sam to bam format using SAMtools.

```
$ samtools view -bh -S alignedFile.sam > alignedFile.unsorted.bam

$ samtools sort alignedFile.unsorted.bam alignedFile.sorted.bam
```

Note that a typical GRO-seq experiment has two or more replicates for each experimental condition. Hence, it is important to test that the replicates are highly correlated (Fig. 7).

3.2 Analyzing GRO-Seq Data Using groHMM and Other Computational Tools

GroHMM is a software package in R that can be used to define the boundaries of transcription units from a GRO-seq data using a two-state Hidden Markov Model (HMM) [32]. It also provides additional tools for visualizing and analyzing GRO-seq data. The groHMM package covers basic steps of GRO-seq data analysis, including the generation of wiggle files using the "writeWiggle" function and the creation of metagene (data average) plots using the "runMetaGene" function, as well as more advanced steps, such as predicting the boundaries of actively transcribed regions (transcription units) across the genome de novo (Fig. 2).

The aligned files from Subheading 3.1.2 serve as the input to groHMM. Since GRO-seq data is strand-specific, one can visualize the signals from the plus and minus strands separately. The pipelines

for calling transcription units (using "detectTranscripts"), as well as evaluating (using "evaluateHMMInAnnotations") and tuning the transcript calling, are explained in detail in the tutorial associated with the groHMM package [32]. In a systematic comparison of the performance of groHMM versus other transcription unit callers, such as SICER and HOMER [41, 42], groHMM performed better with respect to coverage of genic and intergenic regions, as well as transcription unit accuracy for both short and long transcripts [32].

3.3 Identification of Active Enhancers from GRO-Seq Data

Transcription from GRO-seq data can be used as a signature to identify active enhancers (here, by "active enhancer," we mean those that are actively transcribed) [7, 33, 43]. This can be accomplished using two approaches: (1) de novo identification of active enhancers using short bidirectional transcript pairs and (2) identification of TF binding sites (from ChIP-seq data) that are actively transcribed. For the de novo identification, bioinformatic approaches can be used to identify motifs for putative transcription factors that drive the formation of those enhancers [7]. In the sections below, we describe how active enhancers can be identified using groHMM, open-source software, and additional scripts in the R and perl programming languages.

The enhancer identification pipelines described herein are implemented in Bash, Perl, and R. The most up-to-date version, with full documentation and examples, is available free of charge under an open-source MIT license via GitHub at: https://github.com/Kraus-Lab/active-enhancers. Note that the various cutoffs described below may have to be tuned for the particular biological system or the particular data set being analyzed.

3.3.1 De Novo Identification of Enhancers Using GRO-Seq Data

We have shown previously that the production of enhancer transcripts can be used to identify active enhancers de novo in the absence of any other genomic information [7]. For these analyses, we have focused on intergenic enhancers to avoid complications in the analysis associated with overlapping gene body transcription. For our purposes, we have searched >10 kb away from the 5′ or 3′ end of an annotated gene [7], although this can be adjusted to recover a greater number of enhancers or those closer to promoters [8]. We have also defined the enhancer transcripts as "short" (i.e., ≤9 kb), as well as unidirectional (i.e., transcript produced from one strand of DNA, but not the other) or bidirectional (i.e., transcript produced from both strands of DNA) [7] (Figs. 3 and 4).

The first step in this analysis is to identify intergenic transcripts from the universe of all transcripts obtained from groHMM [7, 32]. As noted above, we use a cutoff of >10 kb away from either end of annotated genes in order to distinguish enhancer transcription from genic transcription. Here, we show how a set of intergenic transcripts can be identified from a transcript universe using the "intersect" function in BEDtools, a suite of

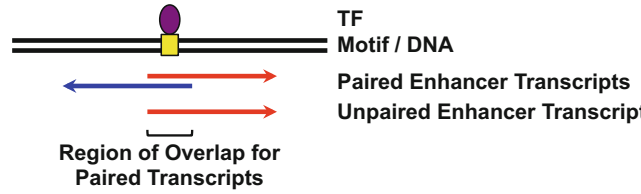


Fig. 3 Schematic representation of an actively transcribed enhancer. Actively transcribed enhancers that form at TF binding sites may produce paired or unpaired enhancer transcripts

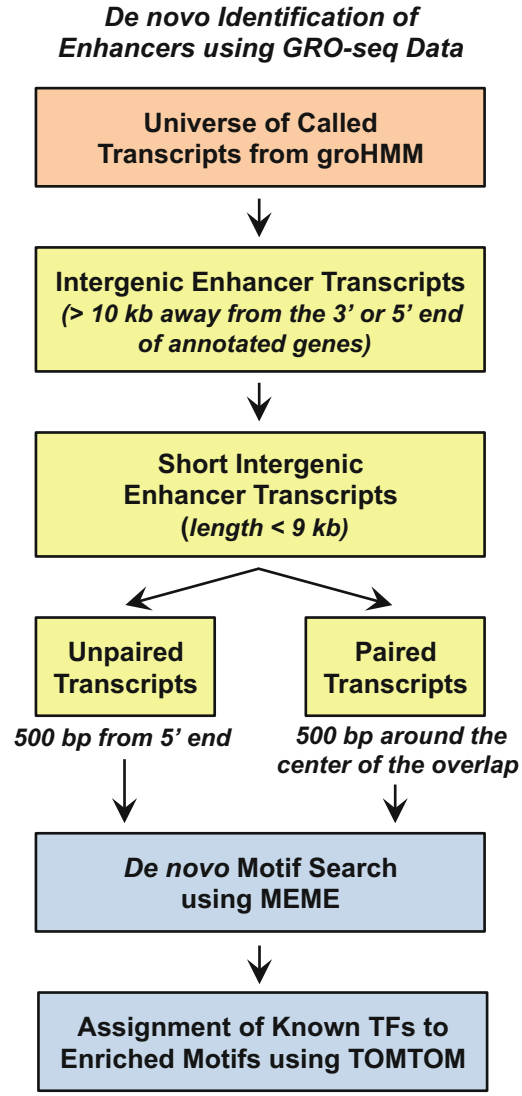


Fig. 4 De novo identification of enhancers using GRO-seq data. Details are provided in the text

different analysis tools that can be used to modify, convert, or compare bed files [36]. The following example illustrates the use of "bedtools intersect" to isolate transcripts that do not intersect (–v) with genic regions. An implementation of the command in a single Bash script is available from the GitHub repository (see below).

Identify intergenic transcripts: The "genic_regions_to_avoid.bed" file contains the genomic coordinates extending 10 kb from the 5′ and 3′ ends of annotated genes. The input files should be sorted before running the bedtools intersect function using the following unix command.

```
$ sort -k1,1 -k2,2n ip.txt ip_sorted.txt
$ bedtools intersect -a transcript_universe_
from_groHMM.txt -b genic_regions_to_avoid.bed
-v > intergenic_transcripts.txt
```

After filtering for transcripts that are intergenic, we use a length cutoff to define and identify enhancer transcripts (Fig. 4). In a previous study, we observed that the median length of transcripts originating from distal ERα enhancers in MCF-7 breast cancer cells is ~9 kb [7]. Hence, we use 9 kb as the length cutoff to define "short" eRNA transcription units and hypothesize that longer transcripts originating from the enhancers are more likely to be bona fide long non-coding RNAs (lncRNAs) [7, 44]. As noted above, enhancer transcription can be unidirectional or bidirectional, depending on the nature of the enhancer. Furthermore, the magnitude of enhancer transcription may correlate directly with the activity of the enhancer [7]. A comparison of active enhancers (with robust uni- or bidirectional transcription) with "inactive" enhancers, as well as their associated genomic features, suggests that it is informative to distinguish these different categories of enhancers [7].

The provided Perl script can be used to identify short intergenic transcripts (i.e., putative enhancer transcripts) and then divide them into short paired (bidirectional) enhancer transcripts. The transcripts remaining in the universe of short intergenic transcripts are considered to be "short unpaired transcripts" [7]. The Perl code is available for download from the GitHub repository (https://github.com/Kraus-Lab/active-enhancers/blob/master/scripts/Define_enhancer_transcripts.pl). It will produce an output of short paired intergenic transcripts together with information about the overlap of the transcript pair.

Identify short intergenic transcripts: The output from bedtools intersect after identifiying intergenic transcripts (intergenic_transcripts.txt) is provided as an input to the Perl script. The final transcripts passing these criteria are written to the "paired_transcripts.txt" file, along with length of overlap "paired_transcripts_overlap.txt" and coordinates of a 1 kb window around the center of the overlap "paired_transcripts_1kb_window_overlap".

```
$ ./Define_enhancer_transcripts.pl -i intergen-
ic_transcripts.txt
-a short_paired_transcripts.txt –b short_
paired_transcripts_overlap.txt –c short_
paired_transcripts_1kb_window_overlap.txt
```

3.3.2 Identification of Known TF Binding Sites That Are Actively Transcribed Using GRO-Seq Data

GRO-seq data can be used to identify known TF binding sites (from ChIP-seq data) that are actively transcribed. This can be accomplished in two ways: (1) by comparing the overlap of transcripts in the universe of transcripts from groHMM with known TF binding sites of interest or (2) by collecting and quantifying the GRO-seq reads that fall within a specified window around known TF binding sites of interest (Fig. 5). With respect to the former, criteria for the location of the TF binding site relative to the cognate enhancer transcript(s) (or vice versa) can be specified. For example, if the focus is on paired/bidirectional enhancer transcripts, one might specify that the TF binding site must be located within the region of overlap of the + strand and – strand transcripts [7].

Pipelines for the global identification of enhancer transcripts associated with known TF binding sites using ERα as an example has been described previously [7]. The analysis is similar to the one described in 3.3.1. However, in this case, the starting point is a set of known TF binding sites, rather than a set of known enhancer transcripts. As described above, the first step is to define intergenic TF binding sites and then search for those that overlap with an enhancer transcript to identify active intergenic enhancers.

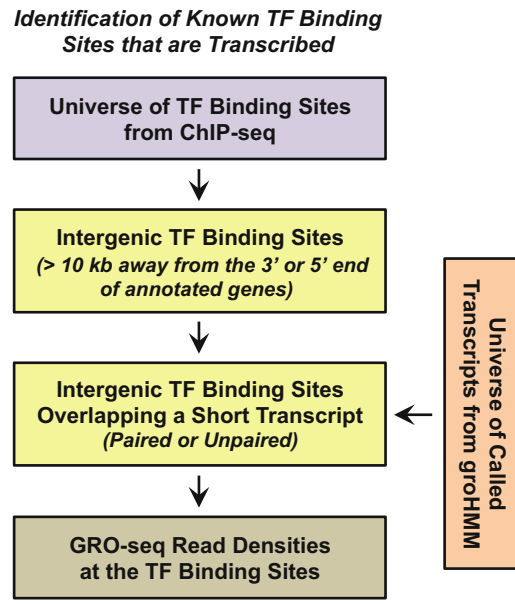


Fig. 5 Identification of known TF binding sites that are transcribed. Details are provided in the text

3.4 Associating Newly Identified Enhancers with TF Motifs

After completing the pipeline for de novo identification of active enhancers using GRO-seq data, as in Subheading 3.3.1 above, one can search in the transcribed region for an enrichment of motifs that suggest putative TFs that may drive the formation of those enhancers [7]. In our analyses, we have focused on (1) a region (e.g., 500 bp) surrounding the center of the overlap between the enhancer transcript pairs for bidirectional/paired enhancer transcripts or (2) a window (e.g., 500 bp) at the 5′ end of unidirectional/unpaired enhancer transcripts (Figs. 3 and 4). The sequences of the genomic regions specified above are extracted from the UCSC genome browser.

Within the regions specified above, motifs for putative TFs can be identified in two ways: (1) a directed approach using software, such as FIMO [45] or MotifScanner [46], which searches for enrichment of known, user-provided TF motifs in the region of interest and (2) a de novo approach using software, such as MEME [47], which searches for the enrichment of specific DNA sequences that can then be matched to known TF motifs using software, such as STAMP [48] or TOMTOM [49]. Motif searches in genomic regions where enhancer transcripts originate, such as those described here, can help in uncovering the TFs that mediated the formation and activity of the enhancers of interest.

3.5 Associating Newly Identified Enhancers with Putative Target Genes

How an enhancer targets and promotes the transcription of its target genes is a fundamental question in gene regulation biology. Such analyses can be readily performed by using a "nearest-neighboring gene" approach. In this approach, the actively transcribed gene (e.g., mRNA gene or lncRNA gene) nearest to an enhancer is assumed to be a target of the enhancer (Fig. 6). While not perfect, this assumption holds well enough to be informative with respect to enhancer function and target gene activation [7, 31]. Alternatively, if genome-wide looping data are available for a particular TF (e.g., from ChIA-PET analyses [16, 50, 51]), then direct associations between enhancers and target genes can be discerned. In either case, the relationship between enhancer transcription and target gene transcription can be determined from GRO-seq data. Furthermore, potential biological functions of a set of enhancers identified using GRO-seq data can be explored by gene ontology (GO) or pathway analyses of the target gene set [31]. Such analyses can reveal the likely biological functions of the target genes and, by extension, the likely biological functions of the enhancers as well (Fig. 6).

3.6 Identifying Cell Type-Specific Enhancers Using GRO-Seq Data

The profiles of enhancer transcripts are highly cell type-specific [32], more so than the profiles of other genomic enhancer data. This cell-type specificity can be used to discern important biological insights. The groHMM-based enhancer identification pipelines described above can be used to identify cell type-specific enhancers

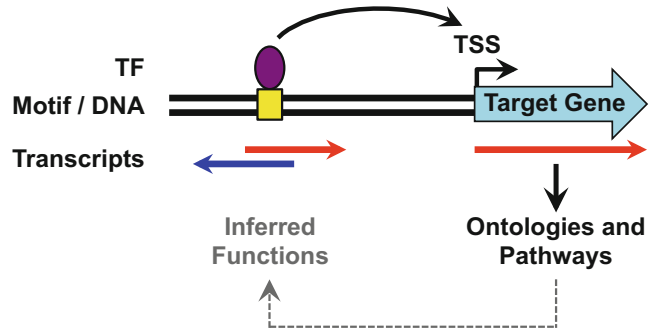


Fig. 6 Analysis of target gene activation and functions. Active enhancers may promote the transcription of nearby genes through looping mechanisms that bring the enhancers and target gene promoters in proximity. Knowledge of the functions of the target genes from ontology analyses can provide clues about the biological functions of the enhancers

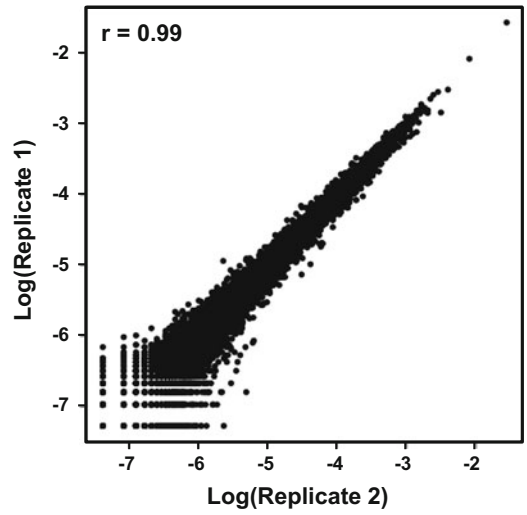


Fig. 7 Correlation plot of two biological replicates of GRO-seq data. A typical GRO-seq experiment has two or more replicates for each experimental condition. Hence, it is important to test that the replicates are highly correlated. Shown here is a Pearson's correlation plot

by comparing GRO-seq data derived from different cell types. Using an approach similar to the one described in Subheading 3.3 above, one can identify the universe of enhancer transcripts expressed in a particular cell type and then compare that universe to the universes of enhancer transcripts expressed in other cell types. These comparisons allow for the identification of enhancer transcripts that (1) are common across various cell types or (2) are unique to a particular cell type. Motif analysis, as described in

Subheading 3.4 above, can be performed for the enhancers producing common or unique transcripts to identify putative TFs that might drive the formation of those enhancers.

For the analysis described in this section, which involves the comparison of multiple GRO-seq datasets to identify cell type-specific enhancers, the library sizes of all the samples should be compared. Appropriate normalization steps should be used to avoid bias due to differences in sequencing depth.

3.7 Integration with Other Genomic Data and Other Bioinformatic Analyses

After identifying the set of active enhancers in a particular cell type, the enhancer information from the GRO-seq data, which includes the genomic location and the magnitude of transcription, can be integrated with data from other genomic approaches. For example, the enrichment of enhancer-related histone modifications (e.g., H3K4me1, H3K27ac) and TF binding from ChIP-seq data or the chromatin state from DNase-seq can be assessed at the GRO-seq-called enhancers (Fig. 1).

As noted above, nearest-neighboring gene analyses can be used to identify putative target genes of the predicted enhancers with subsequent GO and pathway analyses on the potential target genes. The GO and pathway analyses can be performed using tools such as WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) [52] and DAVID [53]. Such analyses can provide insights about the biological functions of GRO-seq-identified enhancers. These "functional" analyses can be facilitated by using GREAT (Genomic Regions Enrichment of Annotations Tool), which assigns biological meaning to a set of noncoding genomic regions by analyzing the annotations of the nearby genes [54]. Users can provide GRO-seq-defined enhancer locations as input in the GREAT web interface and select the "Single nearest gene" option in the association rule settings.

Custom multidimensional analyses can be used to explore the relationships among multiple enhancer-related parameters. For example, we have recently demonstrated how enhancer transcription (from GRO-seq), target gene transcription (from GRO-seq), and TF binding at the predicted enhancer (from ChIP-seq) increase simultaneously in response to an external signal, an observation that can be visualized in a three-dimensional box plot [31]. Of course, the additional analyses described here represent a few of the many ways in which GRO-seq and other genomic data can be mined to explore enhancer functions.

3.8 Validation of Genomic Results Using Enhancer-Specific Molecular Biology Techniques

All of the specific conclusions regarding enhancer formation and function derived from the genomic analyses described here should be validated for individual enhancers using molecular biology approaches. Enhancer features can be tested in locus-specific assays that assess (1) enhancer transcription (e.g., by reverse transcription-qPCR), (2) binding of TFs and enrichment of histone modifications (e.g., by ChIP-qPCR), (3)

chromatin accessibility (e.g., by DNase-qPCR), and (4) looping (e.g., by 3C-qPCR) [7]. The function of the enhancers identified by GRO-seq can be tested in reporter gene assays, where the DNA sequence from an identified enhancer is inserted into a reporter construct. Upon introduction of the enhancer-reporter construct into cells expressing the cognate TF, the presence of the enhancer DNA element should increase reporter activity if it is a functional enhancer [55].

In addition, the function of putative TFs driving the formation of enhancers identified using GRO-seq can be tested in functional assays. For example, the TF should bind to the enhancer (as determined by ChIP-qPCR) and RNA-mediated knockdown of the TF should abolish enhancer formation and function (e.g., loss of enhancer transcription and a reduction of enhancer-associated histone modifications). Furthermore, the functions of GRO-seq-identified enhancers can be tested using enhancer deletion assays in cells, in which the enhancer DNA is deleted (or mutated) using CRISPR/Cas9 and the impairment of enhancer function and target gene transcription is assessed using the qPCR-based locus-specific assays described above. Ultimately, the function of each enhancer identified and examined in detail should be tested using genetic models in vivo [56].

Acknowledgments

The authors thank Minho Chae and Hector L. Franco for helpful comments and suggestions about enhancer identification using GRO-seq, as well as this manuscript. The enhancer-related work in the Kraus lab is supported by grants from the NIH/NIDDK and the Cancer Prevention and Research Institute of Texas (CPRIT).

References

1. Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15:272–286
2. Wamstad JA, Wang X, Demuren OO et al (2014) Distal enhancers: new insights into heart development and disease. Trends Cell Biol 24:294–302
3. Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12:283–293
4. Pennacchio LA, Bickmore W, Dean A et al (2013) Enhancers: five essential questions. Nat Rev Genet 14:288–295
5. Spitz F, Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13:613–626
6. Heinz S, Romanoski CE, Benner C et al (2015) The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol 16:144–154
7. Hah N, Murakami S, Nagari A et al (2013) Enhancer transcripts mark active estrogen receptor binding sites. Genome Res 23:1210–1223
8. Luo X, Chae M, Krishnakumar R et al (2014) Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFalpha signaling revealed by integrated genomic analyses. BMC Genomics 15:155

9. Savic D, Roberts BS, Carleton JB et al (2015) Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/enhancer-binding protein beta binding sites. Genome Res 25(12):1791–1800
10. 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
11. Buenrostro JD, Giresi PG, Zaba LC et al (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10:1213–1218
12. Flores O, Deniz O, Soler-Lopez M et al (2014) Fuzziness and noise in nucleosomal architecture. Nucleic Acids Res 42:4934–4946
13. Song L, Zhang Z, Grasfeder LL et al (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21:1757–1767
14. Carter D, Chakalova L, Osborne CS et al (2002) Long-range chromatin regulatory interactions in vivo. Nat Genet 32:623–626
15. Dekker J, Rippe K, Dekker M et al (2002) Capturing chromosome conformation. Science 295:1306–1311
16. Fullwood MJ, Liu MH, Pan YF et al (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462:58–64
17. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848
18. Hah N, Danko CG, Core L et al (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell 145:622–634
19. De Santa F, Barozzi I, Mietton F et al (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8:e1000384
20. Kim TK, Hemberg M, Gray JM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
21. Wang Q, Carroll JS, Brown M (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol Cell 19:631–642
22. 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
23. Ling J, Baibakov B, Pi W et al (2005) The HS2 enhancer of the beta-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a cis-linked globin promoter. J Mol Biol 350:883–896
24. Spicuglia S, Kumar S, Yeh JH et al (2002) Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes. Mol Cell 10:1479–1487
25. Vieira KF, Levings PP, Hill MA et al (2004) Recruitment of transcription complexes to the beta-globin gene locus in vivo and in vitro. J Biol Chem 279:50350–50357
26. Yamashita R, Sathira NP, Kanai A et al (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res 21:775–789
27. Core LJ, Martins AL, Danko CG et al (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet 46:1311–1320
28. Lam MT, Cho H, Lesch HP et al (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature 498:511–515
29. Wang D, Garcia-Bassets I, Benner C et al (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474:390–394
30. Kwak H, Fuda NJ, Core LJ et al (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339:950–953
31. Franco HL, Nagari A, Kraus WL (2015) TNFalpha signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. Mol Cell 58:21–34
32. Chae M, Danko CG, Kraus WL (2015) groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. BMC Bioinformatics 16:222
33. Danko CG, Hyland SL, Core LJ et al (2015) Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods 12:433–438
34. Martin M (2012) Cutadapt removes adapter sequences from high-throughput sequencing reads. Bioinformatics Action 17(1):10–12, Key: citeulike:11851772 17:10-12
35. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
36. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
37. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

38. Andrews S. (2010) Fastqc. A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
39. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
40. Li R, Li Y, Kristiansen K et al (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714
41. Heinz S, Benner C, Spann N et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576–589
42. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 25:1952–1958. doi: 10.1093/bioinformatics/btp340
43. Fang B, Everett LJ, Jager J et al (2014) Circadian enhancers coordinate multiple phases of rhythmic gene transcription in vivo. Cell 159:1140–1152
44. Sun M, Gadad SS, Kim DS et al (2015) Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. Mol Cell 59:698–711
45. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017–1018
46. Aerts S, Thijs G, Coessens B et al (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res 31:1753–1764
47. Bailey TL, Boden M, Buske FA et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208
48. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 35:W253–W258
49. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol 8(2):R24
50. Handoko L, Xu H, Li G et al (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 43:630–638
51. 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
52. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33:W741–W748
53. Huang DW, Sherman BT, Tan Q et al (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8:R183
54. Mclean CY, Bristor D, Hiller M et al (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28:495–501
55. Heldring N, Isaacs GD, Diehl AG et al (2011) Multiple sequence-specific DNA-binding proteins mediate estrogen receptor signaling through a tethering pathway. Mol Endocrinol 25:564–574
56. Meyer MB, Benkusky NA, Onal M et al (2015) Selective regulation of Mmp13 by 1,25(OH) D, PTH, and Osterix through distal enhancers. J Steroid Biochem Mol Biol

Chapter 11

Evaluating the Stability of mRNAs and Noncoding RNAs

Ana Carolina Ayupe and Eduardo M. Reis

Abstract

Changes in RNA stability have an important impact in the gene expression regulation. Different methods based on the transcription blockage with RNA polymerase inhibitors or metabolic labeling of newly synthesized RNAs have been developed to evaluate RNA decay rates in cultured cell. Combined with techniques to measure transcript abundance genome-wide, these methods have been used to reveal novel features of the eukaryotic transcriptome. The stability of protein-coding mRNAs is in general closely associated to the physiological function of their encoded proteins, with short-lived mRNAs being significantly enriched among regulatory genes whereas genes associated with housekeeping functions are predominantly stable. Likewise, the stability of noncoding RNAs (ncRNAs) seems to reflect their functional role in the cell. Thus, investigating RNA stability can provide insights regarding the function of yet uncharacterized regulatory ncRNAs. In this chapter, we discuss the methodologies currently used to estimate RNA decay and outline an experimental protocol for genome-wide estimation of RNA stability of protein-coding and lncRNAs. This protocol details the transcriptional blockage of cultured cells with actinomycin D, followed by RNA isolation at different time points, the determination of transcript abundance by qPCR/DNA oligoarray hybridization, and the calculation of individual transcript half-lives.

Key words RNA stability, RNA half-life, Transcription inhibitors, Actinomycin D, Pulse labeling with uridine analogs, Noncoding RNAs

1 Introduction

Steady-state RNA levels in the cell are regulated by the balance between DNA transcription and posttranscriptional events, such as RNA processing, splicing, editing, transport, and degradation. Alterations in the gene expression regulation are intimately related to changes in RNA stability. Several external stimuli and signals, such as hormone activation and infection with pathogens can affect the decay rate of specific mRNAs [1–3]. Likewise, different cellular processes such as cell cycle progression, cell differentiation, and embryonic development are modulated by changes in RNA stability [2, 3].

Different methods have been used to estimate RNA stability of individual transcripts or genome-wide [2, 4–13] (*see* Table 1). Among those, the use of transcriptional inhibitors such as

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_11, © Springer Science+Business Media New York 2017

Table 1

Overview of methods to estimate RNA stability

Method	Advantages	Limitations	References
Transcription inhibitors	Simple and cheap method	May be toxic to the cells; can interfere with the determination of transcript half-lives	[14, 15]
Actinomycin D (5–10 μg/ml)	Fast uptake (minutes)	Poor selectivity; inhibits all RNA polymerases. Stock solution in DMSO	[4, 15]
α-amanitin (50 μg/ml for RNP II inhibition and 250/ml for RNP III inhibition)	Highly selective for RNAP II and RNAP III. Stock solution in water	Slow uptake (hours)	[15, 19]
DRB (100 μM)	Fast uptake (minutes)	Stock solution in DMSO. Short RNAs can escape of the transcription inhibition	[15]
Metabolic pulse labeling with uridine analogs	Less toxic to cells; can be used to measure stability under a variety of physiological or stress conditions	Methods of purification is more laborious and expensive	[14, 29]
4sU (200–500 μM)	Fast uptake (1–2 h)	Inhibition of cell growth in prolonged culturing. Can cause mismatch in the RNA sequence	[6, 10, 14]
EU (0.1–0.5 mM)	Fast uptake in high doses (30–60 min or 1–24 h, depending of the concentration)	Inhibition of cell growth in prolonged culturing	[8, 14]

actinomycinD(ActD),α-amanitin,and5,6-dichloro-1-d-ribofuranosyl--benzimidazole (DRB) have been the most widely employed [14]. Actinomycin D preferentially intercalates into GC-rich DNA sequences and inhibits the transcription progression of all eukaryotic RNA polymerases (RNAP) [15]. Alpha-amanitin binds to RNAPs II and III preventing the nucleotide incorporation and translocation of the nascent transcript [16, 17]. RNAP II is 100-fold more sensitive to α-amanitin than RNAP III [18] and a number of RNAP III transcribed genes apparently inhibited by α-amanitin are in fact indirectly regulated by direct α-amanitin inhibition of RNAP II-transcribed RNAP III regulatory proteins [19]. The adenosine analog DRB is a protein kinase inhibitor that

affects the transition between early and productive RNAP elongation, inhibiting the production of long RNAs but did not affect the transcription of short RNA [20]. Differently from actinomycin D and α-amanitin that directly target DNA templates and RNAPs, respectively, DRB affects transcriptional factors (e.g., CDK9, DSIF, NELF) that associate with RNAP II [20]. DRB also affects transcriptional activity of RNAP I by blocking early rRNA processing steps [21].

Following transcription inhibition with one of the drugs mentioned above, RNA decay may be estimated by monitoring the reduction in the abundance of individual or thousands of transcripts over a time course using quantitative methods such as Real-Time PCR (qPCR) or DNA microarray/RNA deep sequencing (RNAseq) assays, respectively [2, 4, 5, 9, 12, 13, 15].

By combining the blockage of global transcription with transcription inhibitors and subsequently monitoring ongoing RNA decay over time using DNA microarray technology, earliest studies estimated the stability of a large numbers of genes simultaneously and correlated their RNA decay rates with specific biological functions [22–24]. Despite simple, widely used, and frequently considered the standard method to determine RNA decay rates [6, 7, 25], transcription inhibitors are toxic and may cause negative impacts on the cellular physiology, which admittedly may affect the precise determination of the RNA decay rates [14, 15] (*see* Table 1).

Alternative strategies to measure RNA stability are based in the metabolic pulse labeling of cellular transcripts with uridine analogs, such as 4-thiouridine (4sU), 5-ethynyluridine (EU), and 5-bromouridine (BrU) [6–8, 10, 11]. Uridine analogs are added to the cell culture medium to be incorporated into newly synthesized RNAs that are collected in one unique time point [6, 7, 10] or at sequential time points after removal of metabolic label [8, 11]. The newly synthesized labeled RNAs can be purified from the total RNA and quantitated using qPCR, DNA microarrays, or RNAseq [14] (*see* Table 1). In single time experiments, the transcript half-lives can be calculated based on the determination of two ratios: newly transcribed (uridine labeled) RNA/total RNA and pre-existing (unlabeled) RNA/total RNA, considering the duration of labeling [6, 7, 10]. This indirect estimation of RNA decay, as well the yield of purification of RNAs containing uridine analogs represent limitations of this approach, which can cause discrepancy in the half-life measurements [14]. A pulse with uridine analog followed by collection at sequential time points is more suitable to measure RNA decay. In this case, the half-life is estimated by determining the decrease in metabolically labeled RNA over time [14]. In general, uridine analogs are less toxic to the cells; however, prolonged culture with 4sU or EU may cause cell growth inhibition [11]. BrU has less harmful effects compared with 4sU or EU, but its RNA labeling kinetics is slower [14]. As a potential disadvantage of the 4sU-labeling method, it

allows base-pairing with guanine instead of adenine, which may compromise the correct estimation of transcript abundance based on RNAseq [14].

Studies comparing RNA half-lives estimated using actinomycin D transcription arrest or metabolic pulse labeling with uridine analogs have showed good agreement [1, 6, 7, 10, 25]. In general, these studies reported better correlations between the two methods for short-lived transcripts, which may be explained by the short duration of the actinomycin D treatment. Of note, it has been reported that to measure accurate transcript half-lives using transcription inhibitors, especially of long-lived transcripts, the duration of time course following inhibition should be greater than 6 h [5].

Genome-wide studies of RNA stability have contributed with novel information regarding the association between RNA half-lives and the functional pathways exerted by their encoded products in the cell [1, 2, 4–7, 10, 11]. It has been noted that subsets of genes with certain combinations of mRNA and protein stability are enriched for specific biological processes [10]. Unstable mRNAs encoding short-lived proteins that require fast induction and repression are usually associated with regulatory functions such as transcription factors signal transduction components, chromatin-modifying enzymes, cytokines, and oncogenes [1, 2, 4–7, 10, 11]. On the other hand, stable mRNAs encoding stable proteins are more frequently associated with constitutive cellular processes such as protein translation, cell respiration, and central metabolism, which do not require dynamic regulation for their proper function [1, 2, 4–7, 10, 11].

Similarly to protein-coding mRNAs, it has been proposed that the stability of ncRNAs may reflect their functional role in the cell; ncRNAs involved in housekeeping functions such as transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), small Cajal body-specific RNAs (SCARNAs) have long half-lives [11]. Conversely, some regulatory long (>200 nt in length) ncRNAs (lncRNAs), such as CDKN2B-AS1, HOTAIR, TUG1, and GAS5, display shorter half-lives [11]. In the last decade, it has become apparent that the eukaryotic genomes encode a plethora of ncRNAs that include intergenic lncRNAs (lincRNAs), enhancer-associated RNAs (eRNAs), antisense and intronic lncRNAs [26]. It is plausible that investigation of RNA stability can be an useful tool to provide clues regarding the biology of these poorly characterized noncoding RNAs classes. In fact, recent studies, two employing actinomycin D [4, 5] and another one using pulse labeling with BrU [11], have investigated globally the stability of lncRNAs in eukaryotic cells. These three studies reported that the stability of lncRNAs is comparable to those of mRNAs, with half-lives that vary over a wide range, comprising unstable and stable transcripts [4, 5, 11]. Interestingly, the class of intronic lncRNAs, i.e., those with the same orientation as the host mRNA, comprise a more

heterogeneous group of transcripts that can be distinguished according to their stability profiles, including short-lived ($t_{1/2} < 1$ h) splicing lariats as well as stable ($t_{1/2} > 3$ h) [4].

In summary, determining profiles of RNA stability in genome-wide scale can be a powerful tool to reveal new aspects from the biology of protein-coding and noncoding transcripts. In this chapter, we describe in detail a protocol to estimate RNA stability in eukaryotic cells following transcription inhibition with actinomycin D. It can be used to investigate RNA decay rates of transcripts individually of genome-wide. The outlined protocol includes cell plating, RNA isolation and removing of contaminating DNA, RNA analysis, reverse transcription followed by qPCR, DNA microarray hybridization, and RNA stability analysis.

2 Materials

1. Cell culture medium for HeLa cells: Dulbecco's Modified Eagle's Medium (DMEM) modified to contain l-glutamine 4 mM, glucose 4500 mg/l, sodium pyruvate 1 mM, sodium bicarbonate 1500 mg/l, and fetal bovine serum 10%. If another cell lineage is used, prepare the appropriate cell culture medium.
2. Phosphate-buffered saline (PBS: pH 7.4: NaCl 137.93 mM, KCl 2,67 mM, Na_2HPO_4 8.06 mM e KH_2PO_4 1.47 mM.
3. Actinomycin D.
4. Dimethyl sulfoxide (DMSO).
5. TRIzol.
6. Chloroform.
7. Ethanol absolute.
8. Ethanol 75%.
9. Isopropanol.
10. DEPC-Treated Water (RNAse-free water).
11. β-Mercaptoethanol.
12. Mini RNA Isolation Kit.
13. cDNA First-Strand Synthesis Mix.
14. SYBR Green PCR Master Mix.
15. Agilent Low Input Quick Amp Labeling Kit (Agilent Technologies).
16. Custom Agilent DNA oligoarray platform (GEO GPL19372) [4] or other custom oligoarray platform of choice.
17. 15 ml conical centrifuge tube.
18. 1.5 ml ribonuclease (RNAse)-free tube.
19. 0.2 ml RNAse-free tube.

20. Optical 96-Well Reaction Plate.
21. Optical Adhesive Film.
22. CO_2 chamber.
23. Laminar flow hood.
24. Refrigerated centrifuge for 1.5 ml tubes.
25. Clinical centrifuge for 15 ml conical tubes.
26. Real-Time PCR System.
27. High-Resolution Microarray Scanner.
28. Hybridization Oven.
29. Feature Extraction software (Agilent Technologies).

3 Methods

3.1 Cell plating

Subculture the cells in the appropriate growth medium for a minimum of 24 h before transcription inhibition to ensure normal cell metabolism. Cells should be seeded in 10 cm-diameter culture dish to reach a 50–70% confluence at the start of treatment (*see* **Notes 1–3**). Prepare six replicates for each time point (3 for test and 3 for mock samples) (*see* **Note 4**).

3.2 Actinomycin D Treatment

1. Prepare an actinomycin D stock solution in a concentration of 1 mg/ml in DMSO (*see* **Note 5**).
2. Prepare enough culture medium with actinomycin D for all time points of the treatment (e.g., 0, 1, 3, 6, and 8 h). For each time point, consider 10 ml of culture medium containing 10 μg/ml actinomycin D. Add 100 μl of the 1 mg/ml stock solution to each 10 ml of medium.
3. Each treatment time point should contain negative control (mock) replicates. Prepare enough mock medium for all treatment time points. Add 100 μl of DMSO to each 10 ml of mock medium.
4. To start the treatment, drain culture medium from all 10 cm-diameter culture dishes. Wash cells with 5 ml PBS. Collect time 0 h (*see* **step 7** below). Add 10 ml of 10 μg/ml actinomycin D or mock medium to the remaining cell dishes.
5. Incubate the cells at 37 °C with 5% CO_2 (*see* **Note 6**).
6. At each time point, drain medium from treatment and mock replicate dishes and wash the cells with 5 ml of PBS. Add 1 ml of TRIzol directly on the dish and pipette up and down, and squirt it around the surface of the dish to remove adherent cells from the surface (*see* **Note 7**). Transfer the cell lysate in TRIzol to a 1.5 ml RNAse-free tube.

7. Store the lysates at −70 °C until use or proceed immediately to RNA isolation.

3.3 RNA Isolation (TRIzol Protocol)

1. Incubate the cell lysate samples for 5 min at room temperature (*see* **Note 8**).
2. Add 200 µl of chloroform to each tube and cap.
3. Agitate vigorously the tubes by inversion for 15 s.
4. Incubate 3 min at room temperature.
5. Centrifuge the tubes at 12,000 × *g* for 15 min at 4 °C.
6. Transfer aqueous upper phase to new tubes (*see* **Note 9**).
7. Add 500 µl of isopropanol per tube and shake by inversion.
8. Incubate 10 min at room temperature.
9. Centrifuge the tubes at 12,000 × *g* for 10 min at 4 °C.
10. Discard the supernatant without disturbing the RNA pellet.
11. Wash the RNA pellet with 1 ml of ethanol 75 %.
12. Invert the tubes carefully three times.
13. Centrifuge the tubes at 7500 × *g* for 5 min at 4 °C.
14. Discard the supernatant and dry the RNA pellet at room temperature for 10 min (*see* **Note 10**).
15. Resuspend the RNA pellet in 100 µl of RNAse-free water.
16. Store the RNA samples at −70 °C until use or proceed immediately to the DNAse treatment.

3.4 Removal of Contaminating DNA

This protocol is based in the illustra RNAspin Mini RNA Isolation Kit with modifications (*see* **Notes 11** and **12**).

1. Prepare binding buffer. For each sample, prepare a mix of 350 µl of buffer RA1, 250 µl of 100 % ethanol and 3.5 µl of β-mercaptoethanol per RNA sample to be treated with DNAse. Make enough mix to all samples.
2. Add 603.5 µl of the mix to 100 µl of each RNA sample.
3. Pipet up and down 2–3 times and transfer on the liquid to a RNAspin Mini Column (blue column).
4. Centrifuge for 30 s at 8000 × *g*. Transfer the column to a new collection tube.
5. Add 350 µl of Membrane Desalting Buffer (MDB).
6. Centrifuge for 1 min at 11,000 × *g*. Discard the flow-through and return the column to the collection tube (*see* **Note 13**).
7. Prepare DNase I reaction mixture. For each sample, add 10 µl reconstituted DNase I (included in the kit) to 90 µl DNase Reaction Buffer. Make enough mix to all samples and mix the solution gently by pipetting up and down (*see* **Note 14**).

8. Apply 95 µl of the DNase I reaction mixture to the center of the column.
9. Incubate for 1 h at room temperature (*see* **Note 15**).
10. Add 200 µl of Wash Buffer I to the column (*see* **Note 16**).
11. Centrifuge for 1 min at 11,000 × *g*. Place the column into a new collection tube.
12. Add 600 µl of Wash Buffer II to the column (*see* **Note 17**).
13. Centrifuge for 1 min at 11,000 × *g*. Discard flow-through and place the column back into the collection tube.
14. Add 250 µl of Wash Buffer II to the column.
15. Centrifuge for 2 min at 11,000 × *g*.
16. Place the column into a new collection tube and centrifuge for 1 min at 11,000 × *g* (*see* **Note 18**).
17. Place the column into a 1.5 ml RNAse-free tube.
18. Add 50 µl of RNase-free water to the center of the column and incubate for 1 min at room temperature.
19. Centrifuge at 11,000 × *g* for 1 min.
20. Collect the eluted RNA and reapply to the center of column (*see* **Note 19**).
21. Centrifuge at 11,000 × *g* for 1 min.
22. Keep the RNA samples on ice. Proceed to RNA quantitation or store at −70 °C until use (*see* **Note 20**).

3.5 RNA Analysis

1. Determine the RNA concentration and purity by measuring the absorbance at 260, 280, and 230 nm (*see* **Note 21**). Apply the formula A260 × dilution × 40 = µg RNA/ml to determine concentration. Store the RNA samples at −70 °C.
2. Use the 28S and 18S rRNA bands (eukaryotic samples) to estimate the integrity of the RNA. The 2:1 ratio (28S:18S) is a good indication that the RNA is not degraded (*see* **Note 22**).

3.6 Transcript Quantification by Reverse Transcription Followed by Quantitative PCR (RT-qPCR)

This step is important to validate the quality of time course samples that will be used to estimate transcript half-lives and should be executed prior to perform large-scale measurements by DNA microarray of RNAseq.

1. Prepare reverse transcription reactions (20 µl of volume) using the kit SuperScript III First-Strand Synthesis SuperMix.
2. Add 1 µl of 50 µM oligo$(dT)_{20}$ primer and 1 µl of annealing buffer in a 0.2 ml RNAse-free tube.
3. Add 1 µg of RNA correspondent to each time point of treated and control samples in a volume of up to 6 µl or bring up volume to 6 µl with RNase-free water (keep the samples on ice) (*see* **Note 23**).

4. Mix by pipetting gently up and down.
5. Heat mixture to 85 °C for 3 min and subsequently to 65 °C for 5 min (*see* **Note 24**). During this time, prepare a mix containing 10 µl of 2× First-Strand Reaction Mix and 2 µl of SuperScript III/RNAseOUT Enzyme Mix per tube. Make enough mix to all samples.
6. After the RNA denaturing step, transfer the reactions to ice for at least 1 min.
7. Collect the contents of the tube by brief centrifugation.
8. Add 12 µl of the mix above to each tube and pipette gently up and down.
9. Incubate the reactions at 50 °C for 50 min.
10. Inactivate the reactions by heating at 85 °C for 5 min.
11. Store the cDNA samples at –20 °C or carry on directly to cDNA dilution by adding 100 µl of RNAse-free water (keep the samples on ice) (*see* **Notes 25** and **26**).
12. Assemble qPCR reactions by adding 5 µl of diluted cDNA sample, 5 µl of the forward and reverse primers mix of interest in the appropriated concentration and 10 µl of Power SYBR Green PCR Master Mix in an optical 96 well plate. Make triplicate reactions for each sample. Carefully seal the plate with optical adhesive film and spin down before running the plate in the real time PCR thermocycler using default parameters.
13. Data analysis should be made using the relative quantification method (delta Ct method) [27]. At each time point, the treated sample should be normalized to the untreated mock sample. The normalized time points should be expressed relative to time 0 h, which is set to 100%.
14. Calculate the transcript decay rates as described in Subheading 3.8 below.
15. As a control, verify the efficiency of transcription inhibition by measuring the half-life of *c-Myc* and compare with the values reported in the literature (typically <1 h) [2, 4, 5, 7] (*see* **Note 27**).

3.7 Genome-Wide Transcript Quantification by DNA Oligoarray Hybridization

We describe here a two-color experimental design to detect the abundance of protein-coding and long ncRNAs using a custom Agilent oligoarray platform (GEO GPL19372) [4] and the protocol recommended by the array manufacturer (Low Input Quick Amp Labeling Kit). We note the same protocol is suitable for other Agilent oligoarrays and can be easily adapted to other array platforms.

Briefly, Cy3-labeled cRNA targets were generated by T7-poly-dT in vitro transcription using an equivalent amount of total RNA (200 ng) from each time points sample treated with actinomycin D. Similarly, Cy5-labeled targets were generated from an RNA pool

comprising equal amounts of RNA from all replicates treated with a DMSO vehicle in all the time points (except time 0 h), and used as a reference (*see* **Note 28**).

For each replicate at each time point, Cy3-labeled actinomycin-treated cRNA targets were combined with Cy5-labeled cRNA targets from the reference RNA pool and incubated with individual arrays. After 17 h incubation at 65 °C, the arrays were washed according to the Agilent SSPE wash protocol v. 2.1 and scanned with a High-Resolution Microarray Scanner. RNA abundance measurements were obtained using the Feature Extraction software (Agilent, version 9.5).

3.8 RNA Stability Analysis

1. Only oligoarray probes confidently detected in the reference sample (RNA pool from mock controls) should be considered for further analysis. A good criterion is to limit the analysis to probes deemed as detected above the background by a 2-sided *t*-test (in the case of Agilent arrays use "IsPosAndSignif" flag in Feature Extraction software output), and with an intensity signal at least twofold greater than the local background in at least 2 of the 3 replicates from each time point (*see* **Note 29**).
2. For each valid probe, calculate the intensity ratio (Cy3-actinomycin-treated sample/Cy5-reference RNA) referent to each time point (0, 1, 3, 6, and 8 h).
3. For each probe, normalize the time course expression data set by the expression level that was measured prior to treatment (0 h), which should be set to 1.
4. Identify a subset of stable transcripts, i.e., those that show apparent increase in their abundance along the time course following transcription inhibition (*see* **Note 30**). Based on this set, calculate a correction factor for each time point such that the averaged normalized expression of the stable gene is set to 1. Scale the intensity values of all transcripts at each time point based on the calculated correction factors. For each transcript, use both the "one-phase exponential decay" and "linear regression" models to fit the intensity signals values versus time and calculate the RNA half-lives (*see* **Note 31**). The model that provides the larger R^2 value should be considered. Transcripts with an R^2 less than 0.7 are not reliable and should not be considered for further analysis (*see* **Notes 32–34**).

4 Notes

1. It is very important that the cells are not confluent to ensure that they have homogeneous access to the actinomycin D containing medium during the treatment.
2. For HeLa cells, seed 5×10^5 cells per 10 cm-diameter dish 24 h before the actinomycin D treatment.

3. Spread the cells using a rapid and smooth back and forth motion. Do not spread the cells by swirling the media in a circular motion; this results in clumping of cells in the middle of the dish.
4. It is important that the time course includes the point 0 h. In order to get reliable results, at least five time points should be prepared and the extension of the treatment should be >6 h to permit a reliable estimation of half-life of long-lived transcripts. Higher the number of points greater will be the accuracy of your measure of half-life. All the points should be made considering one plate for actinomycin D treatment and other for the mock. In the case, we used the time points 0, 1, 3, 6, and 8 h to make the stability experiments.
5. Aliquots of a concentrated (1 mg/ml) actinomycin D stock solution are expected to be stable for at least a month at –20 °C and protect from the light.
6. It should be avoided opening the incubator, especially if performing short time course series.
7. For cell lines growing in suspension transfer the cell containing medium to a 15 ml conical centrifuge tube and spin for 5 min at 800 × *g*. Discard the supernatant and wash the cells with 5 ml of PBS. Resuspend cells in 1 ml of TRIzol. Lyse the cells by pipetting up and down several times.
8. This step allows complete dissociation of the nucleoprotein complex.
9. Avoiding touching in the interphase or organic layer with the pipette to reduce DNA and RNAse contamination of isolated RNA.
10. Alternatively, remove the traces of ethanol 75% using the pipette and proceed to resuspension in water. Caution not to dry the pellet completely because the RNA can lose solubility.
11. This step can be realized using other RNA purification kits with DNAse treatment or DNAse treatments outside column.
12. Contaminating DNA can interfere with RNA quantitation and subsequent gene expression analysis.
13. If the column outlet has come into contact with the flow-through, discard the flow-through and centrifuge again for 30 s at 11,000 × *g*. Efficient salt removal will make the subsequent DNase digestion more effective.
14. Note: During the DNAse reconstitution, avoid vigorous mixing of the DNase I enzyme because it is sensitive to mechanical agitation. Add the indicated volume of RNase-free water to the DNase I vial and incubate for 1 min at room temperature. Gently swirl the vial to completely dissolve the DNase I. Dispense into aliquots and store at –20 °C. The frozen working solution is stable for 6 months. Do not freeze/thaw the aliquots more than three times.

15. Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare) protocol indicates the DNAse treatment for 15 min, but we recommend extending this time to 1 h to increase the efficiency of the treatment.
16. Wash Buffer I will inactivate DNase.
17. Before the use of Wash Buffer II, add the indicated volume of ethanol absolute to the Wash Buffer II concentrate. Store Wash Buffer II at room temperature for up to 1 year.
18. This step helps to eliminate traces of ethanol in the column that could interfere in posterior reactions.
19. This step increases the yield of the RNA recovery.
20. We suggest to check the efficiency of DNAse treatment with a polymerase chain reaction (PCR) (40 cycles) using 1 μg of RNA as a template and primers to alpha-tubulin or any other gene. Extension and annealing conditions must be determined according the primers used. Use 30–50 ng of genomic DNA as a positive control and use a negative control to exclude the possibility of DNA contamination from the PCR reagents, water, and pipettes. For amplicons with up to 400 nucleotides, it is indicated running the samples in 2% agarose gels for better separation of template RNA present in the PCR reaction. If necessary, repeat the DNAse treatment of the RNA samples. Primers for human α -tubulin: α-tubulin _ F: TCAACACCTTCTTCAGTGAAACG; α-tubulin _R: AGTGCCAGTGCGAACTTCATC.
21. We suggest using NanoDrop spectrophotometer (Thermo Scientific) to check the RNA purity and concentration. Absorbance at 280 and 230 nm indicates the presence of protein and phenol contaminants, respectively. 260/280 and 260/230 ratios should be greater than 2.
22. We suggest using capillary electrophoresis in Bioanalyzer equipment (Agilent) to check the RNA integrity. An RNA Integrity Number equal or greater than 8 is recommended.
23. The amount of total RNA to be used as template to the cDNA reactions can be less than 1 μg; the manufacturer protocol suggests using 0.1 pg to 5 μg. However, considering that, in general, ncRNAs are expressed at lower levels compared to the mRNAs we indicate of at least 1 μg of total RNA per reaction. To increase the yield of cDNA, we recommend setup 2 reactions for each sample, each one with 500 ng of RNA, and combining the cDNA products.
24. The additional heating step to 85 °C (in relation to the manufacturer protocol) decreases RNA secondary structures, which is a critical step prior to carrying out the cDNA synthesis.

25. The cDNA dilution volume should be calculated according to the abundance of the transcripts of interest that will be measured in the next step. In general, ncRNAs are expressed at lower levels compared to the mRNAs. We recommend that a test is performed upfront to determine the appropriate volume of dilution for the qPCR experiment. The transcript of interest should be detectable in the diluted cDNA from time point 0 h at a cycle threshold low enough (equal or lower that Ct 28) so that it can still be reliably detected in several time points of treatment. This is necessary since it must be considered that the amount of the transcript will decrease over the sequential time points.
26. cDNA is less stable than normal dsDNA and must be always keep on ice. Avoid repeated freezing/thawing cycles.
27. Human c-*Myc*_F: TCAAGAGGTGCCACGTCTCC and human c-*Myc*_R: TCTTGGCAGCAGGATAGTCCTT
28. Reference design avoids the effects of dye bias in the results, eliminating the necessity of dye-swap. All the subsets of samples can be compared because all samples are hybridized with the same labeled reference [28].
29. Although this filter may exclude transcripts expressed at lower levels, we recommend its use to avoid the misinterpretation of microarray measurements since small changes in probe intensity after transcription blocking may be artifactually considered as evidence of high stability [2, 4].
30. Since the same amount of total RNA from each time point used in the hybridizations, the abundance of the most stable RNAs appears to increase over the time. The calculation of a correction factor to account for those genes is used to correct for biases introduced by the apparent relative increase in transcript abundance due to the overall decrease in RNA mass along the actinomycin D treatment [2, 4, 5].
31. Use the GraphPad Prism software version 5.04 or a similar software package. In general, one-phase exponential decay is more appropriate for modeling short-lived transcripts and linear regression is more appropriate for long-lived transcripts [4, 5].
32. To verify that the transcript half-lives were properly modeled and not biased by the dynamic range of the intensity values measured in the oligoarrays, it is recommended to perform a correlation analysis (Spearman) between the calculated transcript half-lives and the transcript abundance prior to treatment (time 0 h). There should be no significant negative correlation between transcript half-lives and expression levels, which could be an indicative that the background signal is preventing the correct estimation of the decrease in transcript abundance along the time course of transcription inhibition, which could be misinterpreted as evidence of high stability [4, 5].

33. The accuracy of the estimated transcript half-lives will be better for those values that lie within the time window of the transcription inhibition experiment [2].
34. As an additional control to evaluate the biological robustness of the RNA stability data, we suggest an analysis of overrepresentation of specific Gene Ontology (GO) terms assigned to the mRNAs for which valid half-life measurements were calculated. GO terms associated with regulatory functions should be enriched among the most unstable mRNAs and terms associated to the housekeeping functions should be enriched among the most stable mRNAs [2, 4, 5, 10, 11].

References

1. Rabani M, Levin JZ, Fan L et al (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol 29(5):436–442
2. Sharova LV, Sharov AA, Nedorezov T et al (2009) Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res 16(1):45–58
3. Chen CY, Ezzeddine N, Shyu AB (2008) Messenger RNA half-life measurements in mammalian cells. Methods Enzymol 448:335–357
4. Ayupe AC, Tahira AC, Camargo L et al (2015) Global analysis of biogenesis, stability and subcellular localization of lncRNAs mapping to intragenic regions of the human genome. RNA Biol 12(8):877–892
5. Clark MB, Johnston RL, Inostroza-Ponta M et al (2012) Genome-wide analysis of long noncoding RNA stability. Genome Res 22(5):885–898
6. Dolken L, Ruzsics Z, Radle B et al (2008) High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA 14(9):1959–1972
7. Friedel CC, Dolken L, Ruzsics Z et al (2009) Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic Acids Res 37(17):e115
8. Ideue T, Adachi S, Naganuma T et al (2012) U7 small nuclear ribonucleoprotein represses histone gene transcription in cell cycle-arrested cells. Proc Natl Acad Sci U S A 109(15):5693–5698
9. Redrup L, Branco MR, Perdeaux ER et al (2009) The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development 136(4):525–530, dev.031328 [pii]
10. Schwanhausser B, Busse D, Li N et al (2011) Global quantification of mammalian gene expression control. Nature 473(7347):337–342. doi:10.1038/nature10098
11. Tani H, Mizutani R, Salam KA et al (2012) Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res 22(5):947–956
12. Beckedorff FC, Ayupe AC, Crocci-Souza R et al (2013) The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation. PLoS Genet 9(8):e1003705
13. DeOcesano-Pereira C, Amaral MS, Parreira KS et al (2014) Long noncoding RNA INXS is a critical mediator of BCL-XS induced apoptosis. Nucleic Acids Res 42(13):8343–8355
14. Tani H, Akimitsu N (2012) Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. RNA Biol 9(10):1233–1238. doi:10.4161/rna.22036
15. Bensaude O (2011) Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? Transcription 2(3):103–108
16. Brueckner F, Cramer P (2008) Structural basis of transcription inhibition by alpha-amanitin and implications for RNA polymerase II translocation. Nat Struct Mol Biol 15(8):811–818
17. Kaplan CD, Larsson KM, Kornberg RD (2008) The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. Mol Cell 30(5):547–556
18. Weinmann R, Raskas HJ, Roeder RG (1974) Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus

genome late in productive infection. Proc Natl Acad Sci U S A 71(9):3426–3439

19. Raha D, Wang Z, Moqtaderi Z et al (2010) Close association of RNA polymerase II and many transcription factors with Pol III genes. Proc Natl Acad Sci U S A 107(8):3639–3644
20. Nechaev S, Adelman K (2011) Pol II waiting in the starting gates: regulating the transition from transcription initiation into productive elongation. Biochim Biophys Acta 1809(1):34–45
21. Burger K, Muhl B, Harasim T et al (2012) Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. J Biol Chem 285(16):12416–12425
22. Lam LT, Pickeral OK, Peng AC et al (2001) Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anticancer drug flavopiridol. Genome Biol 2(10):RESEARCH0041
23. Raghavan A, Ogilvie RL, Reilly C et al (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. Nucleic Acids Res 30(24):5529–5538
24. Yang E, Van Nimwegen E, Zavolan M et al (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res 13(8):1863–1872, 13/8/1863 [pii]
25. Maekawa S, Imamachi N, Irie T et al (2015) Analysis of RNA decay factor mediated RNA stability contributions on RNA abundance. BMC Genomics 16:154
26. St Laurent G, Wahlestedt C, Kapranov P (2015) The Landscape of long noncoding RNA classification. Trends Genet 31(5):239–251
27. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29(9):e45
28. Simon RM, Dobbin K (2003) Experimental design of DNA microarray experiments. Biotechniques Suppl:16–21
29. Munchel SE, Shultzaberger RK, Takizawa N, Weis K (2011) Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. Mol Biol Cell 22(15):2787–2795

Chapter 12

A Novel Method to Quantify RNA–Protein Interactions In Situ Using FMTRIP and Proximity Ligation

C. Zurla, J. Jung, E.L. Blanchard, and P.J. Santangelo

Abstract

RNA binding proteins (RBP) and small RNAs regulate the editing, localization, stabilization, translation, and degradation of ribonucleic acids (RNAs) through their interactions with specific cis-acting elements within target RNAs. Here, we describe a novel method to detect protein–mRNA interactions, which combines FLAG-peptide modified, multiply-labeled tetravalent RNA imaging probes (FMTRIPs) with proximity ligation (PLA), and rolling circle amplification (RCA). This assay detects native RNA in a sequence specific and single RNA sensitive manner, and PLA allows for the quantification and localization of protein–mRNA interactions with single-interaction sensitivity.

Key words FMTRIPS, PLA, RCA, Posttranscriptional regulation, mRNA binding proteins

1 Introduction

As a gene is being transcribed, posttranscriptional events take place, including cotranscriptional mRNA processing (capping, splicing, 3′ end processing), nucleocytoplasmic export, and mRNA localization, prior to translation, mRNA stabilization, translational regulation, and decay. Different sets of RNA binding proteins (RBPs) are associated with mRNAs during each stage of their maturation [1–3]. These interactions depend on the RBPs availability in response to cellular or extracellular cues, as well as mRNA localization within different cellular compartments, which determines the assembly of temporally and spatially dynamic RNP complexes called messenger ribonucleoproteins (mRNPs) [4, 5]. Their specific molecular composition determines "the mRNA state," which is whether a transcript is immediately translated, transported to specific sites prior to translation, or transported to specific sites for storage and translation or tagged for degradation through quality control mechanisms.

Protein–mRNA interactions are characterized by an inhomogeneous spatiotemporal distribution within cells and across individual cells. Therefore, their investigation requires the development of

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_12, © Springer Science+Business Media New York 2017

ultrasensitive techniques. Single molecule approaches have emerged as powerful tools to resolve complex cellular processes [6–8], which are otherwise masked by ensemble averaging provided by most biochemical methods [9–11]. The main goal in the study of protein–RNA interactions at the single molecule level is to permit the simultaneous detection and quantification of both the mRNA and its interactions with specific trans-acting factors with single molecule and single-interaction sensitivity. In the present review we provide a detailed protocol of a novel method with the potential to achieve these objectives.

The RNA imaging strategy that we utilize is based on the use of MTRIPs, previously described in Santangelo et al. [12, 13]. MTRIPs are fluorescently labeled tetravalent single RNA sensitive probes consisting of a neutravidin core and four biotinylated fluorescently labeled oligonucleotides for specific mRNA targeting. MTRIPs are delivered to the cytoplasm of cultured cells via reversible membrane permeabilization using Streptolysin-O. Their chimeric 2′OMe-RNA nature ensures an optimal level of affinity to bind target mRNAs without inhibiting their function and metabolism, and without the toxicity observed for other chemical modifications. To date, MTRIPs have been successfully utilized for the quantification of endogenous mRNAs such as β-actin, c-myc, and polyA+ transcripts, and the colocalization with regulatory proteins and RNA granules in both fixed and living cells [14–16]. MTRIPs were also utilized to study the RSV genome, its organization in viral particles, its localization in infected cells and colocalization with viral and host proteins [17, 18]. In the protocol we describe here, mRNAs are identified using FMTRIPs, a peptide-modified version of MTRIPs, where a flag-tagged neutravidin constitutes the scaffold for the tetravalent probes (Fig. 1).

Typically, proteins are imaged using immunofluorescence (IF) or overexpression with fusion Fluorescent proteins. Colocalization and correlation functions are often utilized to observe the cellular

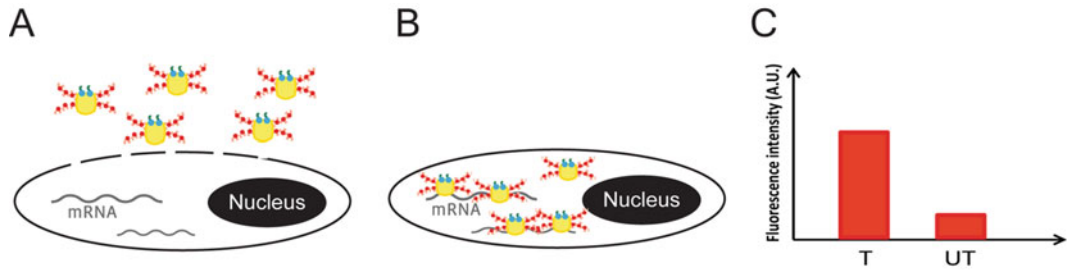


Fig. 1 Delivery of Flag-tagged MTRIPs (FMTRIPs): (**a**) FMTRIPs are composed by a flag-tagged (*green*) neutravidin core (*yellow*) and four fluorescently labeled (*red*) oligonucleotides. They are delivered into live cells via SLO-mediated reversible membrane permeabilization. (**b**) FMTRIPs rapidly bind to target mRNAs while cells recover after SLO treatment. (**c**) Single mRNAs bind multiple probes, and can be recognized by the enhanced signal to noise ratio (T = FMRTIPs bound to mRNA, UT = unbound FMTRIPs)

localization of mRNAs and related proteins and infer their association. However, these methods of analysis are poor indicators of physical interactions, because of limits in the optical resolution of the microscopes commonly utilized for these measurements. Here, we described a different approach based on a proximity ligation assay (PLA) that permits to quantify specific mRNA–protein interactions in situ.

PLA is a commercially available kit, whose overall approach was described in Soderberg et al. [19, 20]. It is generally utilized to study protein–protein interactions in fixed cells and tissues [21, 22]. The sample is initially incubated with primary antibodies specific for the proteins of interest and then the oligonucleotide-labeled proximity probes are added. If <40 nm apart, the oligonucleotides on the proximity probes come together to form a template for a circularized DNA strand by ligation. One of the proximity probes oligonucleotides then serves as template for the rolling circle amplification (RCA), which results in a coiled single-stranded DNA. The PLA product is detected by hybridizing complementary fluorescently labeled oligonucleotides. In PLA experiments using FMTRIPs one antibody targets the protein of interest, and the other one the flag tag on neutravidin (Fig. 2).

Using this method, we visualized and quantified interactions the genomic RSV RNA (labeled with FMTRIPs) and the viral N protein [23]. The results validated previous evidence obtained both via colocalization analysis and super resolution microscopy [17, 18]. We also interrogated the localization and frequency of interactions of native mRNAs with RBPs involved in posttranscriptional regulation such as HuR and TIA1, at both native and modulated protein levels [23, 24]. The results of our studies allowed to introduce a novel mechanism for fine-tuning of programmed cell death 4 (PDCD4), a tumor suppressor gene, in the regulation of protein levels to prevent neoplastic transformation [24]. Last, this method was also used to quantify the interactions between native mRNAs and microtubules, microfilaments and intermediate filaments, which are correlated with mRNA

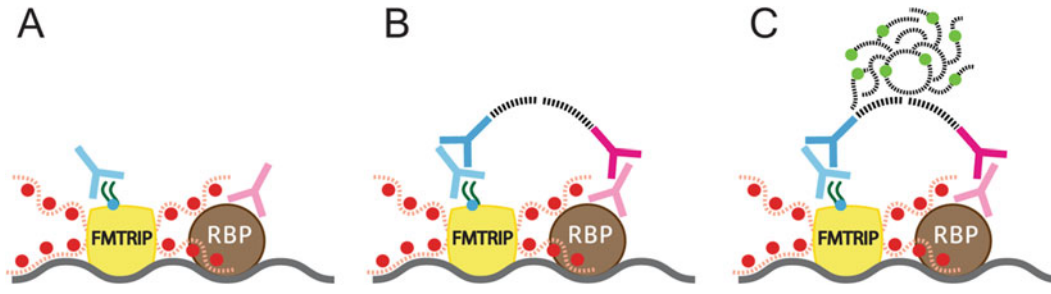


Fig. 2 PLA to detect protein–mRNA interactions using FMTRIPs: (**a**) Antibodies (*light blue* and *magenta*) bind to the Flag tag on FMTRIP and to the RNA binding protein (RBP) of interest. (**b**) Proximity probes (*dark blue* and *magenta*) bind to the primary antibodies and ligation occurs (**c**) A cy5-equivalent hybridized product (*green*) is synthesized via RCA

translation, during oxidative stress conditions [25]. The results we obtained were consistent with evidence obtained by our lab and other groups [15, 26–28].

2 Materials

2.1 Cell Culture and Reagents

1. Cells of choice.
2. High glucose Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS, 100 μ/ml Penicillin and 100 μg/ml streptomycin or other complete media.
3. 24-Well plates.
4. No 1.5 glass coverslips 10 mm.
5. Phosphate buffered saline without Calcium and Magnesium (PBS-Ca^{2+}-Mg^{2+}).
6. Optimem.
7. 1% Paraformaldehyde in 1× PBS (PFA).

2.2 FMTRIPs Assembly and Delivery

1. Biotinylated oligonucleotides with internal amino groups.
2. cy3b NHS ester or, alternatively, Dylight650 NHS ester.
3. Dimethyl Sulfoxide (DMSO).
4. 0.1 M Sodium Bicarbonate.
5. Neutravidin.
6. Flag tag-hyNic (Solulink).
7. S-4FB (Solulink).
8. 10× Modification and 10× Conjugation buffers (Solulink).
9. 2 μ/ml Streptolysin-O.
10. Tris (2-carboxyethyl)phosphine.
11. 1× Phosphate buffered saline (PBS).
12. 3 and 30 kDa centrifugal filters.

2.3 Proximity Ligation Assay

1. Primary antibodies of choice.
2. Primary antibody solution: 0.25% gelatin, 0.5% Triton X-100, 0.5% donkey serum and 1% BSA in PBS.
3. PLA probes solution: 0.05% Tween-20 in PBS.
4. PLA kit (Olink Bioscience).
5. 0.2% Triton X-100.
6. Modified blocking solution: 0.5% Tween-20, 0.1% Triton X-100, 0.1% gelatin, 2% donkey serum and 1% bovine serum albumin (BSA) in PBS.
7. PLA mounting media with DAPI.

8. Slides holders.
9. Beaker.
10. Humidity chamber.
11. Shaker.

2.4 Microscope Operation and Processing

1. Zeiss Axiovert 200 M with an UltraVIEW Spinning disk Confocal microscope (PerkinElmer).
2. 60× numerical aperture (NA) 1.4 Plan-Apochromat objective (Zeiss).
3. Flash 4.0v2CMOS camera (Hamamatsu).
4. Imaging and processing via the Volocity software (PerkinElmer).

3 Methods

3.1 Neutravidin Labeling with Flag Peptides

Neutravidin represents the core of MTRIPs, since it allows for the tetramerization of the biotinylated mRNA-targeting oligonucleotides. In PLA studies, ideally, the neutravidin should also provide for the epitopes recognized by the primary antibody. The primary antibodies for PLA are crucial reagents, which must be carefully selected for purity and specificity. Unfortunately, we could not identify any anti-neutravidin (or streptavidin) antibodies suitable for our studies. As an alternative strategy, we opted for a flag-peptide-based system, since several highly specific anti-flag antibodies are commercially available. FMTRIPs are a Flag peptide-modified version of MTRIPs. The following protocol, schematically illustrated in Fig. 3a, is optimized to produce neutravidin molecules labeled, on average, with 2 Flag-peptides, which ensures maximum PLA efficiency. Indeed, we observed that three Flags per neutravidin caused reduced PLA efficiency, probably due to steric hindrance or self-quenching of the fluorophores used for detection [23]. The use of HyNic-4FB system permitted to perform a controlled reaction, which results in a stable, covalent bond between the peptide and the neutravidin. The degree of labeling can be quantified after neutravidin conjugation to both 4FB and Flag-peptide, for quality control.

1. Dilute 80 µl of 10 mg/ml neutravidin in 1× Modification buffer to have a 2 mg/ml solution.
2. Add 4.9 µl of 20 mg/ml S-4FB in DMSO (corresponding to 30× molar excess).
3. Incubate the reaction for 2 h at room temperature.
4. Add 400 µl of 1× conjugation buffer and remove unincorporated 4FB using a 30 kDa filter (3 times at 10,000 × *g* for 5 min).

5. Recover the neutravidin-S4FB by reverse spinning (2 min at 1000 × *g*) and adjust the concentration of the solution to 2 mg/ml using 1× conjugation buffer (*see* **Note 1**).
6. Add 5 μl of Flag-HyNic 10 mg/ml in DMSO (corresponding to 3× molar excess).
7. Incubate overnight at room temperature.
8. Add 350 μl of PBS and clean up using a 30 kDa filter (3 times at 10,000 × *g* for 5 min).
9. Recover the neutravidin-Flag by reverse spinning (2 min at 1000 × *g*).
10. Determine the molar substitution ratio by measuring the ratio of the concentration of the S4-FB-HyNic bond and the concentration of neutravidin. The reaction between 4-FB reagent and HyNic-modified proteins leads to the formation of a traceable absorbance signal at 350 nm with a molar extinction coefficient of 24,000 mol^{-1} cm^{-1} (*see* **Note 2**).
11. Dilute the Flag-tagged neutravidin to 6 μM in 1× PBS and store at 4 °C.

3.2 Oligonucleotide Labeling with Fluorophores

The oligonucleotides utilized for FMTRIPs consist of a 5′ biotin modification, a linker sequence, composed typically by 6 thymidine nucleotides, and a ~20 bases-long antisense hybridization sequence, composed by 2′-O-Methyl RNA nucleotides, and three to four C6-amino modified thymines. The linker sequence is included to extend the hybridization region from the neutravidin, facilitating the binding to target mRNAs. The hybridization sequence should have a GC content of about 50%, should not form secondary structures or primer dimers (which can be verified using software like MFOLD), and should be aligned against known transcripts, to ensure specificity for the mRNA of interest. The antisense sequence should not overlap known binding sites for regulatory proteins or miRNAs. The C-6 amino modified thymines are used as sites of conjugation of fluorophores, typically Cy3b or Dylight 650, which are characterized by a long fluorescent lifetime and good resistance to photobleaching. We typically utilize 3 FMRTIPS per mRNA, preferentially targeting the 3′UTR. The following protocol describes the oligonucleotide-labeling procedure we currently utilize, which yields a DOL of about 2–3 fluorophores per oligonucleotide (Fig. 3b).

1. Resuspend cy3b in DMSO at 25 mM concentration. The resuspended cy3b can be subsequently stored at −20 °C in ready-to-use aliquots.
2. Resuspend the biotinylated oligonucleotide in nuclease free water at 450 μM
3. Combine 6 μl of oligonucleotide, 3.5 μl of fluorophore and 37.5 μl of 0.1 M Bicarbonate buffer

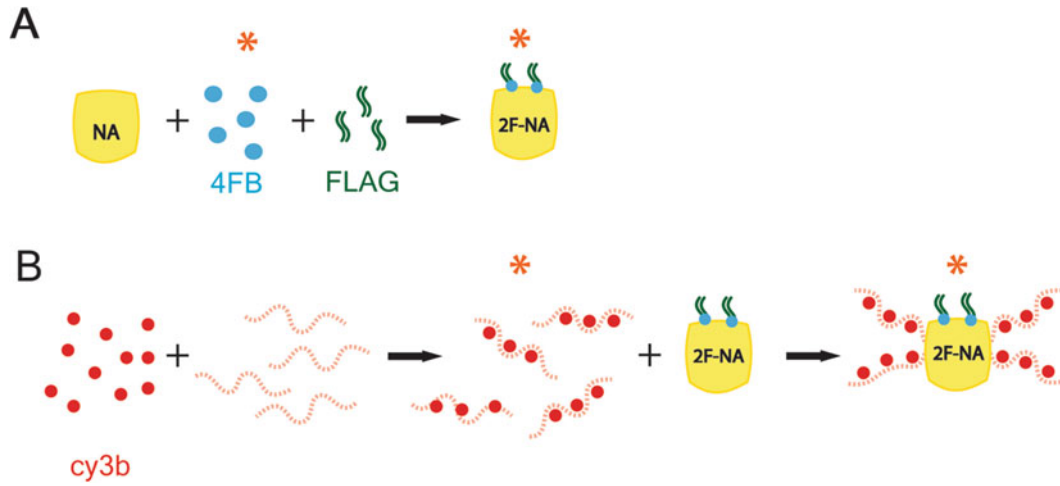


Fig. 3 Details of FMTRIP composition and assembly: (**a**) neutravidin (*yellow*) is modified with 2 HyNic-Flag-tags (*green*) via the 4FB Linker (*blue*). (**b**) Biotinilated oligonucleotides are fluorescently labeled with cy3b (*red*) and subsequently incubated with Flag-tagged neutravidins to obtain functional FMTRIPs. *Red* asterisks represent "checkpoints" where reaction products can be quantified

4. Vortex gently for 4 h protected from light at room temperature
5. Remove unincorporated dyes by centrifugation three times in 1× PBS using 3 kDa filters (20 min at 10,000 × *g*).
6. Reverse spin in a new collection tube to retrieve the labeled oligonucleotide (2 min at 1000 × *g*).
7. Determine oligonucleotide molar concentration and degree of labeling using UV–Vis. Measure the absorbance at 260 nm and at 565 nm. The molar concentration of nucleic acids and dye is given by [Concentration] = Abs × dilution factor × Ɛ (Ɛ cy3b=150,000 M^{-1} cm^{-1}; Ɛoligo is provided by the manufacturer). The degree of labeling (DOL) is given by the ratio of [oligo]/[dye] (*see* **Notes 3** and **4**).
8. Dilute the labeled oligonucleotide at 30 μM in 1× PBS and store at −20 °C in ready-to-use aliquots.

3.3 FMTRIPs Assembly

The following protocol describes the final step to obtain functional FMTRIP molecules (Fig. 3b). FMTRIPs should be freshly prepared and purified before delivery to live cells. If different hybridization antisense sequences are utilized, each FMTRIP should be assembled and purified separately. At the end of this section we will discuss two assays that can be utilized to verify the successful assembly of the probes.

1. Combine 5 μl of 30 μM cy3b-labeled oligonucleotide and 5 μl of 6 μM neutravidin-flag.
2. Incubate for 1 h at room temperature protected from light.

3. Remove excess of oligonucleotide by centrifugation three times in 1× PBS using 30 kDa (5 min at 10,000 × *g*).
4. Recover the FMTRIPs by reverse spinning (2 min at 1000 × *g*). The concentration of the FMTRIPs is about 1 μM.
5. Using the same procedure, assemble MTRIPs (containing neutravidin without Flag-peptide that will be utilized as a control experiment). *See* **Note 5**.

Successful assembly of FMTRIPs should be verified in preliminary measurements, to ensure that the conjugation of flag-peptides does not compromise the ability of neutravidin to bind to the biotinylated oligonucleotides. This can be tested using, first of all, a gel retardation assay where the migration of biotinylated and fluorescently labeled oligonucleotides is compared to that of FMTRIPs in a native 20 % TBE polyacrylamide gel. The latter, will result in a band shifted toward higher molecular weights than the former. The effect of the flag-peptide modification can be easily observed by comparing the bands obtained after assembly of MTRIPs (with unmodified neutravidin) and FMTRIPs (Fig. 4a). Successful assembly of FMTRIPs can also be verified after deposition of the probes on glass coverslip. The probes are incubated for 30 min in growth media at low concentration (1 nM) in order to deposit single molecules. The media is removed, and the probes are fixed in 1 % PFA and mounted. This is a procedure we routinely utilize to test new imaging molecules, and the samples can be easily analyzed using wide field deconvolution microscopy or confocal microscopy. Experiments performed using a mixture of probes labeled with two different dyes (such as cy3b and Dylight 650) demonstrated that they do not form aggregates because they displayed similar mean fluorescence intensities and no colocalization between the two dyes was observed (Fig. 4b). FMTRIPs can be additionally visualized using immunofluorescence (IF) after fixation and blocking with a flag-tag antibody and a suitable secondary antibody. Experiments performed using cy3b labeled FMTRIPs and Alexafluor 488 labeled anti-flag antibody yielded high colocalization between the two fluorescent signals, demonstrating that the modified neutravidin retain their functionality (Fig. 4c).

3.4 Delivery to Cells

The following protocol describes the procedure to deliver FMTRIPs (and MTRIPs for control experiments) to cells using the pore-forming toxin Streptolysin-O (SLO). We utilized this delivery method successfully with several cell lines and primary cells including A549, Hela, U2OS, Vero, VERO C1008, Hep, LnCaP, RAW 286.7, DU145, MDBK, MDCK, HDF, CEF, MCF-7. Standard working concentrations of SLO are 0.15 μ/ml up to 0.8 μ/ml. The optimal SLO concentration should be determined in preliminary experiments for each cell type, as described at the end of this paragraph.

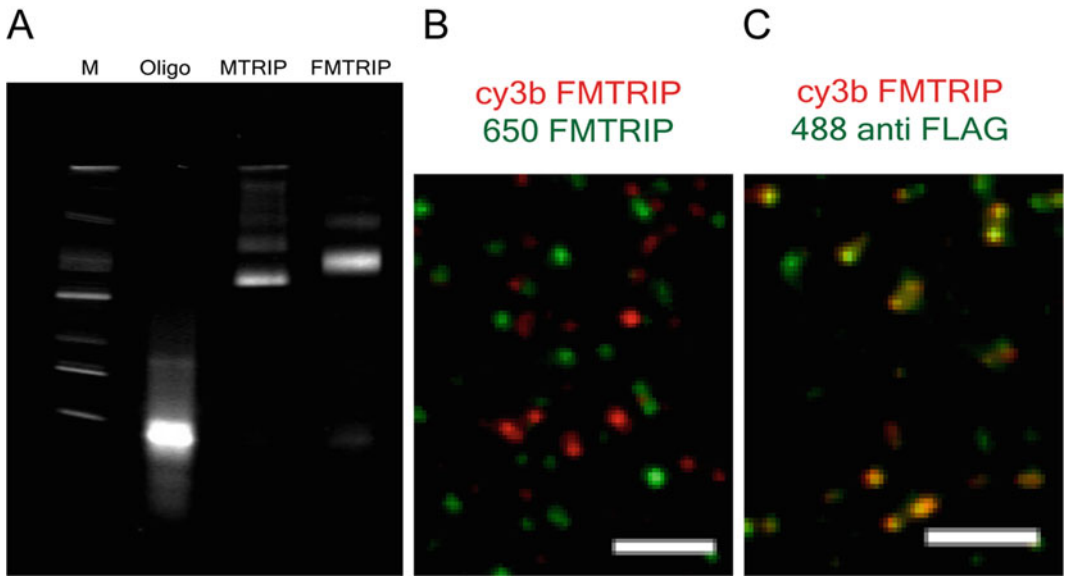


Fig. 4 Flag-tagged neutravidin as a core for functional mRNA imaging probes: (**a**) Migration of oligonucleotides, MTRIPs, and FMTRIPs in a 20% TBE gel demonstating the band shift due to the assembly of functional FMTRIPs. (**b**) Single cy3b (*red*) and 650 (*green*) labeled FMTRIPs do not aggregate or colocalize when deposited on glass. (**c**) Colocalization between cy3b-labeled FMTRIPs (*red*) and Flag immunofluorescence (*green*). Scale bars are 2.5 μm

1. Cells are plated the day prior to the experiment on 10 mm coverslips in 24 well plates (*see* **Note 6**).
2. Activate SLO by adding 1.5 μl TCEP to 100 μl of 2 μ/ml SLO in H_2O and incubate 1 h at 37 °C (*see* **Note 7**).
3. Wash cells once with PBS without Calcium and magnesium
4. Dilute the 100 μl of activated SLO 1:10 in Optimem to have s 0.2 μ/ml solution.
5. Add the FMTRIPs to the activated SLO. The final concentration of the FMTRIPs for delivery is about 30 nM.
6. Add to each well 250 μl probes and incubate 10 min at 37 °C
7. Remove probes and add complete media. Incubate for 15 min at 37 °C for recovery.
8. Fix cells in 1% PFA for 10 min at room temperature (*see* **Note 8**).
9. Wash once with PBS.
10. Permeabilize cells using 0.2% Triton X for 5 min at room temperature.
11. Wash once with PBS.
12. Block using modified blocking buffer for 1 h at 37 °C (*see* **Note 9**).
13. Wash once with PBS for 5 min.

The optimal SLO concentration for efficient delivery can be tested using a given amount (30 nM) of a probe targeting, for example, βactin mRNAs and a scrambled probe (Fig. 5). The former will appear homogeneously distributed in the cell cytoplasm, near the nucleus and around the edges of the cell. RNA granules display heterogeneous intensity profiles, depending on the RNA content of different RNPs. The latter, instead, will appear prevalently distributed around the cell nucleus and not around the edges of the cell. Untargeted probes display uniform intensities, comparable to those of single probes. Low SLO concentrations result in poor probe delivery, with most of the fluorescent signal observed on glass, outside the cell. High SLO concentrations results in sequestration of probes within hollow vesicles in the cell cytoplasm.

3.5 Proximity Ligation Assay

All incubations are performed with coverslips face- down on parafilm in a humidity chamber to keep the reaction volume ~40 μl as recommended by the manufacturer. All washings are done placing the coverslips in a coverslip holder submerged in a beaker containing the wash buffer. For gentle shaking we use a bench rocker set at low speed. Experiments should be designed including two controls: (a) omitting one of the primary antibodies and (b) Using MTRIPs instead of FMTRIPs (Fig. 6).

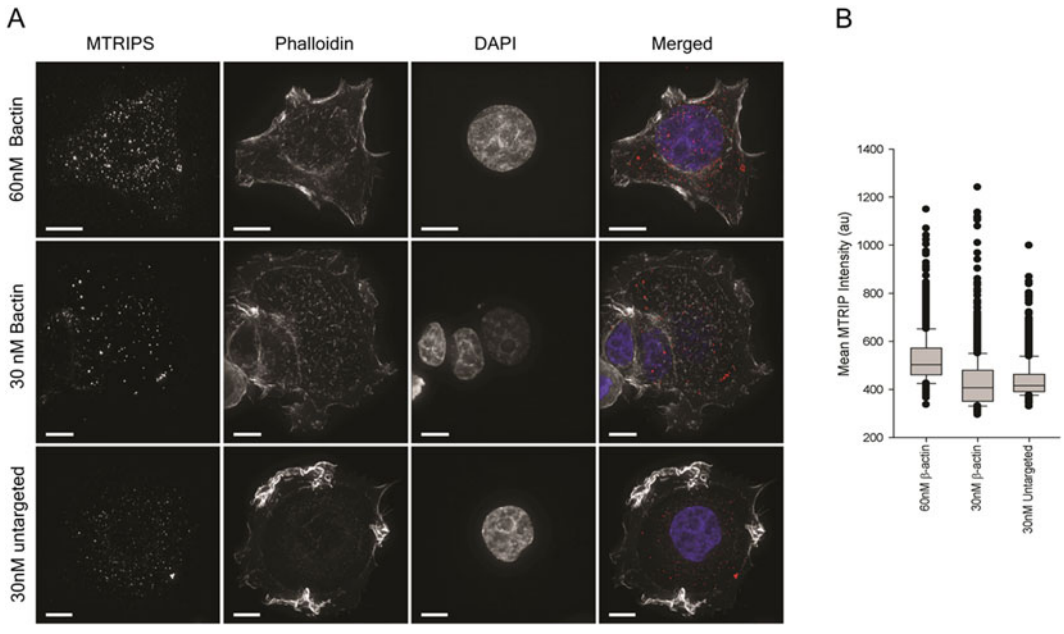


Fig. 5 Observation of efficient SLO delivery of probes: (**a**) Comparison of the distribution of 60 and 30 nM probes targeting βactin mRNA and 30 nM of untargeted probes. Phalloidin was used to identify the contour of the cell, and to demonstrate localization of targeted MTRIPs to the edges of the cell. Nuclei are stained with DAPI. (**b**) Mean MTRIP intensity observed in the experiments in (**a**)

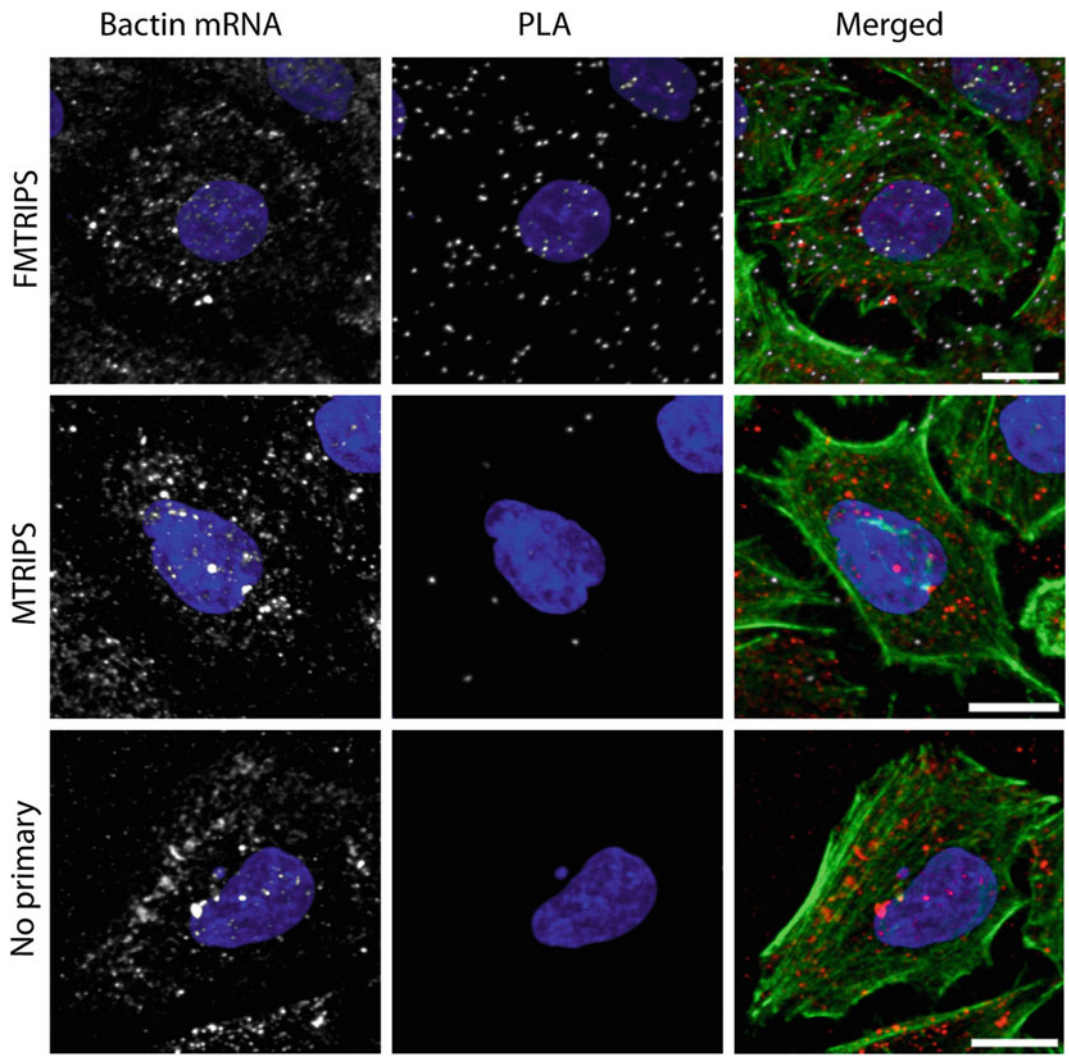


Fig. 6 Visualization of interactions between βactin mRNA and stress F-actin: Three cy3b labeled FMTRIPs (*red*) targeting the 3′UTR of βactin mRNA were delivered to cells using SLO. F-actin was visualized using Phalloidin Alexafluor488 (*green*). For PLA, we used an αFlag primary antibody, an α488 primary antibody, and far-red detection reagent (*gray*). We routinely perform two control measurements, one using MTRIPs and one omitting one of the primary antibodies. Nuclei are stained with DAPI (*blue*). Scale bar is 10 μm

1. Incubate samples with the primary antibody diluted in the primary antibody solution for 30 min at 37 °C (*see* **Note 10**).
2. Wash 10 min in buffer A (provided by manufacturer).
3. Incubate with PLA plus and minus probes for 30 min at 37 °C.
4. Wash 10 min in buffer A.
5. Incubate with the ligase diluted as recommended for 30 min at 37 °C.
6. Wash 10 min in buffer A.

7. Incubate with the Amplification solution diluted as recommended for 100 min at 37 °C (*see* **Notes 11** and **12**).
8. Wash with 1× buffer B (provided by manufacturer) for 20 min.
9. Wash with 0.01× buffer B for 1 min.
10. If immunostaining is required, wash once with 1× PBS before adding the primary antibody. For example, phalloidin staining (which binds to F-actin) can be used to identify the contour of single cells in the sample for subsequent analysis (*see* **Note 13**).
11. Mount with Duolink mounting medium with DAPI and wait for 15 min with samples protected from light.
12. Seal with nailpolish before imaging. Slides should be stored at –20 °C.

3.6 Detection and Analysis

In the current paragraph we will describe the detection and analysis procedure we perform utilizing the Volocity software (PerkinElmer) using a spinning disk confocal microscope. We also used laser scanning confocals as well as wide field deconvolution microscopes successfully. Analysis is performed on single cells identified by the mRNA signal or immunofluorescence.

1. Record image stacks at 300 nm intervals to adequately sample volumes.
2. Import the acquired files in the Volocity analysis software and linearly contrast enhance for display
3. In a single cell, determine the mRNA volume based on standard deviation intensity of the probes.
4. Identify the PLA puncta as "Objects" by their standard deviation intensity then separate into individual punctae using the "separate touching object" tool
5. The PLA punctae can be further filtered based on their size and maximum intensity (*see* **Note 14**).
6. The PLA frequency can be measured as the ratio of the number of PLA punctae and the fluorescent FMTRIP volume. In this way, interactions are normalized to the mRNA signal, allowing comparison of the quantification between cells (*see* **Note 15**)
7. Perform statistical analysis on the results using a software like Sigma plot.

3.7 Conclusions and Future Perspectives

The use of FMTRIP and PLA offers a relatively simple approach to the investigation of protein–mRNA interactions in single cells with single-interaction sensitivity. Protein–mRNA interactions were quantified by measuring the PLA frequency, defined as the ratio of the number of PLA punctae and the mRNA volume (identified by the fluorescent probes labeling the mRNAs). The limits of traditional colocalization measurements can be overcome using PLA because the

antibodies are not utilized for direct observation of a protein but their specific binding to target proteins is necessary to start a reaction which amplifies the signal rendering the interaction visible. The result of the assay can be observed using a deconvolution or a confocal microscope. Various fixatives can be utilized to preserve specific cellular components without affecting the specificity and quality of the results. Crucial reagents in the assay are the primary antibodies specific for the Flag tag on FMTRIPs and the protein of interest, which have to be carefully tested in initial assays. Potentially all protein involved in translational regulation can be examined, provided "a good" antibody, such as TTP, TIAR, or CUG-BP, but also proteins involved in mRNA decay like DCP1a or the proteasome components. The effect of RBPs on miRNA binding could additionally be addressed analyzing the interactions between the RNA-induced silencing complex (RISC) or Ago2 and RBPs. Identifying the localization of these interactions and whether RBPs bind cooperatively or competitively with other RBPs can help establish a model system for examining how changes in RBPs modulate mRNAs and their translation.

Critical aspects of the presently described method can be improved. PLA relies on two enzymatic reactions catalyzed by the ligase and phi29 DNA polymerase. Although these enzymatic reactions are generally predictable, they do not allow for optimization. Ensuring uniform enzymatic reactions across all the samples is virtually impossible. Potentially, this could be overcome by using a method that does not rely on enzymatic reactions, but yet ensures signal amplification. The proximity dependent Hybridization chain reaction method (proxHCR) recently described by Koos et al. [29] might represent a useful implementation to FMTRIP-based mRNA imaging for the detection of protein–mRNA interactions. In proxHCR when two oligonucleotide hairpins conjugated to antibodies bind in close proximity, they can be activated to reveal an initiator sequence. This starts a chain reaction of hybridization events between a pair of fluorophore-labeled oligonucleotide hairpins, generating a fluorescent product. This method would allow for detection of protein–protein or RNA–protein interactions without the need of enzymes. In conclusion, the quantitative characterization of the distribution of interactions for various mRNAs and RBPs provided by our method would be helpful in establishing the cell-to-cell variability for posttranscriptional regulatory events and how they contribute to mRNA "state."

4 Notes

1. The molar substitution ratio (number of S4-FB per neutravidin) can be measured using the 2-HP reagent according to the manufacturer protocol. The reaction between the 2-HP reagent and 4FB-modified proteins leads to the formation of a

traceable absorbance signal at 350 nm with a molar extinction coefficient at 18,000 l/mol cm.

2. The concentration of neutravidin should be determined using a BCA assay or a Bradford assay. The concentration of neutravidin cannot be measured via UV–Vis at 280 nm in a reliable way because of the S4-FB-HyNic bond.
3. High DOL often imply over-labeling, which causes quenching of fluorescence and/or inefficient binding of the oligonucleotide to its target mRNA.
4. Another dye we routinely use for oligonucleotide labeling is Dylight 650 NHS ester. After oligonucleotide labeling with this dye, unincorporated fluorophores are not efficiently removed using 3 kDa filters, due to the larger molecular weight (>1000 g/mol). We recommend using Zeba desalting columns (Thermo Fisher scientific) according to manufacturer instructions at least three times. $\mathcal{E}$ 650 = 250,000 M^{-1} cm^{-1}.
5. We routinely use FMTRIPs designed to target at least three sequences in the 3′UTR of target mRNAs, near the RPB binding site, since PLA detects interaction between elements <40 nm apart. We observed that increasing the number of FMTRIPs did not significantly affect PLA frequency.
6. Prior to plating, the coverslips are cleaned with ethanol, dried with a kimwipe, and incubated at 37 °C submerged in complete media for at least 30 min to promote cell adhesion.
7. SLO is dissolved in molecular biology grade water to a concentration of 25,000 μ/ml, then aliquot into 2 μ/ml (100 μl each) and stored at −20 °C. Do not use DEPC-treated water and do not freeze/thaw the aliquots, as this will decrease SLO activity.
8. Different fixatives have been tested and were successfully utilized for PLA with FMTRIPs, such as 100% Methanol, 50% Methanol: 50% Acetone or PFA in BRB 80 buffer (ideal to preserve microtubules).
9. In preliminary experiments we tested both the blocking solution provided by the manufacturer and the modified blocking solution described here. While the former resulted in nonspecific signal, the latter efficiently eliminated the background signal.
10. The concentration of the primary antibodies must be determined accurately for PLA (usually dilution 1:500 or higher, depending on the antibodies).
11. The amplification solution contains the “detection reagent,” which is light sensitive. In this step and all the subsequent passages, samples should be protected from light.
12. The detection reagent is available green, red, orange, and far-red and is therefore amenable for different filter sets and potentially ideal for multiplexing. We routinely utilize the far-red one, because it is the brightest and most photostable.

13. We noted that performing immunofluorescence before the PLA protocol yields generally poor results probably as result of the washings with the required buffers.
14. A PLA product is a micron-sized fluorescent puncta with consistent size and intensity for a variety of analytes, antibodies, and cell types.
15. It is important to notice that not every FMTRIP participates in PLA productively for at least two reasons: (a) PLA detects interactions present at the time of fixation and (b) the distance between the FMTRIPs and the antibody against the protein of interest may exceed the distance for proximity ligation. For this reason, PLA represents a powerful tool for quantifying and comparing the relative change of interactions between different experimental conditions, rather than obtaining absolute numbers. Moreover, when performing PLA experiments, mRNAs and proteins are typically under-sampled in order to detect their interactions randomly for comparative analyses. Usually, the FMTRIPs and primary antibodies are used in concentrations that provide minimal nonspecific binding. While this allows for detection of relative differences between samples, it may cause under-sampling of the interactions especially for less abundant mRNAs.

References

1. Keene JD (2007) RNA regulons: coordination of post-transcriptional events. Nat Rev Genet 8:533–543
2. Ho JJ, Marsden PA (2014) Competition and collaboration between RNA-binding proteins and microRNAs. Wiley Interdiscip Rev RNA 5:69–86
3. Keene JD, Tenenbaum SA (2002) Eukaryotic mRNPs may represent posttranscriptional operons. Mol Cell 9:1161–1167
4. Giorgi C, Moore MJ (2007) The nuclear nurture and cytoplasmic nature of localized mRNPs. Semin Cell Dev Biol 18:186–193
5. Gonsalvez GB, Long RM (2012) Spatial regulation of translation through RNA localization. F1000 Biol Rep 4:16
6. Gaspar I, Ephrussi A (2015) Strength in numbers: quantitative single-molecule RNA detection assays. Wiley Interdiscip Rev Dev Biol 4:135–150
7. Lorenz M (2009) Visualizing protein-RNA interactions inside cells by fluorescence resonance energy transfer. RNA 15:97–103
8. Pitchiaya S, Heinicke LA, Custer TC et al (2014) Single molecule fluorescence approaches shed light on intracellular RNAs. Chem Rev 114:3224–3265
9. Mchugh CA, Russell P, Guttman M (2014) Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol 15:203
10. Cook KB, Hughes TR, Morris QD (2015) High-throughput characterization of protein-RNA interactions. Brief Funct Genomics 14:74–89
11. Halbeisen RE, Galgano A, Scherrer T et al (2008) Post-transcriptional gene regulation: from genome-wide studies to principles. Cell Mol Life Sci 65:798–813
12. Santangelo PJ, Lifland AW, Curt P et al (2009) Single molecule-sensitive probes for imaging RNA in live cells. Nat Methods 6:347–349
13. Santangelo PJ, Alonas E, Jung J et al (2012) Probes for intracellular RNA imaging in live cells. Methods Enzymol 505:383–399
14. Zurla C, Lifland AW, Santangelo PJ (2011) Characterizing mRNA interactions with RNA granules during translation initiation inhibition. PLoS One 6:e19727
15. Lifland AW, Zurla C, Yu J et al (2011) Dynamics of native beta-actin mRNA transport in the cytoplasm. Traffic 12:1000–1011

16. Lifland AW, Zurla C, Santangelo PJ (2010) Single molecule sensitive multivalent polyethylene glycol probes for RNA imaging. Bioconjug Chem 21:483–488
17. Lifland AW, Jung J, Alonas E et al (2012) Human respiratory syncytial virus nucleoprotein and inclusion bodies antagonize the innate immune response mediated by MDA5 and MAVS. J Virol 86:8245–8258
18. Alonas E, Lifland AW, Gudheti M et al (2014) Combining single RNA sensitive probes with subdiffraction-limited and live-cell imaging enables the characterization of virus dynamics in cells. ACS Nano 8:302–315
19. Soderberg O, Gullberg M, Jarvius M et al (2006) Direct observation of individual endogenous protein complexes in situ by proximity ligation. Nat Methods 3:995–1000
20. Clausson CM, Allalou A, Weibrecht I et al (2011) Increasing the dynamic range of in situ PLA. Nat Methods 8:892–893
21. Soderberg O, Leuchowius KJ, Gullberg M et al (2008) Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. Methods 45:227–232
22. Leuchowius KJ, Weibrecht I, Soderberg O (2011) In situ proximity ligation assay for microscopy and flow cytometry. Curr Protoc Cytom. J. Paul Robinson, managing editor … [et al.] Chapter 9:Unit 9 36
23. Jung J, Lifland AW, Zurla C et al (2013) Quantifying RNA-protein interactions in situ using modified-MTRIPs and proximity ligation. Nucleic Acids Res 41:e12
24. Wigington CP, Jung J, Rye EA et al (2015) Post-transcriptional regulation of programmed cell death 4 (PDCD4) mRNA by the RNA-binding proteins human antigen R (HuR) and T-cell intracellular antigen 1 (TIA1). J Biol Chem 290:3468–3487
25. Jung J, Lifland AW, Alonas EJ et al (2013) Characterization of mRNA-cytoskeleton interactions in situ using FMTRIP and proximity ligation. PLoS One 8:e74598
26. Condeelis J, Singer RH (2005) How and why does beta-actin mRNA target? Biol Cell 97:97–110
27. Sundell CL, Singer RH (1991) Requirement of microfilaments in sorting of actin messenger RNA. Science 253:1275–1277
28. Bassell GJ, Powers CM, Taneja KL et al (1994) Single mRNAs visualized by ultrastructural in situ hybridization are principally localized at actin filament intersections in fibroblasts. J Cell Biol 126:863–876
29. Koos B, Cane G, Grannas K et al (2015) Proximity-dependent initiation of hybridization chain reaction. Nat Commun 6:7294

Chapter 13

In Silico Promoter Recognition from deepCAGE Data

Xinyi Yang and Annalisa Marsico

Abstract

The accurate identification of transcription start regions corresponding to the promoters of known genes, novel coding, and noncoding transcripts, as well as enhancer elements, is a crucial step towards a complete understanding of state-specific gene regulatory networks. Recent high-throughput techniques, such as deepCAGE or single-molecule CAGE, have made it possible to identify the genome-wide location, relative expression, and differential usage of transcription start regions across hundreds of different tissues and cell lines. Here, we describe in detail the necessary computational analysis of CAGE data, with focus on two recent in silico methodologies for CAGE peak/profile definition and promoter recognition, namely the Decomposition-based Peak Identification (DPI) and the PROmiRNA software. We apply both methodologies to the challenging task of identifying primary microRNAs transcript (pri-miRNA) start sites and compare the results.

Key words TSS, Promoter, microRNAs, DPI, PROmiRNA

1 Introduction

Gene expression is regulated at many levels, including chromatin packing, transcription initiation, polyadenylation, splicing, mRNA stability, and others. One of the most important regulatory steps is transcription initiation, which is coordinated by the binding of many proteins to gene promoters and enhancers. Combinations of binding sites determine the expression context of a certain gene and its activity in a certain tissue or condition [1, 2].

The annotation of gene promoters, as well as other transcriptionally-active regulatory sequences is essential to understand biological mechanisms underlying context-specific gene regulatory networks. But what is a promoter exactly and how can it be precisely defined? A promoter is not a clearly defined unit and to this question there is no unique answer, although scientists studying gene regulation largely agree nowadays on the fact that a promoter can be defined as the region surrounding the Transcriptional Start Site (TSS) of a gene which contains regulatory elements and Transcription Factor Binding Sites (TFBBs) necessary to initiate gene transcription [1–3].

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_13, © Springer Science+Business Media New York 2017

The promoter structure of a eukaryotic organism is more complex than a prokaryotic one, with the complexity increasing from single-celled yeast to mammals, and regulatory elements spread over large genomic space [1]. Although eukaryotes have different types of RNA polymerases, RNA polymerase II is responsible for transcription of mRNAs, as well other classes of noncoding RNAs, including some microRNAs and long noncoding RNAs (lncRNAs).

The region of 30–100 nucleotides surrounding the TSS is generally referred to as *core promoter* (Fig. 1) and contains interchangeable sequence elements and general transcription factor binding sites recognized by the preinitiation complex (PIC) which initiates transcription from a loosely conserved Initiator site (Inr). The PIC includes, besides Polymerase II, general initiation factors such as TFIIA, TFIIB, a TATA box binding protein (TBP) which binds specific DNA elements about 25 base pair (bp) upstream of the TSS, and several TBP-associated factors (TAFs). This core promoter may also contain downstream elements like DPE and MTE (in fly), BRE upstream or downstream elements or DCE, downstream core element (in vertebrates) [2, 4].

The region further away (up to 500 bp upstream of the gene TSS) is usually referred to as *proximal promoter* and contains other promoter elements, such as the GC box and/or the CAAT box, as well as more specific TFBS necessary to coordinate transcription in a tissue- and developmental stage-specific manner. TFBSs can also occur in clusters, forming cis-regulatory modules (CRMs) [4, 5].

Distal regulatory elements also influence transcription, including enhancers, active regions which enhance gene transcription, insulators, which mark boundaries between DNA active regions, and silencers, regions which repress gene transcription. These elements are part of the so-called *distal promoter*, which can extend up to several kb from the TSS (upstream and/or downstream) [2, 4]. Finally, in most eukaryotic genomes, chromatin is made of basic units called nucleosomes. A nucleosome is composed of a segment of DNA wrapped around a histone core. Chromatin structure can be tightly wrapped or accessible to proteins: active promoters are usually found in accessible chromatin regions (or nucleosome-free regions) [2].

Earlier experimental methods for promoter identification, such as nuclease protection and primer extension can identify promoters on a gene-by-gene basis and cannot be extended genome wide. Later advances in promoter identification are sequencing methods, such as RACE 5′-tag sequencing of cDNA or mRNA sequences, which rely on reverse transcription, fragmentation, and amplification of cDNAs and alignment to the genome to get information about TSS location [6]. Other high-throughput experimental procedures include hybridization methods, such as oligonucleotide tiling arrays [4].

Back in early 1990s, experimental techniques for promoter identification were costly, labor intensive, time consuming, and not really applicable genome wide. Hence, several in silico methods for

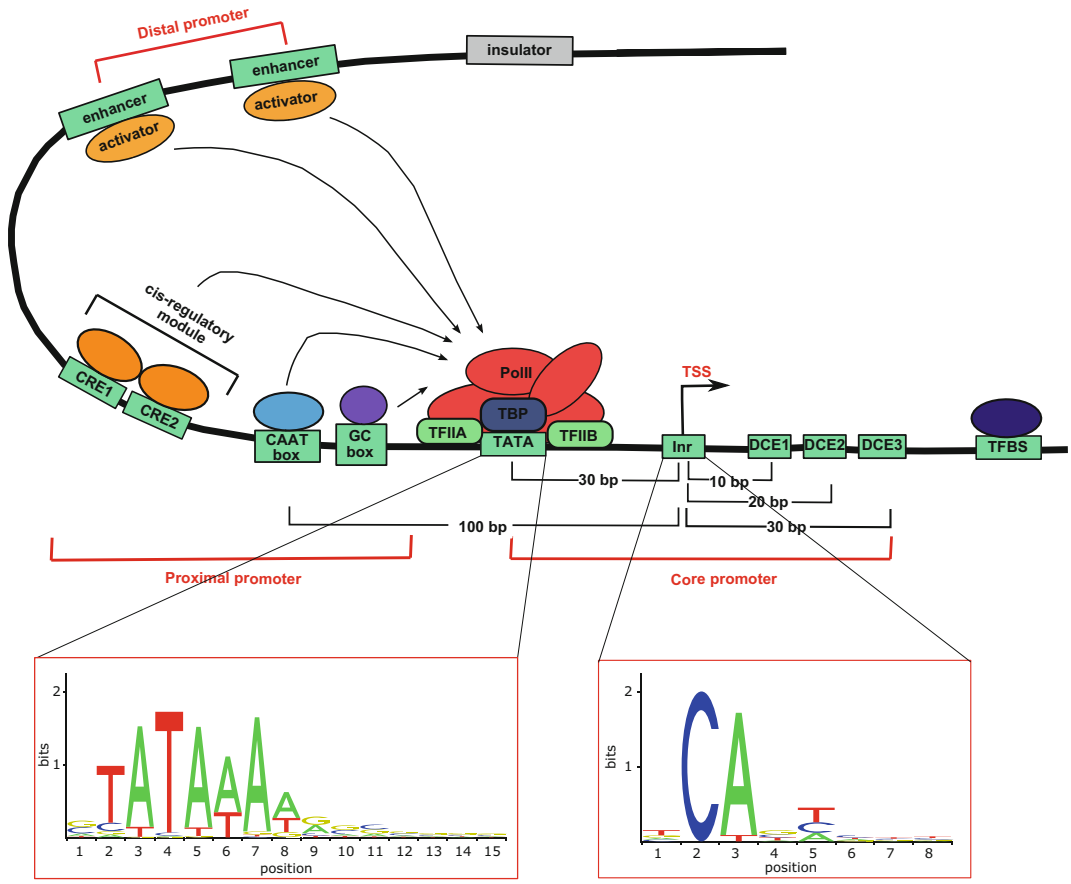


Fig. 1 Summary of regulatory elements and basal transcriptional machinery at a eukaryotic RNA Polymerase II promoter in an open chromatin region. Boundaries between accessible chromatin states are marked by insulators. The region around the Transcription Start Site (TSS) is divided into Core Promoter, Proximal Promoter and Distal Promoter. The Core Promoter contains the regulatory elements necessary to recruit Polymerase II and basal transcription factors (e.g., TFIIA and TFIIB) necessary to activate transcription, as well as the TATA box element (TATA), the Initiator site (Inr) and the downstream core element (DCE). The location of such elements with respect to the TSS is shown here as boxes and their sequence patterns (for the TATA box and Inr only) as logos from the Jaspar database [12]. Some more sequence-specific transcription factors bind to some other sequence elements, such as the CAAT- and the GC box, other Transcription Factor Binding Sites (TFBSs) in the proximal promoter or in enhancer regions. TFBSs can occur in clusters to form cis-regulatory modules (CRMs). Proximal and distal regulatory regions are brought together at TSSs to control the transcription of target genes

promoter predictions were developed to improve the genome annotation when experimental support was not available [7].

The main goal of a promoter recognition algorithm is the computational identification of genomic regions corresponding to 5′ ends of genes in a fast and reliable way, and based on the idea that promoter regions differ in several features (sequence, context, structure) from other genomic features, such as exons, 3′UTRs and intergenic regions [7].

Promoter features can be sequence signals at core-promoter elements and TFBSs, or large-scale features such as CpG islands, k-mer frequency, DNA structure, TFBS density, nucleosome binding, and chromatin modifications. Methods for promoter recognition can be *discriminative*, aiming at finding the optimal classification boundary between promoters and nonpromoters based on some selected features, or *generative* and describe the generative process of the signal. Typical discriminative models use experimentally identified promoter regions or TFBSs from databases as training set for Artificial Neural Networks (ANNs) or support vector machines (SVMs) in order to differentiate promoters from nonpromoter regions. Generative models instead learn signals of promoter elements and/or distance between binding sites from experimentally identified promoters, and apply it to find other regions that score well against the model [7].

Early 1990s' computational methods for promoter prediction combine several sequence patterns (TATA box, Inr, DPE, and BRE motifs) to classify promoter regions versus other genomic sequences [8–10]. Binding specificity is characterized, either by consensus sequences that is, giving the most preferred base at each site position within a binding site, or by Position Weight Matrices (PWMs), which assign a weight to each nucleotide at each position of a putative binding site. New binding sites are scored according to the sum of the scores of the individual positions from the PWM model. Maintained collections of PWMs include TRANSFAC [11] and JASPAR [12]. Methods based on consensus sequences and PWMs might give poor results due to the fact that TFBSs are typically short (5–15 bp long), degenerate and several hits of their consensus/model sequence can be found quite often along the genome just by chance. It became clear during the years that most promoters only have one or a few of the patterns described above, and that some patterns are only found in a small proportion of vertebrate promoters. Therefore it became possible to describe some functional groups of promoters in great detail from TFBS consensus sequences, but the false discovery rate remained high when attempting to detect core promoters genome wide [2, 7].

The late 1990s are signed from advances in algorithms or strategies for pattern finding: promoter prediction methods are not based anymore only on a collection of putative binding sites, but the so-called *context features*, i.e., k-mer content extracted from DNA sequences of promoters, are incorporated in both generative and discriminative models [7]. These algorithms are inspired by linguistic and are based on the rationale that promoter and nonpromoter regions differ in their word content. K-mers may correspond to known biological signals (e.g., TATA box), but they might also correspond to yet unknown promoter signals. PromoterInspector [13] and Promoter2.0 [14] are tools which use k-mers with variable gaps or wildcards to distinguish promoters from nonpromoters. For a comprehensive list refer to [1] and [7].

Since 2001, with the first genome projects and the sequencing efforts of the human genome, people realized that promoter recognition algorithms lack sensitivity and specificity when applied genome wide [7]. The observation that promoter features can be so diverse between different promoter subclasses changed the perspective by which computational algorithms looked at promoter prediction. In particular, CpG islands, regions of vertebrate genomes defined primarily by the lack of methylation at CpG doublets, were observed to be a large-scale signal present in about 70% of the human promoters [2], and gained more and more interest. Also, classifiers that analyze CpG-rich and CpG poor promoters separately achieve better sensitivity and specificity as the two classes seem to have different properties at sequence level [7].

In addition, people started appreciating that TFs recognize DNA-binding regions not only at sequence level, but that the conformation and structure of the DNA play a crucial role in guiding DNA-binding proteins to their sites and also influence promoter activity [7]. Hence, structural features, nucleosome positioning preferences, and others started being included, together with sequence patterns, into promoter prediction algorithms. Among them, the Eponine method, one of the best promoter prediction algorithms still nowadays, applies relevance vector machines to capture the most important sequence signals at promoters, represented by a collection of PWMs and positional constraints between them, together with CG content enrichment [15]. In this method category we find McPromoter [16], ProSOM [17] and ARTS [18] superior among others.

More advanced classifiers are ensemble methods, such as PromoterExplorer [19], CoreBoost [20], MetaProm [21] and EnsemPro [22], which combine results from multiple classifiers on multiple features in order to achieve more robust predictions.

Although it had been suggested for several years that epigenomic features, such as histone acetylation, methylation marks, and nucleosome positioning can provide an extra layer of information beyond DNA sequence features, only after 2001 such signals started to be systematically exploited for correctly locating gene promoters in open chromatin regions. Indeed, although promoters differ in their motif content or GC content, properties such as nucleosome-free regions and epigenetic features around the TSS are quite common to all active promoters [23].

Promoter recognition methods also benefit from the search of evolutionarily related sequences by looking for regions of conservation upstream of annotated genes. However, such methods can only identify homologous promoters when sequence conservation is present, but might miss nonconserved promoters [4].

The aforementioned methods predict promoters using various features but the true promoter usage has to be validated in a context-dependent manner. Recently, thanks to the advent of

next-generation sequencing technologies combined with Chromatin Immunoprecipitation (ChIP-Seq) technology [24], and nascent transcript capturing methods, such as Cap Analysis of Gene Expression coupled to NGS sequencing [25, 26] or Global run on sequencing (GRO-Seq) [27], several promoter recognition methods have moved from being purely predictive approaches based on DNA sequence or structure-related features to be data-driven, i.e., to use the observed genome-wide signals, to unravel mechanisms of transcriptional regulation instead of pure sequence features. For example, the epigenetic mark H3K4me3 and the acetylation of H2 have been identified as a hallmark of active promoters, and computational methods for promoter recognition have begun exploiting this information systematically [28, 29].

Comparative methods, as well as prediction methods based only on "first principles" (DNA sequence and structure) do not identify the conditions where certain promoters are activated. Cap Analysis of gene Expression (CAGE) instead allows high-throughput identification of 5′ ends of capped mRNA in a tissue-specific manner, allowing the localization of the associate core promoters, as well as measuring promoter usage in different states [25, 30].

In this chapter we will focus on the identification of genome-wide signals from the CAGE technology and their importance in promoter recognition. Therefore, in the following we will introduce the CAGE technology and the different FANTOM Consortiums in detail, as well as the algorithms for reliable CAGE peak recognition. Subheading 3 describes in detail the steps of the in silico analysis of CAGE data, focusing on the DPI method for CAGE signal recognition [31], and the PROmiRNA software [32], for miRNA promoter predictions. Subheading 3.7 compares the two methods for the specific task of miRNA promoter recognition.

1.1 The CAGE Techniques and the FANTOM Consortium

Cap Analysis of gene Expression (CAGE) allows the identification of transcriptional starting points genome wide by sequencing 5′ ends from full-length cDNA libraries and mapping back those sequences to the genome, thus determining regions corresponding to active promoters of coding and noncoding transcripts, as well as active enhancers. In detail, in its first version the method uses cap-trapper full-length cDNAs to attach linkers to their 5′-ends. This is followed by cleavage of the first 20 base pairs by class II restriction enzymes, PCR, concatamerization and cloning of the CAGE tags. Sequenced CAGE tags mapped to the genome are then used to identify the TSSs of annotated or novel transcriptional units specific to each tissue, cell or condition, as well as the analysis of differential promoter usage [25]. Compared to RNA-seq or microarray, CAGE allows the separate analysis of multiple promoters linked to the same gene. In fact, most genes have more than one TSS and the regulatory inputs or TFs that determine TSS choice and activity in a particular tissue are diverse.

FANTOM stands for the Functional Annotation Of Mammalian genome and is an international research consortium founded in the year 2000 to assign functional annotations to the full-length complementary DNAs (cDNAs) that were collected during the Mouse Encyclopaedia Project at RIKEN. Research at FANTOM has proceeded in three phases. FANTOM began with the establishment of an annotation pipeline that developed and expanded quickly into more transcriptome and functional analysis.

Only in the second phase, with the FANTOM3, the Consortium started using the CAGE technology to study transcriptional initiation genome wide. FANTOM3, which focused on identifying transcribed components of mammalian cells, improved the estimation of the total number of genes and their alternative transcript isoforms in both human and mouse, and revealed that about 70 % of the genome is transcribed as RNA, confirming the existence of thousands of noncoding RNAs (ncRNAs) [30]. This led us to gain new insights into how transcription initiation works and to revise central dogmas of Molecular Biology, projecting us into an "RNA world," whose functional implications are still partially to be discovered.

More in detail, in the FANTOM3 145 mouse and 41 human libraries are analyzed; CAGE tags of size 20–21 bp are derived from transcripts sequenced in proximity of the cap site. Amplified tag libraries contain between 50,000 and 100,000 tags. Clones are sequenced with Sanger sequencing techniques and their unique mapping positions on the genome identify putative TSSs. Clusters of overlapping tags define promoter strength and shape. Based on these data, Carninci et al. [30] classify tag clusters into different shapes, ranging from single-peak TSSs to broad or bimodal tag distributions, corresponding to different promoter contexts [30]. Given that the data constitute a quantitative profiling of relative promoter usage across tissues and cell types, it is observed that alternative promoter usage is higher than expected, with the majority of protein-coding genes having two or more alternative promoters, especially in brain tissues [33, 34].

In the era of high-throughput sequencing, the FANTOM4 Consortium develops deepCAGE (CAGE followed by deep sequencing of the tags). The CAGE method is adapted to the 454 Life Sciences (Roche) GS20 sequencer and the main difference consists in the fact that cloning is no longer necessary, as after amplification and concatenation the tags can be directly sequenced, generating libraries of up to two million tags [26]. The focus of the FANTOM4 also shifts from the recovery of transcribed elements to the integration of such components into biological networks for functional analysis in specific contexts such as Leukemia or monocyte differentiation [35].

In FANTOM5, the HeliScopeCAGE technique is introduced, an adaptation of CAGE to single molecule sequencing with the revolutionary HeliScope Single Molecule Sequencer measurements [36].

Such technique opens the door to detailed analysis of gene expression levels and rare cell populations, providing the community with a promoter expression atlas where expression profiles are determined at an unprecedented depth and high precision [31]. Unlike earlier sequencers, the Heliscope Sequencer does not employ polymerase chain reaction (PCR) amplification to multiply DNA fragments, a process which can introduce biases into data, instead the reverse-transcribed DNA is sequenced directly, enabling direct, high-precision measurements [36]. The latest CAGE dataset from FANTOM5, includes 573 human and 128 mouse primary cell samples, 152 human post-mortem samples, 271 mouse developmental tissue samples and 250 different cancer cell lines sequenced to a median depth of four million mapped tags per sample [31].

Given that promoter-distal regulatory regions such as enhancers are essential in controlling time- and cell-specific gene regulation, and that they have been shown to be often transcribed by PolII, producing so-called eRNAs, FANTOM5 CAGE data are also used to detect actively transcribed enhancers. Based on the data from hundreds of cell lines and tissues, Andersson et al. identify more than 40,000 enhancer regions, together with their activation levels across human tissues, marked by the presence of bidirectional capped transcripts [37].

1.2 Methods for the Analysis of CAGE Data

Either CAGE data are used to locate active promoters of known genes, or to identify start sites of novel transcripts, or to locate active enhancers, appropriate computational methods are needed to analyze the NGS data and detect transcriptional events above noise. As CAGE tags tend to be clustered, with more or less signal, at active transcription sites, the task of identifying signal-enriched regions is similar to the peak calling step in the analysis of ChIP-seq data. Peak calling methods, such as HOMER [38] can be applied to identify peaks corresponding to initiation events in CAGE data. However CAGE peaks/clusters possess specific features that distinguish them from ChIP-Seq data, so that dedicated methodologies have been developed in the past few years specifically for the analysis of CAGE data. Initial studies of CAGE dataset have employed basic methods for processing mapped CAGE tags and identifying CAGE TSSs [30, 34]. Active promoters have been reconstructed by means of different clustering approaches based either on the proximity of individual TSSs or their density [39]. With the increase of sequencing depth, in order to perform TSS-centered differential expression analysis, normalization approaches, and explicit noise modeling have been introduced [40]. In this chapter we will focus on the Decomposition-based peak identification (DPI) method, especially designed for FANTOM5 CAGE data and methodology of choice for most of the FANTOM5 subsequent analysis (Subheadings 3.1–3.3). As CAGE data can locate both coding and noncoding transcript TSS, we describe the PROmiRNA software, especially

designed for the challenging task of identifying miRNA promoters from either FANTOM4 or FANTOM5 data (Subheadings 3.4–3.6). Although PROmiRNA has several analysis steps in common with other CAGE analysis methodologies (see below), its underlying statistical model is optimized for the de novo detection of lowly expressed TSSs, and this makes it particularly suitable to detect both intergenic and intronic transcription initiation events of transient miRNA primary transcripts, which undergo rapid processing by the Drosha enzyme in the cell nucleus, yielding sparse CAGE tag coverage around true TSSs.

Although not described in this chapter, a relatively new software package which integrates several CAGE analysis workflows is the R/Bioconductor package CAGEr [41]. CAGEr implements various methods for CAGE data processing, it provides several normalization strategies, easy access to published CAGE dataset in several organisms and introduces a novel method for detection of differential TSS usage and promoter shifting in different tissues/contexts.

The main steps of the analysis of CAGE data, common to several CAGE analysis pipelines, can be summarized as follows:

1. *Library preparation and sequencing.* The CAGE technology has evolved during the last 10 years and the different protocols for library preparation and sequencing have been discussed above.
2. *Read mapping.* The first step in the analysis of CAGE data is the mapping of the CAGE tags back to the genome. Depending on the sequencing protocols, different mapping tools and strategies have been employed for this task. In FANTOM3, still based on Sanger sequencing, CAGE tags of 20–21 nt are aligned on the genome using BlastN [42]. Tags mapping on multiple genomic regions are not used for subsequent analysis and only best alignments of at least 18 nt are kept for subsequent analysis [30]. The data from the FANTOM4 are mapped with different tools: for example Valen et al. [34] use BLAST/V alignment programs and only the longest matches without mismatches are selected, whereas matches shorter than 18 nt are discarded and multi-mapping CAGE tags are included according to a computed posterior probability for each mapping location [43]. Balwierz et al. [40] use the same strategy for multi-mapped reads, but CAGE tags are aligned with the Kalign2 alignment tool, which maps tags in multiple passes [44]. Specifically, tags that do not map perfectly to the genome are given as input to a second step, where they are mapped with at most one mismatch or event to a third step, where they are mapped allowing indels. In FANTOM5, sequenced Heliscope reads have different lengths, without associated base quality values and high sequencing error rates (up to 5%) [31]. After removal of reads corresponding to ribosomal RNA, all remaining CAGE reads are mapped to the genome using the probabilistic mapper Delve, which places reads

to single positions in the genome according to a computed probability of being a true match from a Hidden Markov Model [31, 45].

3. Analysis of CAGE tag peaks: the most important step in CAGE data analysis is the identification of regions of significant CAGE tag signal, equivalent to clusters of overlapping tags or highly dense tag regions. Most genes are transcribed in different isoforms that use different TSSs arranged typically in local clusters spanning regions from few to over 100 bps. Depending on the application, different methods deal differently with the question: what defines a Tag Cluster (TC)? On the FANTOM3 data, Carninci et al. grouped individual CAGE tags that had identical sequences into a representative CAGE tag [30]. Representative CAGE tags with the same starting position define a CAGE tag-defined transcriptional start site (CTSS) (*see* **Note 1**). As the focus of the FANTOM3 is to characterize all distinct transcription initiation events, the authors simply cluster CAGE tags whose genomic mapping overlap by at least 1 bp in Tag Clusters (TCs) (*see* **Note 2**). The PROmiRNA methods (extensively described in Subheading 3.4) defines TCs in a similar way, except that it joins together in the same cluster also tags which do not overlap with each other, but are closer than 20 bp from one another. This allows recovering much more sparse tag signal as the one generated by transient microRNA primary transcript TSSs. More sophisticated approaches to define tag clusters include the Paraclu algorithm [39] and the method from Balwierz et al. [40]. The Paraclu algorithm is based on the observation that core promoters do not have a single TSS, but a distribution of initiation sites clustered at multiple scales as a consequence of multiple regulatory processes. The Paraclu algorithm aims at finding these clusters, at multiple scales, among transcription initiation events observed at specific locations in the genome by finding maximal scoring segments with a density of more than d events per nucleotide. Afterwards, an inhomogeneous HMM is learned from dominant TSS (clusters associated to at least five transcription initiation events) to determine sequence preferences of TSSs and apply the trained model to discover new TSSs genome wide. The approach in Balwierz et al. takes into account expression profiles of TSSs across different samples and finds clusters of nearby co-expressed TSSs by using Bayesian hierarchical clustering. More in detail, their goal is to define Transcription Start Clusters (TSCs) of contiguous TSSs such that expression profiles of clustered TSSs are the same among tissues up to measurement noise. In FANTOM5, given the much higher sequencing depth compared to previous studies, a simple clustering procedure such as the one used by Carninci et al. or PROmiRNA, would generate very long clus-

ters. In DPI [31] the authors first group CTSSs (*see* **Note 1**) from different tissues into tag clusters according to the procedure from Carninci et al., and then try to separate distinct transcriptional events inside each cluster by means of Independent Component Analysis (ICA, *see* **Note 3**). All steps of DPI analysis are described in detail in Subheading 3.1.

4. TSS-centered differential expression: to quantify the expression of individual TSSs and enable comparison between samples, raw tag counts have to be normalized. Many studies based on deepCAGE use the number of tag per million (TPM) values, which is the simplest normalized measure also used on the FANTOM5 data and widely used in other high-throughput sequencing tools [46]. One of the more sophisticated approaches [40] is based on the observation that the reverse cumulative distribution of the number of tags per TSS follows a power-law distribution with a very good approximation. Therefore, CAGE tag counts across different samples are transformed to match a common reference power-law distribution. Normalization can be performed at promoter level (cluster tag counts are normalized) or at individual TSS level. For example, in order to take into account substantial differences in the total numbers of read counts, PROmiRNA counts the number of overlapping 5′-ends at each bp position and performs per-position quantile normalization across tissues, inspired by normalization methods for microarray analysis [14, 32]. After applying any of the aforementioned normalization procedures, normalized CAGE tag counts can be used to perform differential expression profiling at single TSS or promoter level.
5. Assignment of tag clusters to genes: unnormalized or normalized CAGE tag clusters (also referred to as CAGE peaks) from a CAGE analysis tool can be used to define transcription start sites of novel transcripts or assign active promoters to known genes. When trying to assign CAGE peaks to known annotation, one needs to define a distance cutoff to assign a peak to the closest gene. In [31] the authors assign a peak to a known transcript if its 5′ end is within 500 bp from the defined peak. Such distance cutoff is arbitrary and depends on the research application. Obviously it can happen that more than one peak is assigned to the same gene, or the same peak is within a certain distance to more than one transcript. Solutions to such situations differ according to the research motivation.

2 Materials

The purpose of this chapter is to give details on practical aspects regarding the computational analysis of CAGE data and usage of the DPI and PROmiRNA software, with emphasis on the CAGE

peak calling step. As the analysis with these two software is based on raw or preprocessed data from the FANTOM4 and/or the FANTOM5 Consortium, in this paragraph we provide some details about data sources and specific data files. The PROmiRNA software was originally designed to recognize miRNA human promoters from FANTOM4 deepCAGE data, but can also be applied to data from the FANTOM5 in both human and mouse.

FANTOM4 raw data, mapped data, as well as tag count files and processed files containing the annotation of the detected promoters can be downloaded at the following link: http://fantom.gsc.riken.jp/4/download/Tables/.

To facilitate data interpretation and integration, as well as navigate through the FANTOM4 dataset, Severin et al. developed the EdgeexpressDB database [47]. Such source not only collects alternative promoters and gene expression patterns across tissues, but also provides a regulatory network view of the data, including regulating factors and microRNAs.

All FANTOM5 data, including visualization and web-based tools, different data access points, CAGE raw data (fasta sequences), mapped CAGE data (bam files), as well as processed data from human and mouse samples, including position and expression of CAGE peaks, are precomputed and available on the following website http://fantom.gsc.riken.jp/5/. In particular, two specialized tools allowing exploring relations between the data, namely ZAMBU, which is useful if one wants to investigate the relationship between CAGE tag distributions and expression profiles [31, 48], and STARR, a semantic tool to explore relationships between promoters, genes, samples, and TFBSs [49].

When interested in sample-specific CAGE peak information, one can download the corresponding CAGE tag starting site file (ctss file) from the aforementioned website, and then use DPI to identify CAGE peaks based on such input file. The GM12878 ctss files used as input to DPI for the example shown in this paper and the GM12878 CAGE bam file used as input for PROmiRNA are downloaded from http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.cell_line.hCAGE/.

The DPI software (extensively described in Subheading 3.1) is available for download on Github http://github.com/hkawaji/dpi1/.

The PROmiRNA software (extensively described in Subheading 3.4) can be freely downloaded at http://promirna.molgen.mpg.de together with the *external_data.tar.gz* directory.

3 Methods

This section will focus on computational methods for peak calling and promoter identification from CAGE data. First, we will introduce the Decomposition-based Peak Identification (DPI) method,

especially designed for FANTOM5 CAGE data and applicable to both promoter and active enhancer recognition [31]. Second, we will introduce the PROmiRNA software, especially designed for miRNA promoter recognition from both FANTOM4 and FANTOM5 CAGE data [32]. Third, we will show as an example, the results from applying both DPI and PROmiRNA to identify miRNA promoter of expressed miRNAs in the Gm12787 B-lymphoblastoid cell line.

3.1 The DPI Algorithm

The main steps of the DPI algorithm are illustrated in Fig. 2a and described in detail below. The intermediate output files generated from each step are schematically described in Fig. 2b.

1. *Input*. The input to the DPI algorithm is represented by one or more ctss files (*see* **Note 1**) from tissue-specific mapped tags (Fig. 2a, Step 1). These correspond to CAGE profiles at individual biological states. Only ctss supported by two or more CAGE 5′-end reads in a single profile are used by DPI.
2. *Definition of CAGE tag clusters (TCs)*. Cage tags are clustered based on proximity to each other. Input ctss from different tissues are first merged to produce an accumulated CAGE profile (Fig. 2a, Step 2). Selected ctss, supported by no less than two reads are grouped together into the same cluster if they are within 20 bp from each other.
3. *TC Decomposition*. Due to higher sequencing depth compared to previous CAGE dataset, step 2 may produce very long tag clusters, which might contain several transcription start sites. To correct for this, DPI uses Independent Component Analysis (ICA, *see* **Note 3**) on clusters wider than 50 bp (or with a coverage higher than 50 tag counts), in order to decompose the overall signal into distinct TSS signals (Fig. 2a, Step 3). Within each cluster, ICA infers the number of underlying signals which correspond to 95% of the signal variance (and up to a maximum of 5 independent components) and represent individual ctss intensity patterns.
4. *Scaling*. The signal in each inferred independent component is downscaled by 10% of the intensity of its highest ctss. This step is performed in order to avoid detecting "too much signal" in proximity of very active TSSs, where a continuous but modest read coverage is observed (Fig. 2a Step 4).
5. *Smoothing*. At this stage, DPI applies a Gaussian kernel to smooth each independent signal component in each cluster and detect candidate peaks where the signal is higher than the median of each signal component (Fig. 2a Step 5).
6. *Merging*. Inferred peaks are merged if they overlap with each other (Fig. 2a Step 6).

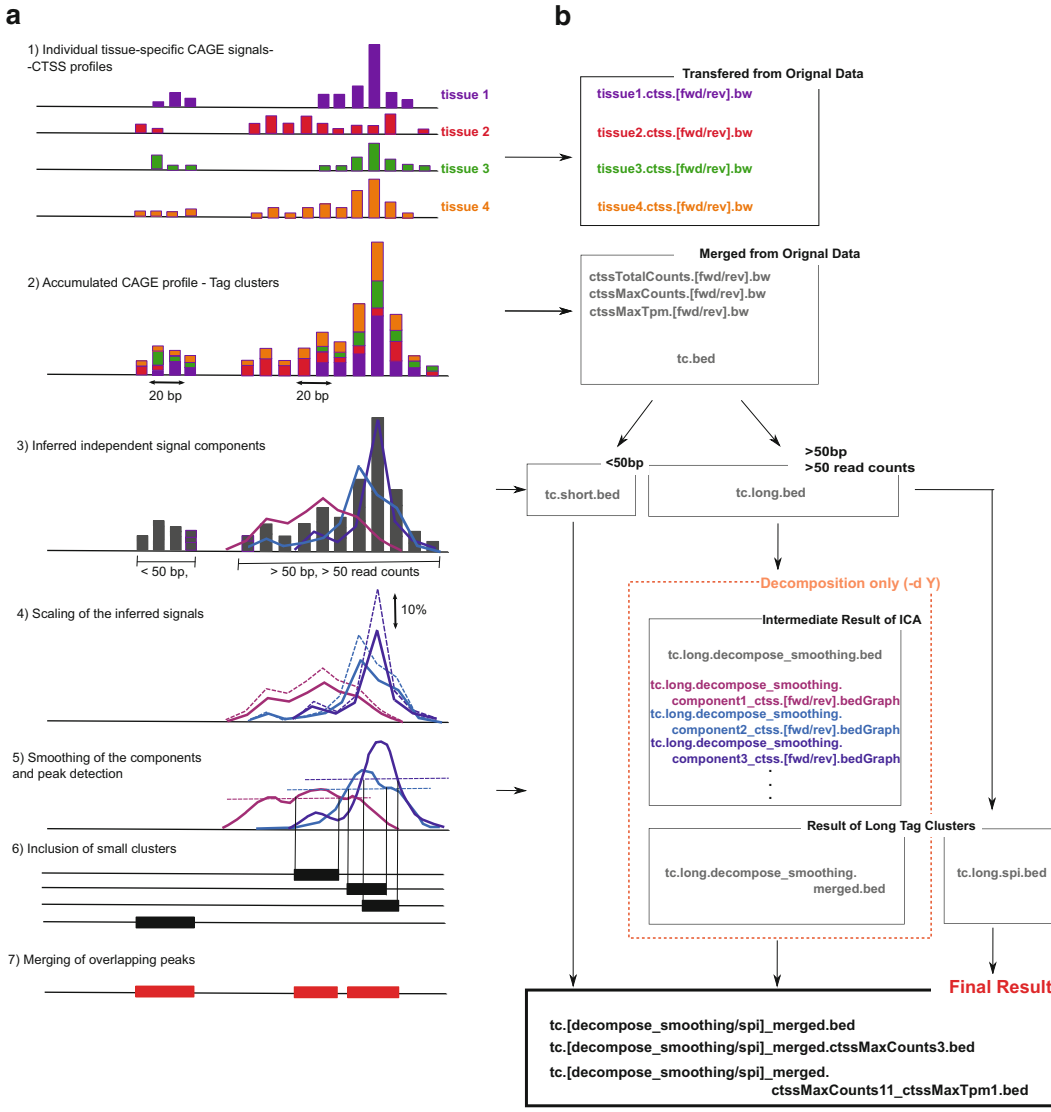


Fig. 2 (**a**) Different steps of the DPI workflow, from parsing of the input ctss files to the final CAGE peaks; (**b**) Intermediate and final output files from the DPI pipeline. Adapted by permission from Macmillan Publishers Ltd: Nature [31], copyright 2014

7. *Output*. Finally, aggregated peak regions are reported together with short clusters (<50 bp) which were not selected in Step 3 for ICA processing (Fig. 2a Step 7). In order to minimize the fraction of peaks mapped to internal exons and enrich for promoter regions, DPI applies a tag threshold to define robust and permissive output peaks, based on the assumption that genuine TSSs have a higher number of 5′ tags starting at the same position than random regions along the transcript. A fold enrichment of at least 2.0 over random regions (equivalently peaks with a single ctss supported by at least 11 reads) defines

the robust cutoff, while the more permissive cutoff corresponds to a fold enrichment of 0.7 (single ctss supported by at least three reads in at least one CAGE profile). Both "robust" peaks and "permissive" peaks are reported by DPI.

Although it is not a part of the DPI pipeline, the detected peaks can be used to quantify tissue-specific expression of transcription start regions. In [31] the authors, after applying the DPI pipeline to the FANTOM5 data, count the number of tags whose 5′ ends start within the boundary of a "robust" peak in that tissue. In order to compare TSS activity between tissues, read counts are transformed into TPM (tag per million) values and normalizing factors are estimated using the relative log expression (RLE) method implemented in the EdgeR R package [50].

3.2 Practical Usage of the DPI Software

DPI runs on the Unix/Linux system with Grid Engine without installation. If Grid Engine is not available, one can still use it without the decomposition step (see below). Before using it, one should insure that the following languages/software are available on the system:

Ruby (https://www.ruby-lang.org)

R (http://cran.r-project.org/) and the R package fastICA (http://cran.r-project.org/web/packages/fastICA/index.html)

Command line bigWig tools (http://hgdownload.cse.ucsc.edu/admin/)

BEDtools (https://github.com/arq5x/bedtools2)

Importantly, one should declare these tools in the system environment.

Download DPI from github using command line:

```
> git clone https://github.com/hkawaji/dpi1.git
```

A packed shell script: *DPI_DIR/dpi1/identify_tss_peaks.sh* is included in the package. One can view detailed package information, parameters, and output explanation by running this script:

```
> DPI_DIR/dpi1/identify_tss_peaks.sh
```

Before peak calling, prepare the following input files:

Chromosome size file in BEDTools (should be automatically provided):

```
BED_DIR/genomes/YOUR_SPECIES.genome
```

ctss files in bed format downloaded from FANTOM http://fantom.gsc.riken.jp/5/datafiles/latest/

After specifying the output folder, simply run:

```
> DPI_DIR/dpi1/identify_tss_peaks.sh -g genome -i CTSS FILE -o OUTPUT_DIR -d Y/N
```

where -d is an optional parameter and is set to "N" by default. When –d is specified to Y, the decomposition step will be performed (see DPI algorithm from Subheading 3.1). In general, DPI takes as input multiple .ctss files: one can simply put all .ctss files in one folder and set the input parameter as:

```
-i'CTSS FOLDER/*.ctss.bed.gz'
```

3.3 The DPI Output

After running DPI as shown above, the output consists of three folders: *outCounts*, *outTpm* and *outPooled*. *outCounts*, and *outTpm* contain bigwig files for each individual input ctss file, with the value being tag counts and tags per million (TPM), respectively. Tags on forward strand (fwd) and reverse strand (rev) are reported separately. The *outPooled* folder contains the following result files for the intermediate steps illustrated in Fig. 2b:

bigwig files correspond to pooled individual ctss files (Fig. 2b Step 2):

ctss.[MaxCounts/MaxTpm/TotalCounts].[fwd/rev].bw

bed files for all/long/short tag clusters (Fig. 2b Step 3):

tc.[-/long/short].bed.gz

If the decomposition parameter –d is specified, bed/bedGraph files from decomposition step will be generated. (Fig. 2b Step 4/5):

```
tc.long.decompose_
smoothing.*.[bed/bedGraph].gz
```

The merged peak files (Fig. 2b Step 6/7):

peaks with robust threshold, i.e., more than 10 ctss tags and no less than 1 TPM (Fig. 2b Step 6/7):

```
tc.[decompose_smoothing/spi]_
merged.[ctssMaxCounts11/ctssMaxCount11_
ctssMaxTpm1].bed.gz
```

peaks with permissive threshold, i.e., more than 2 ctss tags (Fig. 2b Step 6/7):

```
tc.[decompose smoothing/spi]_
merged.ctssMaxCounts3.bed.gz
```

3.4 The PROmiRNA Software

Due to fast Drosha cleavage in the nucleus, miRNA primary transcript TSSs are hard to identify from sparse CAGE tag coverage with conventional methods. The PROmiRNA algorithm combines CAGE tag counts and several promoter sequence properties into a statistical model, in order to identify miRNA promoter at high sensitivity, while distinguishing them form transcriptional noise. The main steps of the PROmiRNA methodology are illustrated in Fig. 3a and described in detail below.

1. *Input*. The input to PROmiRNA is represented by more than one tag alignment file, one for each tissue, in bed format. These are used to build the tissue-specific CAGE tag profiles up to 50 kb upstream of annotated miRNA precursors. Such

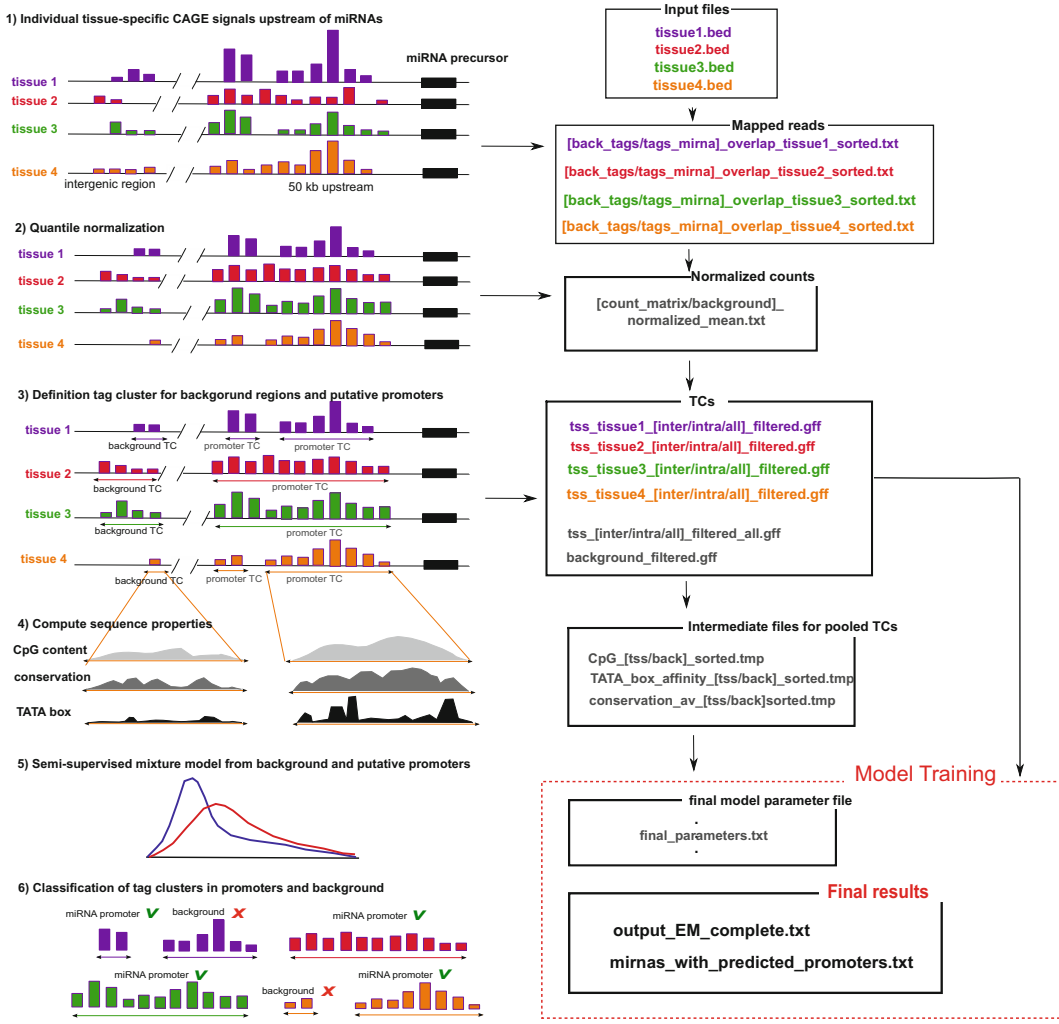


Fig. 3 (**a**) Different steps of the PROmiRNA workflow, from parsing of the input bed file to the final list of CAGE peaks corresponding to miRNA promoters; (**b**) intermediate and final output files from the PROmiRNA pipeline

profiles represent CAGE read coverage (or tag counts) at 1 bp resolution (Fig. 3a, Step 1).

2. *Tag-count normalization.* In order to make tag counts comparable across tissues, row counts at each bp position are quantile-normalized. In detail, position-specific tag counts from each sample are transformed to match a common reference distribution, randomly chosen from the available libraries. Normalized tag counts can be interpreted as expression values at TSS level at 1 bp resolution (Fig. 3a, Step 2).
3. *Cage Tag clusters (TCs)—Identification of putative promoter regions.* CAGE tags are grouped into clusters if the overlap between their genomic coordinates is at least 1 bp (or they are within a distance of 20 bp from each other). Normalized tag

counts inside each cluster are summed up. Tag clusters whose genomic coordinates overlap with the TSSs of other annotated transcripts other than miRNA host genes are filtered out (not shown). Tag clusters located in the 50 kb region upstream of miRNA precursors which do not overlap any known TSS define putative miRNA promoter regions in a tissue-specific manner. Tag clusters located in randomly selected intergenic regions are defined in the same way and are interpreted as non-promoters, therefore assumed to represent background noise (Fig. 3a, Step 3).

4. *Sequence properties of tag clusters.* The statistical model of PROmiRNA computes a prior probability for each TC of being a real promoter, based on the following sequence properties computed in the 1000 bp genomic regions around the center of each defined TC (Fig. 3a, Step 4): Normalized CpG content, computed as described in [32];

   Average PhastCons conservation score across on a 46-way vertebrate alignment downloaded from the UCSC Genome Browser (*see* **Note 5**); Affinity for a TATA box protein, computed by means of the TRAP program (cite) and based on the position-specific scoring matrix (pscm) downloaded from the Jaspar database (http://jaspar.genereg.net/, Jaspar ID: MA01082);

   Genomic proximity score of the defined TC to the miRNA precursor.

5. *PROmiRNA's mixture model.* Pooled TCs from all tissues, together with their normalized tag counts and computed sequence properties, are fed into a semisupervised mixture model which automatically learns, through an EM algorithm, the optimal separation between TCs corresponding to promoters and TCs corresponding to background. TCs from random intergenic regions are interpreted as "exact" negative examples by the model (supervised part), while TCs upstream of miRNAs are nonlabeled examples (unsupervised part) which might either belong to the miRNA promoter class or to the background noise (Fig. 3a, Step 5).

6. *MiRNA promoter assignment.* TCs upstream of miRNA promoters are classified as miRNA promoters, if the computed posterior probability from the model is higher than 0.5, otherwise they are classified as background. The main output of the PROmiRNA software is a list of predicted promoters, for each miRNA gene, together with their genomic coordinate (Fig. 3a, Step 6).

3.5 Practical Usage of the PROmiRNA Software

PROmiRNA runs on every Linux/Unix environment. Before using it, one should make sure that the following languages and tools are available on the system:

python 2.7 (https://www.python.org/download/releases/2.7/).

PROmiRNA does not run with Python 2.6 or Python 3.x

R >=2.12.1 (http://cran.r-project.org)

Perl >= 5.12 (http://perl.org/get.html) and BIO:Graphics perl module (http://search.cpan.org/dist/Bio-Graphics-2.34/)

BEDtools (http://code.google.com/p/bedtools)

cd-hit (http://weizhong-lab.ucsd.edu/cd-hit/download.php)

ANNOTATE 3.04 (http://trap.molgen.mpg.de/download/TRAP/ANNOTATE-3.04.01.tar.gz)

After unzipping and placing the *external_data* directory into the *PROmiRNA* main directory we can have a brief look at the PROmiRNA subdirectories structure. The PROmiRNA folder contains four subdirectories:*PROmiRNA/src*, it contains all necessary code to run PROmiRNA*PROmiRNA/miRBase*, it contains miRNA annotation files downloaded from the miRBase database [51]. *PROmiRNA/external_data*, it is further divided into two subdirectories, *bed_files*, where input files to the software (tag alignments) in bed format should be placed (*see* **Note 4**), and *Phastcons*, where chromosome-wise PhastCons conservation files should be placed (*see* **Note 5**). This directory contains also other data files necessary for the analysis: genome files for the organism under study (a chromosome size file, e.g., *hg19.chrom.sizes* (*see* **Note 6**), a fasta file for the whole genome, e.g., *hg19*.fa and its corresponding index file, e.g., *hg19.fa.fai* (*see* **Note 7**)); the annotation of the repetitive regions for the organism under study (e.g., *hg19_repeats.bed*, *see* **Note 8**); a gtf file containing Ensembl gene annotation (e.g., *Homo_sapiens.GRCh37.66 gtf*, *see* **Note 9**)*PROmiRNA/Data*, it contains all intermediate output files and it is divided into four subdirectories, namely *gff_files*, where all tissue-specific, as well as pooled TCs are stored, *fasta*, where fasta sequences of TC regions are stored, *background*, where TCs and computed sequence properties for the background TCs are stored, *overlap_files*, which stores intermediate overlap files between miRNA genomic coordinates and CAGE tags in different tissues and *matrix_file*, where intermediate matrices of read counts before and after quantile normalization are stored. This *Data* directory contains many other intermediate files, the most important being discussed in the next session. The PROmiRNA software can be used in two different modes, depending on the application:

1. Testing mode. Given a set of genomic regions in gff format, test if they contain one or more miRNA promoters, based on a pretrained PROmiRNA model. After defining the output directory where all intermediate files and final results will be placed, run

   ```
   > python test_new_regions.py <out_dir> <input_file>
   ```

   For example: given the *test_regions.gff* provided inside the PROmiRNA directory, and setting output_dir = *test_regions*, run:

```
> python test_new_regions.py test_regions
test_regions.gff
```

In the output directory, the main result files from this command are:

output_EM.txt, it contains promoter regions, as well as background TCs, with their respective genomic coordinates, normalized tag counts, values of the computed features and prior and posterior probabilities from the model.

miRNA_predicted_promoters.txt, it reports, for each miRNA in the input file, genomic coordinates of the predicted promoter TCs, together with normalized CAGE tag counts and genomic distance of the TC from the miRNA precursor.

2. Training mode. The original PROmiRNA model is trained on FANTOM4 data on the human assembly hg19. If you want to use PROmiRNA with new CAGE libraries (e.g., FANTOM5 or Encode data) or on a new assembly / organism, we strongly suggest to re-train the PROmiRNA model.

   After downloading the necessary files (*see* **Notes 4–9**), retrieve the miRNA annotation from miRBase [51]:

   ```
   > python download_mirbase_annotation.py <org> <v>
   ```

   Where *org* is the official three-letter code for the organism identifier (e.g., *hsa* for human) and *v* indicates the miRBase release number. This command downloads the following annotation files in the PROmiRNA/miRBase directory:

   [org].gff2, a gff file containing the genomic coordinates of all precursor miRNAs for a specified organism *org*;

   miRNA.txt, annotation file containing information about each mature miRNA (e.g., accession, species, genomic sequence..);

   mirna_context.txt, annotation of the genomic context of a miRNA (intergenic, intron, exon, 3′ UTR, 5′ UTR)

   The training itself is done via:

   ```
   > python PROmiRNA.py <genome>
   ```

   Where `<genome>` refers to the genome assembly specified for promoter prediction, e.g., hg19.

3.6 The PROmiRNA Output

Although PROmiRNA produces many intermediate files during both testing and training, the most important output files are summarized below and illustrated in Fig. 3b.

1. Files reporting the overlap between CAGE tags and regions upstream of miRNAs / random intergenic regions for each tissue, sorted by genomic position:

   ```
   PROmiRNA/Data/overlap_files/[back_tags/tags_
   mirna]_overlap_<tissue>_sorted.txt
   ```

2. Matrix file of the normalized CAGE tag counts across tissues for both putative TSS positions and random intergenic regions:
   `PROmiRNA/Data/matrix_file/[count_matrix/background]_normalized_mean.txt`
3. TC cluster files for candidate promoter regions (both tissue-specific and pooled across tissues), further distinguished in intragenic, intergenic, and all TCs:
   `PROmiRNA/Data/gff_files/tss_<tissue>_[inter/intra/all]_filtered.gff`
   A similar file is provided for background TCs:
   `PROmiRNA/Data/background/background_filtered.gff`
4. Files of computed sequence properties for putative TSSs and background TCs:
   `PROmiRNA/Data/gff_files/CpG_tss_sorted.tpm, PROmiRNA/Data/gff_files/TATA_box_affinity_tss_sorted.tmp, PROmiRNA/Data/gff_files/conservation_av_tss_sorted.tmp PROmiRNA/Data/background/CpG_back_sorted.tpm, PROmiRNA/Data/backgrouns/TATA_box_affinity_back_sorted.tmp, PROmiRNA/Data/backgoruns/conservation_av_back_sorted.tmp`
5. File listing the final model's parameters learned during PROmiRNA training:
   `PROmiRNA/Data/final_parameters.txt`
6. Files reporting the final promoter predictions (see previous section)
   `PROmiRNA/Data/output_EM_complete.txt, PROmiRNA/Data/mirnas_with_predicted_promoters.txt`

3.7 Case Study: Prediction of miRNA Promoters in Gm12878

To show an example of application of both DPI and PROmiRNA we used both tools to identify miRNA promoters in the B-lymphoblastoid cell line Gm12878. Although DPI is designed to define CAGE peaks genome wide, and not tuned to specifically find miRNA promoters, we can nonetheless assign DPI peaks to miRNA genes by looking at the defined DPI peaks in the 50 kb region upstream of annotated miRNAs.

3.7.1 Application of PROmiRNA to Gm12878 CAGE Data

For PROmiRNA, alignment files in bam format (two biological replicates) for the Gm12878 cell line were downloaded at

http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.cell_line.hCAGE/

For the sake of simplicity we will rename these files to *Gm12878_rep1.bam* and *Gm12878_rep2.bam*. As PROmiRNA

requires input alignments in bed format, the bam file were converted to bed format using the Bedtools:

```
> bamToBed -i Gm12878_[rep1/rep2].bam >
Gm12878_[rep1/rep2].bed
```

The two bed files were placed in the *PROmiRNA/external_data/bed_files* directory

miRNA promoters in the Gm12878 cell line from miRBase version 20 and human assembly hg19 were predicted as follows:

```
> python download_mirnabse_annotation hsa 20
> python PROmiRNA hg19
```

The miRNA promoter predictions were listed in the output file *PROmiRNA/Data/mirnas_with_predicted_promoters.txt*. This file reports the union of predicted promoters from the two replicates. In order to derive a specific and strict list of promoters, and minimize the number of false positives we applied the following constraints:

- only promoter predictions common to the two replicates were retained
- only promoter predictions in open chromatin regions were retained

In order to fulfill the second criterion we downloaded DNAseI hypersensitivity peak sites for the Gm12878 cell line from the ENCODE website http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks/. For the sake of simplicity we will rename this file to *DNAseI_ Gm12878.bed*.

After converting PROmiRNA predicted promoters to gff format using a customized simple script (*promoters_Gm12878.gff file*) we computed the overlap between DNAseI hypersensitivity sites and promoters' genomic coordinates (extended by 100 bp upstream and downstream) by means of the Bedtools:

```
> windowBed -a promoters_Gm12878.gff file -b DNAseI_Gm12878.bed -w 100 -u >
promoters_Gm12878_dnase_validated.gff
```

3.7.2 Application of DPI to Gm12878 CAGE Data

The cell-specific ctss file is downloaded at http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.cell_line.hCAGE/ and given as input file to DPI.

```
> identify_tss_peaks.sh -g human.hg19.genome -i'DATA FOLDER/*.ctss.bed.gz'
-o ./result -d Y
```

The output files of the predicted Tag Clusters genome wide are:

```
tc.decompose_smoothing_merged.ctssMaxCounts11_ctssMaxTpm1.bed (robust cutoff)
```

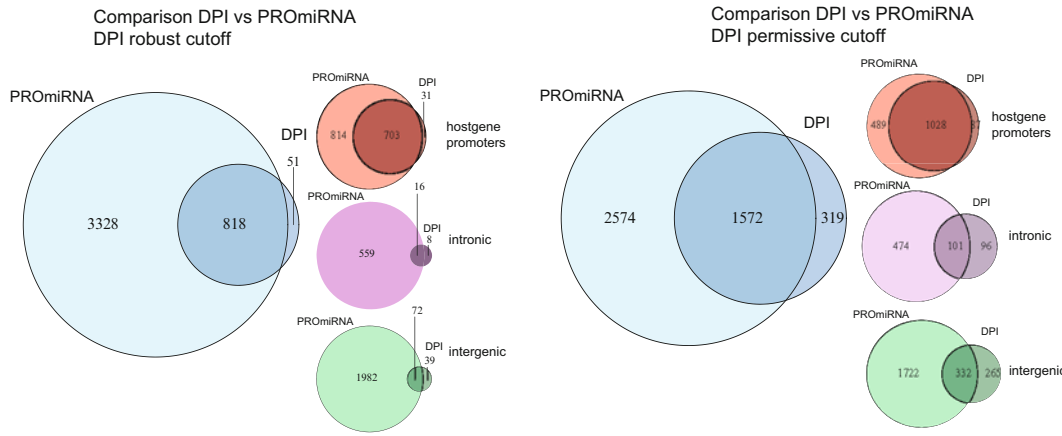


Fig. 4 Comparison between DPI and PROmiRNA for the miRNA promoter prediction task in the Gm12878 cell line. Overlap between predictions is shown for different promoter classes and for two sets of DPI predictions: set 1—where a robust cutoff of 2.0 TPM expression has been applied to the representative CTSS of a tag cluster and set 2—where a more permissive cutoff of 0.7 TPM has been applied. The overlap between DPI and PROmiRNA is improved when considering the DPI set 2 (permissive cutoff) given the lower expression values of miRNA promoters compared to gene promoters. The biggest overlap is observed for miRNA host gene promoters in both cases (DPI set 1 and set 2), whereas the overlap between the two tools is limited when it comes to the prediction of intergenic and intronic miRNA promoters

tc.decompose_smoothing_merged.ctssMaxCounts.bed (permissive cutoff)

In order to compare DPI predictions with PROmiRNA predictions we considered only DPI peaks up to 50 kb upstream of annotated miRNAs from miRBase v 20. We also filtered DPI peaks according to their overlap with DNAseI hypersensitivity regions as done above. The procedure was repeated for the two DPI output files corresponding to both the robust and permissive cutoff on the read counts.

The results from the comparison are summarized in Fig. 4. First of all, we observe that the largest overlap between DPI and PROmiRNA predictions is reached with the sets of DPI peaks at the permissive cutoff. This strengthen the argument that miRNA promoters are lowly detected compared to the protein-coding gene promoters due to fast processing of the miRNA primary transcripts, and a strict cutoff on the read counts will not allow their genome-wide identification.

The overlap between the tools is very high for miRNA host gene promoters, i.e., the promoters of protein-coding genes hosting miRNA hairpins inside their transcripts, and lower for intergenic and independent intragenic miRNA promoters, This underlines the fact that miRNA promoter prediction is still a challenging task compared to protein-coding gene promoter prediction.

Overall, PROmiRNA returns many more predictions than DPI: this might be due to the fact that PROmiRNA, unlike DPI, does not filter out genomic positions where only one tag maps, but includes them in the subsequent analysis, generating inevitably more predictions. While this is necessary in order to capture TSSs of lowly expressed miRNA primary transcripts which harbor promoter features in their sequences, it can happen that a certain fraction of PROmiRNA predictions represent false positives. However, most of predictions might represent real alternative miRNA promoters which need further investigation and validation.

4 Notes

1. The input files to DPI are CTSS files in bed format. CTSS stand for CAGE Transcription Start Site, and a CTSS file stores the 5′-end positions of the representative CAGE tags which start at the same genomic position on the same strand, together with the total number of representative CAGE tags at that position. A representative CAGE tag is a group of tags which have identical sequence (and therefore identical genomic mapping) [30].
2. A Tag Cluster (TC) is a cluster of overlapping TSSs, which are within 20–21 bp of each other. A TC genomic regions spans from the 5′-end of its most 5′-end tag, to the 3′ end of its most 3′-end tag. Two adjacent but non-overlapping tags contribute to separate TCs unless they are bridged by another tag. For a more detailed definition and some examples see [30].
3. Independent component analysis (ICA) is a useful method in signal processing, which is used to decompose a multivariate signal into subsignals (*see* Fig. 5), when knowing/assuming that the subsignals are independent and non-Gaussian, and by maximizing the statistical independency of the subsignals. The input of ICA is n observations of mixed signal, and each observation is a linear mixture of the original signals. The ICA technique is exemplified in Fig. 5. The top panel shows a simulated signal, which consists of $m = 2$ independent non-Gaussian source signal components, over time. Assume that we observed $n = 3$ independent observations of this mixed signal (middle panel). By applying the ICA technique we are able to decompose the observed mixed signal in two estimated signal components, which correspond to the original signal we want to reconstruct (lower panel).
   DPI assumes that each long tag cluster peak corresponds to a mixed signal (i.e., independent CAGE profiles). The m subsignals come from different transcriptional starting sites. The expression level of each TSS is independent from the others, and the signals are assumed to be non-Gaussian. Thus, ICA is

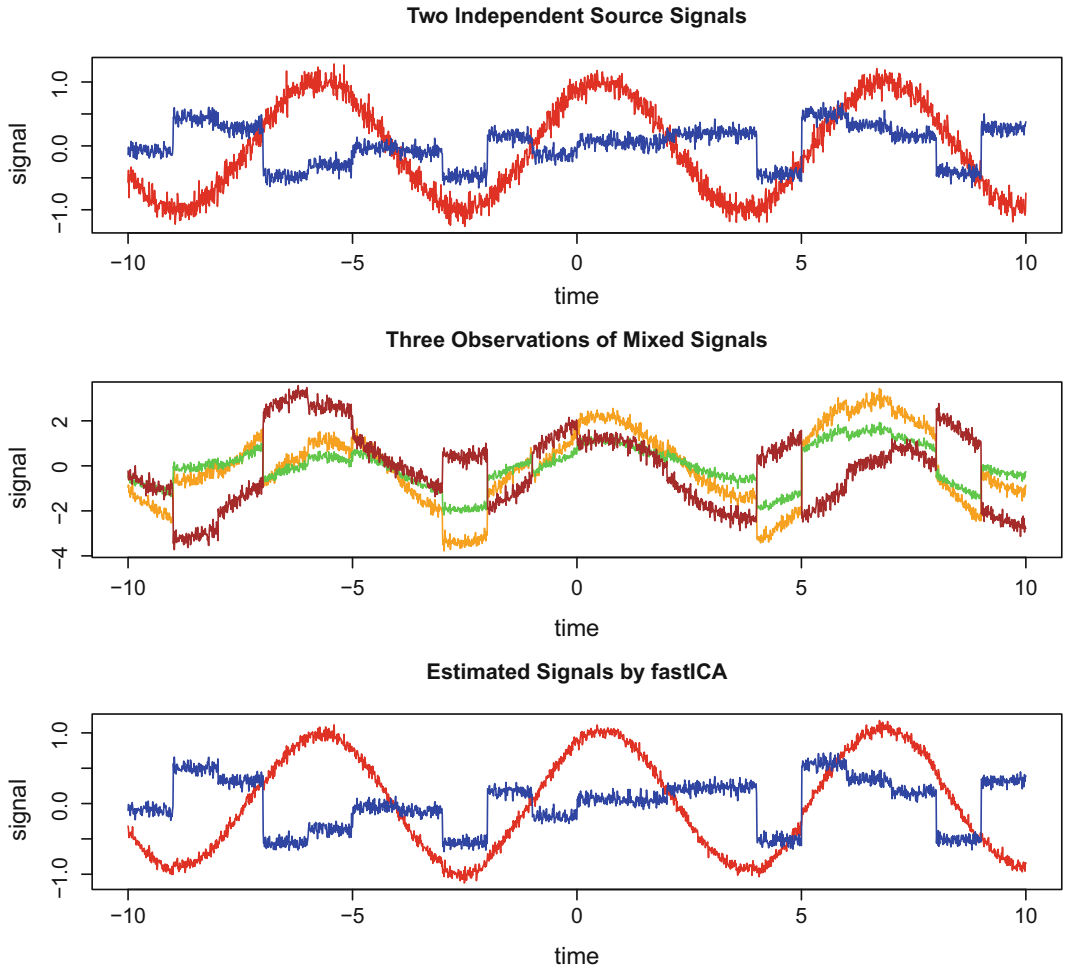


Fig. 5 Upper panel. Simulation of a mixed signal over time. Middle panel. Independent observations of a noisy mixed signals. Lower panel. Reconstruction of the two independent components of the noisy mixed signal using FastICA, an efficient R implementation of Independent Component analysis

suitable to separate mixed TSSs peaks into single TSS peaks. Different tissue types correspond here to the *n* observations.

DPI calculates ICA using the R package *fastICA*, an efficient and popular algorithm for finding an orthogonal rotation of the data [52]. In *fastICA*, the non-Gaussianity is measured as a proxy for the statistical independency using approximations to negative-entropy, which is robust and fast to compute.

4. The input files to PROmiRNA are CAGE tag alignments on the genome of interest in bed format (one bed file for each library), with six columns: *chromosome* (in UCSC format, e.g., chr1), *start* (5′-end of the aligned tag), *stop* (3′-end of the aligned tag), *tag_identifier* (or any other string), *number of tags* (number of identical tags mapping exactly at those position), strand.

5. When training PROmiRNA with new CAGE libraries on a new organism, the external annotation data (provided for hg19 with the current version of PROmiRNA) has to be built from scratch. Phastcons files in WigFix format for each chromosome can be downloaded at http://hgdownload.cse.ucsc.edu/goldenPath/hg19.

   For example, for hg19 the link is: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate

   WigFix files need to be converted to the binary wib format. This can be done with the *wigFix2wib.pl* script provided in *PROmiRNA/src*. Example of usage:

   ```
   > wigFix2wib.pl inFile1.wigFix[.gz][in-
   File2.wigFix]…]
   ```

   The generated *.wib files have to be placed in the directory *PROmiRNA/external_data/Phastcons* before using the software.

6. To retrieve the *chrom.sizes* file for your organism of interest use the *fetchChromSizes* script from http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/fetchChromSizes.

   Example of usage:

   ```
   > fetchchromSizes <db> <db>.chrom.sizes
   ```

   <db> corresponds to one of the ucsc databases (e.g., hg18, hg19, mm9, etc.). Place the *<db>.chrom.sizes* file in the *PROmiRNA/external_data* directory.

7. Sequence fasta files for each chromosome can be downloaded at http://hgdownload.cse.ucsc.edu/goldenPath/<db>/chromosomes/, where <db> corresponds to one of the ucsc databases (e.g., hg18, hg19, mm9, etc.). Pool the individual *chrom.fa* files into a common file *<db>.fa* using the Linux command *cat*. For example:

   ```
   > cat chr1.fa, chr2.fa, …… > hg19.fa
   ```

   Place the <db>.fa file in the *PROmiRNA/external_data* directory.

   Afterwards, create a fasta index file from <db>.fa using the samtools [53]:

   ```
   > samtools faidx PROmiRNA/external_
   data/<bd.fa>
   ```

8. PROmiRNA excludes repetitive regions when forming TCs from CAGE tags. In order to allow that, it requires a file in bed format listing the genomic coordinates of annotated repetitive regions for the genome of interest. A repeat file can be downloaded from the UCSC Genome Browser with the following instructions:
   - Go on the UCSC website (https://genome.ucsc.edu) and select *Table Browser* on the left menu;

- Select the organism and the genome assembly. For example, for the human assembly hg19 select *Mammal* in the *clade* field, "Human" in the *genome* field, *GRCh37/hg19* in the *assembly* field, *RepeatMasker* in the *track* field, "genome" in the *region* field and *BED-browser extensible data* in the *output format* field.
- Click the *get output* button to retrieve the desired file in bed format
- The repeat file has to placed in the *PROmiRNA/external_data* directory.

9. In order to annotate host gene and intronic promoters for intragenic miRNAs, PROmiRNA requires a gene annotation file in gtf Ensembl format. Such a file can be downloaded from the Ensembl ftp site (www.ensembl.org/info/data/ftp/index.html) and has to be placed in the *PROmiRNA/external_data* directory.

Acknowledgements

We would like to thank Xintian You for useful criticism and proofreading of the manuscript. The project was supported by the Freie Universität Berlin within the Excellence Initiative of the German Research Foundation.

References

1. Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. Genome Res 7
2. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 6
3. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5
4. Yella VR, Bansal M (2014) In silico Identification of Eukaryotic Promoters. In: Systems and synthetic biology
5. Abeel T, Saeys Y, Bonnet E et al (2008) Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res 18
6. Sandelin A, Carninci P P, Lenhard B et al (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet 8
7. Zeng J, Zhu S, Yan H (2009) Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. Brief Bioinform 10
8. Kondrakhin YV, Kel AE, Kolchanov NA et al (1995) Eukaryotic promoter recognition by binding sites for transcription factors. Comput Appl Biosci 11
9. Hutchinson GB (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. Comput Appl Biosci 12
10. Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249
11. Matys V, Kel-Margoulis OV, Fricke E et al (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 1
12. Mathelier A, Zhao X, Zhang AW et al (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res
13. Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297

14. Knudsen S (1999) Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics 15
15. Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12
16. Ohler U, Niemann H, Liao G et al (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics 17
17. Abeel T, Saeys Y, Rouzé P et al (2008) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. Bioinformatics 24
18. Sonnenburg S, Zien A, Rätsch A (2006) ARTS: accurate recognition of transcription starts in human. Bioinformatics 22
19. Xie X, Wu S, Lam KM et al (2006) PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. Bioinformatics 22
20. Zhao X, Xuan Z, Zhang MQ (2007) Boosting with stumps for predicting transcription start sites. Genome Biol 8
21. Wang J, Ungar LH, Tseng H et al (2007) MetaProm: a neural network based meta-predictor for alternative human promoter prediction. BMC Genomics 8
22. Won HH, Kim MJ, Kim S et al (2008) EnsemPro: an ensemble approach to predicting transcription start sites in human genomic DNA sequences. Genomics 91
23. Valen E, Sandelin A (2011) Genomic and chromatin signals underlying transcription start-site selection. Trends Genet 27
24. Johnson DS, Mortazavi A, Myers AM et al (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316
25. 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 U S A 100
26. Ravasi T, Suzuki H, Cannistraci CV et al (2010) An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140
27. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322
28. Wang X, Xuan Z, Zhao X et al (2009) High-resolution human core-promoter prediction with CoreBoost_HM. Genome Res 19
29. Megraw M, Pereira F, Jensen TH et al (2009) A transcription factor affinity-based code for mammalian transcription initiation. Genome Res 19
30. Carninci P, Sandelin A, Lenhard B et al (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38
31. (Dgt) FCaTRPaC (2014) A promoter-level mammalian expression atlas. Nature 507
32. Marsico A, Huska MR, Lasserre J et al (2013) PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. Genome Biol 14
33. Gustincich S, Sandelin A, Plessy C et al (2006) The complexity of the mammalian transcriptome. J Physiol 575
34. Valen E, Pascarella G, Chalk A et al (2009) Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res 19
35. Consortium F (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet 41
36. Kanamori-Katayama M, Itoh M, Kawaji H et al (2011) Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res 21
37. Andersson R, Gebhard C, Miguel-Escalada I et al (2014) An atlas of active enhancers across human cell types and tissues. Nature 507
38. 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
39. Frith MC, Valen E, Krogh A et al (2008) A code for transcription initiation in mammalian genomes. Genome Res 18
40. Balwierz PJ, Carninci P, Daub CO et al (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol 10
41. Haberle V, Forrest AR, Hayashizaki Y et al (2015) CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res 43
42. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25
43. Faulkner GJ, Forrest AR, Chalk AM et al (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics 91
44. Lassmann T, Frings O, Sonnhammer EL (2009) Kalign2: high-performance multiple

alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37

45. Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 489
46. Kadota K, Nishiyama T, Shimizu K (2012) A normalization strategy for comparing tag count data. Algorithms Mol Biol 7
47. Severin J, Waterhouse AM, Kawaji H et al (2009) FANTOM4 EdgeExpressDB: an integrated database of promoters, genes, microRNAs, expression dynamics and regulatory interactions. Genome Biol 10
48. Severin J, Lizio M, Harshbarger J et al. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat Biotechnol 32
49. Lizio M, Harshbarger J, Shimoji H et al (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol
50. Robinson MD, Mccarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26
51. Griffiths-Jones S, Grocock RJ, Van Dongen S et al (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 1
52. Hyvärinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10
53. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25

Chapter 14

Bioinformatics Pipeline for Transcriptome Sequencing Analysis

Sarah Djebali, Valentin Wucher, Sylvain Foissac, Christophe Hitte, Evan Corre, and Thomas Derrien

Abstract

The development of High Throughput Sequencing (HTS) for RNA profiling (RNA-seq) has shed light on the diversity of transcriptomes. While RNA-seq is becoming a de facto standard for monitoring the population of expressed transcripts in a given condition at a specific time, processing the huge amount of data it generates requires dedicated bioinformatics programs. Here, we describe a standard bioinformatics protocol using state-of-the-art tools, the STAR mapper to align reads onto a reference genome, Cufflinks to reconstruct the transcriptome, and RSEM to quantify expression levels of genes and transcripts. We present the workflow using human transcriptome sequencing data from two biological replicates of the K562 cell line produced as part of the ENCODE3 project.

Key words Transcriptome sequencing, Protocols, RNA-seq, Bioinformatics workflow

1 Introduction

1.1 RNA-Seq Technology

The application of HTS technologies for cDNAs (RNA-seq) allows to characterize the myriad of RNA molecules transcribed in a given cell or tissue at a specific time point [1]. The so-called transcriptome sequencing provides a unique snapshot of all expressed transcripts in a particular condition and thus informs about fundamental biological processes such as (1) the transcribed coding (mRNAs) and noncoding RNAs (ncRNAs), (2) novel alternative isoforms, chimeric genes/transcripts, and (3) the levels of expression of these RNAs.

Depending on the experimental protocol, RNA-seq can be used to target specific categories of RNAs based for instance on their sizes (long vs. short RNAs), their molecular properties (RNAs with a polyA tail vs. ribosomal RNAs), or their cellular compartments (cytoplasmic vs. nuclear RNAs) [2].

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_14, © Springer Science+Business Media New York 2017

However, most of the active sequencing platforms worldwide currently rely on a technology that generally does not produce the entire sequence of a nucleic acid: the process can only generate sequences—called "reads"—of a limited length from the extremities of each RNA molecule. Therefore a fragmentation step is generally included in the protocol in order to allow any position of the transcript to be potentially sequenced. When both extremities of each fragment are read the process is called Paired-End sequencing, and pairs of reads are obtained. The read length and the size distribution of the sequenced fragments are important features of the process.

Typically, tens of million of reads of 100–200 bp are generally produced from fragments of about 200–400 bp. These reads need to be:

- Mapped back onto a reference genome taking into account splice sites (mapping step).
- Assembled into exon–intron structures (transcriptome reconstructions step).
- Used to quantify known and/or novel transcripts or genes (quantification step).

However, given the depth provided by current sequencing machines and although numerous and efficient bioinformatics tools dedicated to this task exist, dealing with such massive amounts of data remains a challenge.

1.2 The STAR: Cufflinks: RSEM pipeline

Here we present a commonly used bioinformatics pipeline to process RNA-seq reads using STAR [3] for mapping sequences, Cufflinks [4] for transcript model reconstruction and RSEM [5] for transcript and gene quantifications (*see* Fig. 1). This pipeline quantifies annotated genes and transcripts, however the commands we provide are general enough to be easily extended to quantify both known and novel transcripts.

These programs, actively maintained by their developers, are widely used by the community including international consortia such as ENCODE3 [6], Blueprint [7] or TCGA (http://cancergenome.nih.gov). Moreover, benchmarks done as part of the RGASP project (RNA-seq Genome Assessment Project) [8, 9], or as part of the ENCODE3 evaluation (under review), show that they yield favorable performances while limiting computational needs.

1.3 Alternative Pipelines

Nevertheless, there are many alternative programs that could be used at each step of the workflow with for instance tophat2/bowtie2 [10] or the GEMtools RNA-seq pipeline [11] for splice-aware read mapping software, stringtie [12] for transcriptome reconstruction, and Flux capacitor [13], eXpress [14] or Sailfish [15] for transcript and gene quantifications.

In the following tutorial, we assume that both a reference genome sequence and a genome annotation are available, allowing

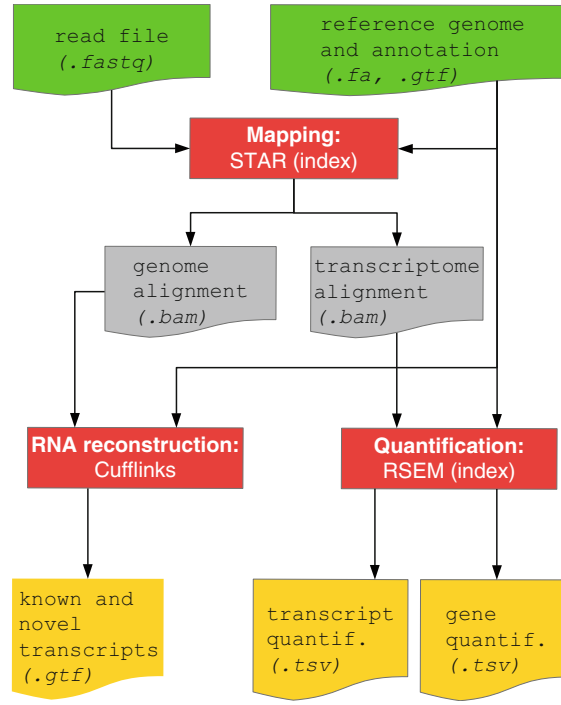


Fig. 1 Pipeline description. Schematic overview of the bioinformatics pipeline described in this protocol. Input files are in *green*, intermediary files in *gray*, output files in *yellow,* and the main steps (bioinformatics tools) in *red*. Using reference genome and annotation files, RNA-seq reads are mapped using STAR to the genome and to the transcriptome. The genome alignment output file is then used by Cufflinks to reconstruct known and novel transcripts. The transcriptome alignment output file is used by RSEM to quantify the levels of expression of genes and transcripts. The index construction required by STAR and RSEM is implicitly represented

the user to implement a genome-guided assembly protocol. In the absence of a reference genome sequence, one would favor de novo transcriptome assembly which involves different algorithms, such as for instance, Trinity [16] or KisSplice [17]. In addition, the biological and bioinformatics protocols may vary with respect to the sequencing technology and the species to be studied. Here, we illustrate the method which uses Illumina paired-end reads (cf. **Note 1**) from vertebrate species samples.

2 Materials

While genome-guided algorithms for transcriptome assembly tend to limit computational resources (which would not be the case for de novo transcriptome assembly), minimal computational resources are still required. We have tested this protocol using a 64-bit Linux system with 32Go of RAM and 8 cores.

In this protocol, all command lines will be written in Courier New police.

Moreover, in order to distinguish biological materials (sequenced reads, reference genome and annotation files) from bioinformatics software, all input "biological files" will be stored in a specific directory named material.

To create this directory, type the following command line:

```
mkdir material
```

2.1 RNA-Seq FASTQ Reads

For this tutorial, we use two biological replicates of the human K562 cell line (adult 53 year female) from the human ENCODE3 RNA evaluation project. The files are available here:

https://www.encodeproject.org/experiments/ENCSR000AEM/

The libraries come from two independent growths of the K562 cell line and the sequencing was done using Illumina Hi-Seq technology with a stranded paired-end read protocol. Only polyA+ RNAs with a size greater than 200 nucleotides were selected, naturally leading to a ribosomal RNA depletion.

Since there are two biological replicates, four read files are available (i.e., two mates x two replicates) with the following IDs ENCFF001RDZ.fastq.gz and ENCFF001RED.fastq.gz for replicate 1 and ENCFF001REF.fastq.gz and ENCFF001REG.fastq.gz for replicate 2. For clarity, we will add a \ when the command line is too long. To download the compressed read files:

```
wget -O ENCFF001RDZ.fastq.gz \
https://www.encodeproject.org/files/
   ENCFF001RDZ/@@download/ENCFF001RDZ.fastq.gz
wget -O ENCFF001RED.fastq.gz \
https://www.encodeproject.org/files/
   ENCFF001RED/@@download/ENCFF001RED.fastq.gz
wget -O ENCFF001REF.fastq.gz \
https://www.encodeproject.org/files/
   ENCFF001REF/@@download/ENCFF001REF.fastq.gz
wget -O ENCFF001REG.fastq.gz \
https://www.encodeproject.org/files/
   ENCFF001REG/@@download/ENCFF001REG.fastq.gz
```

Then, move these four files into the material directory:

```
mv ENCFF001*fastq.gz materials/
```

2.2 Reference Genome and Annotation Files

We will use the human genome assembly version hg19 (aka GRCh37) available as a multifasta file (each sequence corresponding to one chromosome) from the UCSC website [18] here:

```
wget http://hgdownload.cse.ucsc.edu/golden-
   Path/hg19/encodeDCC/referenceSequences/male.
   hg19.fa.gz
```

To decompress the file:

```
gunzip male.hg19.fa.gz
```

Please, note that this file corresponds to the primary assembly of the human male genome, i.e., without haplotypic chromosomes but including chromosome Y (*see* **Note 2**). Repeat sequences are soft-masked which means that all repeats and low complexity regions have been replaced with the lowercase version of their nucleic base.

The reference genome annotation used in this tutorial will be GENCODE [19] which is part of the ENCODE project and whose aim is to annotate all evidence-based gene features on the human genome. The GENCODE gene set is actually the reference human gene annotation used by international projects (ENCODE, 1000 genomes…), its main advantages being its comprehensiveness, since it includes both long noncoding RNA [20] and pseudogene [21] annotations. Providing such a reliable data to the process is fundamental, as several steps of the pipeline (mapping, transcript building, and quantification) are affected by the quality of the annotation.

Generally, the annotation file is stored in a .GFF or GFF3 (General Feature Format) or a .GTF/GFF2.5 (General Transfer Format corresponding to the GFF2.5) which is a 9 columns tab-delimited file storing informations (localization, source, transcriptional orientation) on specific features (gene, transcript, exon, etc.).

Importantly, the version of the annotation must correspond to the genomic sequence used in the pipeline: make sure that the gencode file version matches with the genome assembly (*see* **Note 3**). Here, we retrieved gencode version 19 from (built on the hg19 human genome assembly) usingthe gencode FTP website (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/):

```
wget
ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_hu-
   man/release_
19/gencode.v19.annotation.gtf.gz
gunzip gencode.v19.annotation.gtf.gz
```

Then, both reference files (genome and annotation) are moved to the material directory:

```
mv male.hg19.fa gencode.v19.annotation.gtf ma-
terial/
```

In order to ease the access to these files, it is recommended to create two shell variables ($GENOME and $ANNOTATION for genome and annotation files, respectively) which will point to the absolute path of the files:

```
GENOME=$(readlink -f ./material/male.hg19.fa)
ANNOTATION=$(readlink -f ./material/gencode.
v19.annotation.gtf)
```

2.3 Software Installation

The bioinformatics programs used in this tutorial will be stored in a specific directory named bin. To create this directory and move there, type the following:

```
mkdir bin
cd bin
```

One essential step in each program installation is to make sure that the directory where the program has been installed (or is located) is present in your PATH environment. This dynamic variable lists all the directories that the shell searches through when the user tries to execute a program, thus avoiding the need to use the full path to the program.

2.4 SAMtools

SAMtools [22] are a suite of utilities for manipulating alignments in SAM format (Sequence Alignment/Map) which is the standard format for storing large nucleotide alignments (typically, those encountered in HTS sequencing). In addition, SAMtools are also required by some of the programs described in this chapter such as cufflinks for instance. To install SAMtools, one needs to have zlib and htslib installed.

Check the latest version of the SAMtools on the dedicated website in order to download it:

```
wget -O samtools-1.2.tar.bz2 \ https://github.
   com/samtools/samtools/releases/download/1.2/
   samtools-1.2.tar.bz2
tar xjvf samtools-1.2.tar.bz2
cd samtools-1.2
make
```

As mentioned above, the resulting "samtools" binary is added to the user's PATH environment variable using the following command line where $PWD corresponds to the full path of the current working directory. Please note that this command line assumes a Bourne-Again shell interpreter (bash, *see* **Note 4**).

```
export PATH=$PATH:${PWD}
```

2.5 STAR

Download the latest version of the STAR mapper [3] freely available from its github website: https://github.com/alexdobin/STAR

```
wget -O STAR_2.5.0a.tar.gz \ https://github.
   com/alexdobin/STAR/archive/STAR_2.5.0a.tar.
   gz
tar zxvf STAR_2.5.0a.tar.gz
cd STAR-STAR_2.5.0a/
```

Here, it is either possible to use the precompiled binaries in the ./bin/ directory or to compile the sources such as:

```
cd source
make STAR
export PATH=$PATH:${PWD}
```

2.6 Cufflinks

Like STAR, Cufflinks [4] can either be installed using a precompiled binary release or built from the sources (note that the latter option also requires to install the Boost C++ libraries).

Both Linux and Mac versions are available. Here we download the latest binary version 2.2.1 and then export the cufflinks executable in the PATH.

```
wget http://cole-trapnell-lab.github.io/cuf-
   flinks/assets/downloads/cufflinks-2.2.1.Linux_
   x86_64.tar.gz
tar xzvf cufflinks-2.2.1.Linux_x86_64.tar.gz
export PATH=$PATH:${PWD}
```

2.7 RSEM

For RSEM [5], download the latest archive available on the github website (https://github.com/deweylab/RSEM) and then add it to your PATH. Note that for compatibility with STAR2.5 it is essential to use a version of RSEM that is at least 1.2.25.

```
wget -O RSEM.v1.2.25.tar.gz \ https://github.
   com/deweylab/RSEM/archive/v1.2.25.tar.gz
cd RSEM-1.2.25/
make
export PATH=$PATH:${PWD}
```

3 Methods

Before starting to process the sequence reads, and if this task has not already been performed by the sequencing platform, it is always relevant to assess their quality (*see* **Note 5**). This tutorial demonstrates how to run each step of the pipeline separately, allowing to self- tune each step's parameters with respect to users' needs and to understand problems when they arise. Note that for the mapping and the known gene and transcript quantification parts of the pipeline, bash script and nextflow implementations also exist (*see* **Notes 6** and **7**). In case the RNA-seq experiment has been performed using control RNA spike-ins, they can be used by slightly changing the following protocol (*see* **Note 8**).

3.1 Mapping

STAR uses a suffix array approach to map reads to the genome and to the annotated splice junctions. Reads can be mapped both in a continuous way, i.e., in one block, or in a noncontinuous way, i.e., allowing gaps which can be considered as introns if long enough (see --alignIntronMin option below), in which case the read mapping is called a split-mapping.

STAR

3.1.1 Making the STAR Indices

This only needs to be done once for a given project. The same indices can then be used for all RNA-seq datasets of this project.

3.1.2 Input Files and Arguments

$STARgenomeDir is the directory where the STAR indices will be stored. This directory has to be created with mkdir and given write permissions before the command is run.

- $GENOME is the genome FASTA file.
- $ANNOTATION is the gene annotation in GTF format.
- $threads is the number of threads for parallelizing the task. Here, we fixed it to 8 given the computing resources available, but if one can only use 4 the job will simply take longer (See Mat.).

3.1.3 Command

```
STAR --runThreadN $threads --runMode genomeGen-
erate \
--genomeDir $STARgenomeDir --genomeFastaFiles
   $GENOME \
--sjdbGTFfile $ANNOTATION --sjdbOverhang 100 \
--outFileNamePrefix $STARgenomeDir
```

Note that the --sjdbOverhang option corresponds to the length of the genomic sequence around the annotated junction to be used in constructing the splice junction database, and should be set to the read length minus one.

3.1.4 Output Files

The above command will generate many genome files in the directory $STARgenomeDir, most of which use internal STAR format and are not intended to be utilized by the end user. None of them should be changed. The chrNameLength.txt file contains the chromosome names and lengths and is useful to generate RNA-seq signal files in bigwig format (https://genome.ucsc.edu/goldenpath/help/bigWig.html) from the continuous valued bedgraph files produced by STAR (see below).

3.1.5 Mapping the Reads

Mapping of the reads has to be performed for each RNA-seq dataset of a given project. STAR creates a genome mapping file and a transcriptome mapping file successively, by internally converting genome global coordinates to transcript local coordinates.

3.1.6 Input Files and Arguments

- $STARgenomeDir is the STAR index file directory (see above).
- $read1 is the gzipped FASTQ file of the first mates.
- $read2 is the gzipped FASTQ file of the second mates.
- $nThreadsSTAR is the number of threads.

3.1.7 Command

```
STAR --genomeDir $STARgenomeDir --readFilesIn
$read1 $read2 \
--readFilesCommand zcat --outFilterType
   BySJout --outSAMunmapped Within \
--outSAMtype BAM SortedByCoordinate --outSA-
   MattrIHstart 0 \
--outFilterIntronMotifs RemoveNoncanonical
   --runThreadN $nThreadsSTAR \
--quantMode TranscriptomeSAM --outWigType bed-
   Graph --outWigStrand Stranded
```

The --readFilesCommand zcat option is used to uncompress gzipped read fastq files provided as input, while the --outFilterType BySJout option is used to reduce the number of spurious junctions. The --outSAMunmapped Within and the --outSAMtype BAM SortedByCoordinate are used to produce a standard bam file sorted by coordinates, while the --outSAMattrIHstart 0 and the --outFilterIntronMotifs RemoveNoncanonical are important to produce an output file that is compatible and better to use for cufflinks, respectively. The --quantMode TranscriptomeSAM is used to produce a transcriptome bam file that will be used by RSEM, while the --outWigType bedGraph and the --outWigStrand Stranded optionsproduce stranded continuous valued bedgraph files from the alignments.

STAR allows the user to specify many other options, which can be found in the STAR manual (https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf), however a shorter list of such options, currently used in ENCODE3 and for which we use default values here, are provided in Table 1.

3.1.8 Output Files

STAR produces many output files within the current working directory (note that the output directory can be changed with the --outFileNamePrefix option), the most important of which are the following eight:

- Log.final.out: summary mapping statistics, useful for quality control, with the number and percentage of initial fragments (read pairs) that were mapped, the number and percentage of fragments that are mapped uniquely, that are mapped multiple times (called multimaps, split into the different reasons for that), and unmapped (summarized in Table 2), as well as statistics about splice junction detection.
- Aligned.sortedByCoord.out.bam: the genome BAM file sorted by coordinates.
- Aligned.toTranscriptome.out.bam: the transcriptome BAM file. This file is not sorted, which, in case several threads are used, does not guarantee exact reproducibility of the downstream RSEM quantifications. If exact reproducibility is wanted

Table 1

STAR options which differ between this tutorial and ENCODE3

Option	Meaning	Default value	Encode value
--outSAMattributes	A string of desired SAM attributes, in the order desired for the output SAM	NH HI AS nM	NH HI AS NM MD
--outFilterMultimapNmax	Read alignments will be output only if the read maps fewer times than this value, otherwise no alignments will be output	10	20
--outFilterMismatchNmax	Alignment will be output only if it has fewer mismatches than this value	10	999
--outFilterMismatchNoverReadLmax	Alignment will be output only if its ratio of mismatches to read length is less than this value	1	0.04
--alignIntronMin	Minimum intron size: genomic gap is considered intron if its length>=alignIntronMin, otherwise it is considered Deletion	21	20
--alignIntronMax	Maximum intron size: if 0, max intron size will be determined by (2^winBinNbits)*winAnchorDistNbins	0	1000000
--alignMatesGapMax	Maximum gap between two mates, if 0, max intron gap will be determined by (2^winBinNbits)*winAnchorDistNbins	0	1000000
--alignSJoverhangMin	Minimum overhang (block size) for spliced alignments	5	8
--alignSJDBoverhangMin	Minimum overhang (block size) for annotated spliced alignments	3	1
--sjdbScore	Extra alignment score for alignments that cross database junctions	2	1
--genomeLoad	Mode of shared memory usage for the genome files	NoSharedMemory	LoadAndKeep
--limitBAMsortRAM	Maximum available RAM for sorting BAM. If 0, it will be set to the genome index size. 0 value can only be used with --genomeLoad NoSharedMemory option	0	10000000000

For each such option we provide its name, its meaning, its default value (used here), and the value used in ENCODE3. The mapping results used with one set of values or the other do not vary drastically, and lead to very similar gene and transcript quantifications

Table 2

Mapped fragment statistics

	Mapped		Uniquely mapped		Multi-mapped	
# Fragments	#	%	#	%	#	%
113,327,735	103,116,159	91.0	99,717,493	88.0	3,398,666	3.0

Number of initial, mapped, uniquely mapped, and multi-mapped fragments are provided for the RNA-seq experiment under study

and more than 1 thread is used, this file has to be sorted (as is done in https://github.com/ENCODE-DCC/long-RNASeq-pipeline/blob/master/DAC/STAR_RSEM.sh).

- SJ.out.tab: contains high confidence collapsed splice junctions derived from the split-mapped reads.
- Signal.[type].str[no].out.bg, where type is the type of alignment (either Unique or UniqueMultiple), and where no is the strand number (either 1 or 2), are four stranded bedgraph (BG) files made from the alignments and that can be converted into bigwig files (BW) using the bedGraphToBigWig UCSC tool, for visualization of the mapped read coverage over genes and other regions in the UCSC browser. The syntax of the bedGraphToBigWig command is the following: "bedGraphToBigWig file.bg $STARgenomeDir/chrNameLength.txt file.bw."

For this particular run, we provide both the distribution of mapped reads into genomic domains in (Fig. 2) and an example of coverage plot over a typical gene, *MYC*, in (Fig. 3).

3.2 Transcriptome Reconstruction/ Assembly

3.2.1 Cufflinks

Cufflinks aims at assembling reads mapped to the genome into transcripts, using or not the annotation as a guide. Therefore it uses as input the bam file generated previously, and optionally the reference annotation in GTF format. Even if cufflinks can also provide quantification of expression of the reconstructed transcripts, here, we will only use it as a transcript modeler (while RSEM will be used for quantification).

3.2.2 Input Files and Arguments

- $nthreads is the number of threads that can be used for the computation.
- $ANNOTATION is the gene annotation in GTF format.
- $outdir is the directory where the results will be stored.
- $libtype is the library type (for illumina stranded or unstranded **Note 9**).
- $bam is the genomic BAM file obtained in the previous step.

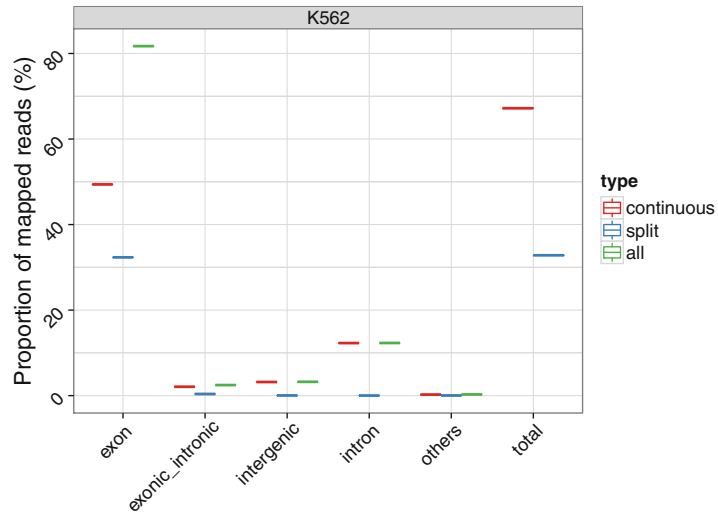


Fig. 2 Mapped read distribution in genomic domains. Primary alignments of reads are partitioned into continuous and split reads and then into the following categories: (1) exonic if they are totally included in exons, (2) intronic if they are totally included in introns, (3) exonic–intronic if they are totally included in genes but not in (1) or (2), (4) intergenic if they are totally included in intergenic regions, and (5) others if they are not in the previous categories. Even if the majority of genes map in a continuous way, the percentage of split-mapped reads is quite high (33 % of the total mapped reads). Most mapped reads fall into exons (82 %), and then introns (12 %), but very few of them lie at exon-intron boundaries and in intergenic regions (<5 %)

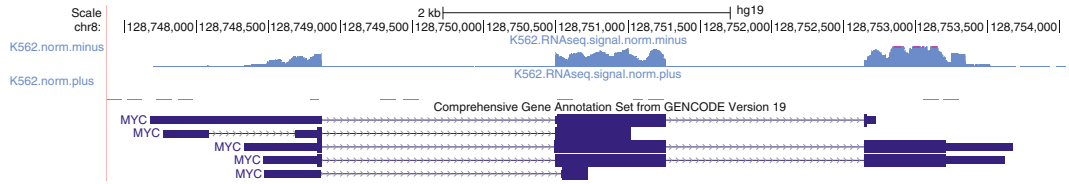


Fig. 3 Read coverage of the MYC gene. This figure shows the + and – strand RNA-seq signal (RNA-seq mapped read aggregation) over the MYC gene in the UCSC browser. As expected from a polyA+ RNA experiment, exons are covered more than introns. Note that this gene has several transcript isoforms annotated in Gencode v19

3.2.3 Command

```
cufflinks -p $nthreads -g $ANNOTATION -o $outdir
-u \
--library-type fr-firststrand $libtype $bam
```

The -g/--GTF-guide option tells cufflinks to use the reference transcript annotation to guide the assembly. Unlike the -G/--GTF option which will ignore alignments that are not present in $ANNOTATION, the -g option allows the identification of novel transcript isoforms.

3.2.4 Output File

$outdir/transcripts.gtf is the file containing the transcript assembly produced by cufflinks.

3.3 Transcript and Gene Quantifications

RSEM uses reads mapped to the transcriptome to quantify the expression of transcripts and genes. It uses an expectation maximization approach to rescue multi-mapped reads based on the location of unique reads in the transcript, in an iterative way that stops when the error made is lower than a threshold.

Preparing the RSEM reference files

This needs to be done only once for a given project. The same reference files will then be used for all RNA-seq datasets of this project.

3.3.1 Input Files and Arguments

- $RSEMgenomeDir is the directory where the RSEM indices will be stored. This directory has to be created by the user before running the command.
- $GENOME is the genome sequence in FASTA format.
- $ANNOTATION is the gene annotation in GTF format.

3.3.2 Command

```
mkdir $RSEMgenomeDir
rsem-prepare-reference --gtf $ANNOTATION
   $GENOME $RSEMgenomeDir/RSEMref
```

3.3.3 Output Files

The above command generates 7 output files starting with the RSEMref prefix in the $RSEMgenomeDir output directory, of which only one is of interest to the user (RSEMref.transcripts.fa) and contains the extracted reference transcripts in Multi-FASTA format. The other ones are either used by RSEM internally (RSEMref.grp, RSEMref.ti, RSEMref.transcripts.fa, RSEMref.seq, RSEMref.chrlist) or useful when mapping is done within RSEM which is not the case here, see --no-bam-output option below (RSEMref.idx.fa and RSEMref.n2g.idx.fa).

3.4 Running the Quantification Process

3.4.1 Input Files and Arguments

- $nThreadsRSEM is the number of threads that can be used for the computation.
- Aligned.toTranscriptome.out.bam is the transcriptome BAM file generated previously by STAR.
- $RSEMgenomeDir is the directory where the RSEM reference files are located.

3.4.2 Command

```
rsem-calculate-expression --bam --no-bam-
output --estimate-rspd \
--calc-ci --seed 12345 -p $nThreadsRSEM --ci-
   memory 30000 --paired-end \
--forward-prob 0 Aligned.toTranscriptome.out.
   bam $RSEMgenomeDir/RSEMref Quant
```

The --bam and --no-bam-output options are used to specify that a transcriptome bam file is provided as input (as opposed to

default FASTQ files), and that the program should not generate a BAM file, respectively. The --estimate-rspd option is used to estimate the read start position distribution (RSPD) from the data, and the --calc-ci option calculates 95% credibility intervals and posterior mean estimates. The --seed 12345 option sets the seed for the random number generators used in calculating posterior mean estimates and credibility intervals, while the --paired-end and the --forward-prob 0 are used to specify that the data is paired-end and stranded with all the first reads coming from the opposite strand of the transcript, respectively. Quant is simply the name of the sample, used to label some output files.

3.4.3 Output Files

The rsem-calculate-expression command generates several output files in the current working directory, described in details in http://deweylab.biostat.wisc.edu/rsem/rsem-calculate-expression.html, the most important of which are the following two:

- Quant.isoforms.results, which is a TSV file containing the expression of the annotated transcripts. The 4 most important columns are 1, 5, 6, and 7 which respectively contain the transcript id, the number of reads assigned to this transcript, and two relative measures of its expression: the TPM (Transcript Per Million) and the FPKM (Fragment Per Kilobase of transcript per Million mapped reads).
- Quant.genes.results, for gene quantifications (with the same kind of file format as for transcript isoforms).

Note that TPM is the native RSEM measure of expression, and should be preferred over FPKM. Indeed while the sum of the FPKMs of all transcripts is not constant across samples, the sum of the TPMs of all transcripts expressed in a given sample is always 1, and therefore constant across samples. These quantification files are very useful for visualizing the distribution of transcript or gene expression in a given sample (Fig. 4), as well as for rapidly extracting the number of transcripts or genes detected in a given experiment (Table 3).

4 Notes

1. Single-end protocol:
   If the protocol is single-end some of the above commands have to be slightly modified:
   - STAR mapping: remove "$read2" in the command.
   - RSEM quantification: remove "--paired-end" from the command.

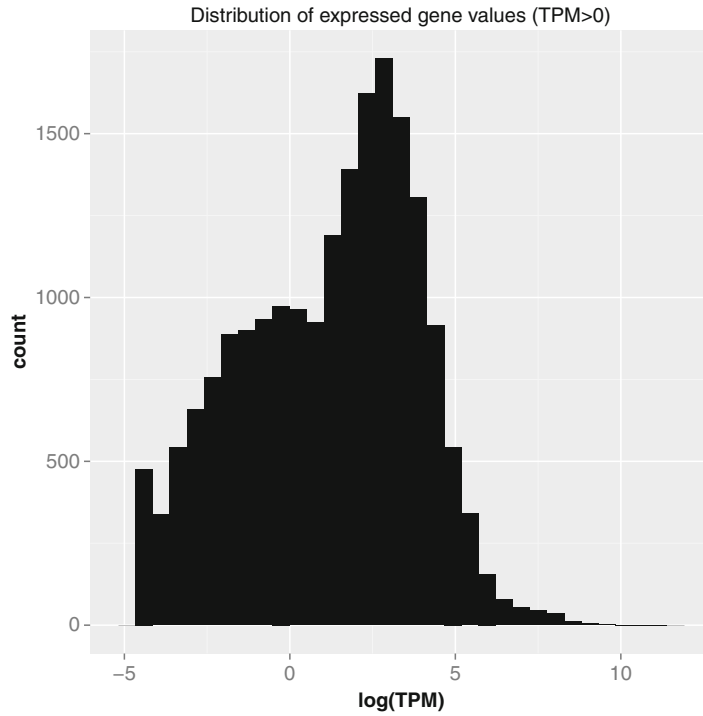


Fig. 4 Gene expression level. Log10 transformed TPM (transcript per million) values of expressed genes is plotted as an histogram. Globally, about 1/3 of the genes are expressed at the threshold of 0 TPM (*see* Table 2)

Table 3

Detected genes and transcripts

		Detected (TPM > 0)			Detected (TPM > 0)	
Super biotype	**# Genes**	**#**	**%**	**# Transcripts**	**#**	**%**
Protein_coding[a]	20,730	14,068	67.9	95,319	43,319	45.4
IncRNA[b]	13,870	3262	23.5	76,684	31,142	40.6
Pseudogene[c]	14,206	1959	13.8	15,343	1929	12.6
SmallRNA[d]	9013	35	0.4	9173	41	0.4
All	57,819	19,324	33.4	196,519	76,431	38.9

Number of annotated and detected genes and transcripts (TPM > 0), for 4 super biotypes (protein_coding, lncRNA, pseudogene, smallRNA), and for all annotated elements. The list of individual Gencode v19 biotypes belonging to each of the 4 super biotypes defined here is indicated at the bottom of the table

[a]IG_C_gene,IG_D_gene,IG_J_gene,IG_V_gene,nonsense_mediated_decay,non_stop_decay,protein_coding,TR_C_gene,TR_D_gene,TR_J_gene,TR_V_gene

[b]3prime_overlapping_ncrna,antisense,lincRNA,processed_transcript,retained_intron,sense_intronic,sense_overlapping

[c]All gencode biotypes containing the term pseudogene

[d]miRNA,misc_RNA,Mt_rRNA,Mt_tRNA,rRNA,snoRNA,snRNA

2. For whole RNA transcriptome sequencing (capturing both polyA+ and polyA- transcripts) it could be of importance to use the unplaced scaffolds when mapping, even if they do not contain any gene, since they act as a sponge of ribosomal reads at the level of the mapping. For instance, the hg19 assembly contains 59 unplaced supercontigs with sizes ranging from 4.2 kb (contig GL000207.1) to 5.5 Mb (for contig GL000207.1) and a total of 6.1 Mb.
3. One primordial task in many HTS bioinformatics analyses is to check whether the chromosome names present in the annotation file correspond to the ones present in the genome file. For instance, genome files from UCSC [18] are not compatible with annotation files from EnsEMBL [23] since the former use the chr_n convention while the latter use the n convention.
4. The command line for exporting the PATH variable depends on the shell language used. In this protocol, we use bash, but for tcsh or csh shells, the syntax should be the following:

```
sentenv PATH $PATH:/path/to/programdir
```

5. Read quality/ QC

   RNA-seq reads deposited in reference databases such as SRA or ENA have generally been precleaned and are normally of good quality, but before using raw sequences generated by sequencing machines, we need to check their quality and possibly clean them to get rid of adapters, contaminants and low quality regions that were introduced in the sequence at various stages of the RNA-seq library preparation.

   Regardless of the sources of the data, the first step consists in assessing read quality. One of the reference tools to this aim is FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) which provides an HTML report on different quality metrics (phred score, GC%, GC and k-mer bias, duplication, adapter contamination, etc.). It can either be launched on the command line (it is then possible to parallelize the processing for big datasets) or in interactive mode for smaller datasets. For example, for the processing of two files on 8 cores:

```
fastqc -t 8 seqfile1 seqfile2
```

   Then we need to clean the reads. If the cleaning steps are less critical in the case of a genome-guided assembly which use local mapping tools (e.g., Bowtie2/STAR), than in the case of de novo assembly, the ultimate goal is still to assign reads to their correct positions and thus remove low quality regions that potentially contain errors and introduce biases in the quantifications.

   Errors should be removed in the reverse order of the sources that have generated them. (1) First the so-called sequencing-related technical errors: low quality read parts and technical

contaminations like adapters. Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) performs an adaptive trimming from the read ends and applies a sliding window over the entire sequence. It preserves information on paired and singletons and performs cleaning from a list of adapters sequences. To use it:

```
trimmomatic PE -threads 8 r1.fq.gz
r2.fq.gz r1_paired.fq.gz r1_unpaired.
fq.gz r2_paired.fq.gz r2_unpaired.fq.gz
ILLUMINACLIP:adaptor_list.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 min-
len:50
```

Secondly, the so-called biological errors related to the library preparation protocol and contamination: polyA-tails, rRNA sequences, mtDNA sequences. Since mitochondrial and ribosomal RNAs could be polyadenylated, the use of a polyA selection is not an absolute guarantee to get rid of all these molecules that often represent more than 90% of the cellular RNAs. Ribopicker (ribopicker.sourceforge.net/) can be used to identify and remove contaminant from a dataset. It can be used in the following way:

```
ribopicker -f r1_paired.fq.gz -dbs bwa_in-
dexed_rna_db
```

Finally, it could be necessary to run the trimmomatic program again to maintain, after biological contamination cleaning, paired read integrity and remove shorter reads.

6. Pipeline implementation using ENCODE3 shell scripts.

   Stand-alone shell scripts exist for the STAR/RSEM part of the pipeline used here, however the parameters used for mapping are slightly different from the ones described above, corresponding to the ones used in the official ENCODE3 long RNA-seq pipeline. Note that these scripts, partially documented, only quantify annotated genes :

   (a) Making the indices and reference files: https://github.com/ENCODE-DCC/long-RNASeq-pipeline/blob/master/DAC/STAR_RSEM_prep.sh.

   (b) Mapping the reads, making bigwigs, and quantifying annotated genes: https://github.com/ENCODE-DCC/long-RNASeq-pipeline/blob/master/DAC/STAR_RSEM.sh.

7. Pipeline implementation using nextflow

   Nextflow (www.nextflow.io) is a programming language that eases the writing of computational pipelines with complex data. A nextflow implementation of the mapping and quantification parts of the above pipeline, called grape [24], exists and

may be simpler to use than each individual step (although it will only quantify annotated genes). In order for grape to run with STAR for mapping and RSEM for quantification, the "starrsem" profile needs to be used.

8. Using control RNA spike-ins.

   RNA spike-ins are synthetic RNA molecules added to the RNA library in known amounts in order to be able to calibrate the expression measurements of annotated genes (see [25] for an example). In case they are available, it is recommended to use them as additional reference genes, even if current normalization strategies using them do not necessarily perform better than others [26]. This implies a simultaneous mapping the reads to the spike-ins at the same time as to the genome and transcriptome altogether. Similarly, the expression quantification should include the spike-ins in the set of known (or known and novel) genes. This involves slightly modifying some of the above commands:

   – STAR indexing: add $fastaSpikeins after $fastaGenome, where $fastaSpikeins is a FASTA file with the spike-ins, e.g., spikes.fixed.fasta.
   – RSEM reference file generation: replace $fastaGenome by $fastaGenome","$fastaSpikeins.

9. Unstranded protocol:

   If the protocol is unstranded, some of the above commands have to be slightly modified:

   STAR mapping: add --outSAMstrandField intronMotif so the intron motif at the boundary of a split-mapped reads can be used to determine the mapping strand (this option generates the XS strand attribute for all alignments containing a splice junction, and eliminates all the ones with an undefined strand), and replace Stranded by Unstranded in the --outWigStrand option for wiggle file generation.

   Cufflinks reconstruction: replace "--library-type fr-firststrand" by "--library-type fr-unstranded."

   RSEM quantification: remove --forward-prob 0 from the command line.

References

1. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nature 10:57–63
2. Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 488:101–108

3. Dobin A, Davis CA, Schlesinger F et al (2012) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21
4. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515
5. Li B, Ruotti V, Stewart RM et al (2010) RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics 26:493–500
6. T.E.P. Consortium, T.E.P. Consortium, O.C. Data Analysis Coordination et al (2013) An integrated encyclopedia of DNA elements in the human genome. Nature 488:57–74
7. Martens JHA, Stunnenberg HG (2013) BLUEPRINT: mapping human blood cell epigenomes. Haematologica 98:1487–1489
8. Steijger T, Abril JF, Engström PG et al (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 10:1177–1184
9. Engström PG, Steijger T, Sipos B et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10: 1185–1191
10. Roberts A, Goff L, Pertea G et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578
11. Marco-Sola S, Sammeth M, Guigó R et al (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9:1185–1188
12. Pertea M, Pertea GM, Antonescu CM et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295
13. Montgomery SB, Sammeth M, Gutierrez-Arcelus M et al (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464:773–777
14. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10:71–73
15. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 32:462–464
16. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512
17. Sacomoto GAT, Kielbassa J, Chikhi R et al (2012) KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics 13(Suppl 6):S5
18. Rosenbloom KR, Sloan CA, Malladi VS et al (2013) ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 41:D56–D63
19. Harrow J, Frankish A, Gonzalez JM et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22:1760–1774
20. Derrien T, Johnson R, Bussotti G et al (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22:1775–1789
21. Pei B, Sisu C, Frankish A et al (2012) The GENCODE pseudogene resource. Genome Biol 13:R51
22. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
23. Cunningham F, Amode MR, Barrell D et al (2015) Ensembl 2015. Nucleic Acids Res 43:D662–D669
24. Knowles DG, Röder M, Merkel A et al (2013) Grape RNA-seq analysis pipeline environment. Bioinformatics 29:614–621
25. Jiang L, Schlesinger F, Davis CA et al (2011) Synthetic spike-in standards for RNA-seq experiments. Genome Res 21:1543–1551
26. Risso D, Ngai J, Speed TP et al (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 32:896–902

Chapter 15

CRISPR/Cas9 Genome Editing in Embryonic Stem Cells

Guillaume Andrey and Malte Spielmann

Abstract

Targeted mutagenesis is required to evaluate the function of DNA segments across the genome. In recent years the CRISPR/Cas9 technology has been widely used for functional genome studies and is partially replacing classical homologous recombination methods in different aspects. CRISPR/Cas9-derived tools indeed allow the production of a wide-range of engineered mutations: from point mutations to large chromosomal rearrangements such as deletions, duplications and inversions. Here we present a protocol to engineer Embryonic Stem Cells (ESC) with desired mutations using transfection of custom-made CRISPR/Cas9 vectors. These methods allow the in vivo modeling of congenital mutations and the functional interrogation of DNA sequences.

Key words Genome engineering, ES cells, CRISPR, Cas9, Structural variants, Indels, Point mutation, Mouse

1 Introduction

Targeted genetic alterations are methods of choice to functionally assess the role of protein-coding genes, amino acid, noncoding RNA or regulatory regions. They have also shown their importance in modeling human congenital mutations in the mouse and other model animals in the past 20 years. However classical homologous recombination methods and subsequent crossing steps are time-consuming and laborious [1]. In recent years, the development of the CRISPR/Cas9 system has allowed extremely efficient targeted mutagenesis in vitro and in vivo and has eased the access to genetic engineering.

The CRISPR/Cas system is derived from the bacterial type-II CRISPR defense mechanism and is based on the hybridization of a guide RNA to a corresponding target DNA sequence [2]. The guide RNA itself contains the hybridizing part, which is variable, and the Cas9 interacting regions, which allows the recruitment of the Cas9 endonuclease at the hybridization site [3, 4]. Increasing the site-specificity, a three nucleotide PAM sequence "NGG," must be located downstream of the hybridization site. Ultimately, the Cas9 induces a Double Strand Break (DSB) at the target site.

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_15, © Springer Science+Business Media New York 2017

Several tools, which can be implemented in an ESC culture system, have been derived from this powerful prokaryotic defense system [5]. A synthetic guide RNA (sgRNA) that can be programmed to target any sequence of the genome, is made of approximately 100 nt. It is divided in a "directing" region (first 20 nts), which will hybridize to a specific genomic site and a Cas9 recruiting region. Once a DSB is induced at the target site, the DSB repair mechanism of the cell will try to repair it through either the Non-Homologous End Joining repair mechanism (NHEJ) or the Homology Directed Repair mechanism (HDR) [6]. As NHEJ is an error-prone repair mechanism, it may result in the induction of indels, deletions, inversions or duplications at the targeted sites. In contrast, the HDR mechanism can be diverted to increase the efficiency of homologous recombination via a targeting cassette bearing homology arms for the surrounding DNA region [7].

The use of a single sgRNA can thereby induce indels from one to several tens of bps and is thus very useful to functionally test small genomic regions or to induce frameshift mutations in coding exons [8]. Second, as mentioned above, the introduction of a recombination cassette that can be specifically introduced at the breakpoint through HDR, will allow the replacement of an endogenous sequence by another one or the introduction of specific DNA sequence (reporter, tags, etc…) [9]. Third, the production of two DSB at distal locations, through the use of two distal sgRNA, allows the deletion, inversion or duplication of the intermediate DNA fragment [10]. This last method allows the production of so-called Structural Variants (SVs), often found in patients with congenital disabilities or cancer.

In this chapter, we describe how to engineer indels, homologously integrated cassettes, deletions, inversion and duplications in an ESC system. We will start with the design of the directing part of the sgRNA that will define the target site of the CRISPR/Cas9-induced DSB. We will then describe the method to clone the oligonucleotides encoding this directing part of the sgRNA into a Cas9 encoding vector. Subsequently, we propose a 2-week method to obtain any of the above-targeted mutations in ESC culture and a way to select positive clones and to avoid mutagenic byproducts. The production of genetically engineered ESC can be the starting point to obtain engineered mice either through blastocyst ESC transfer or through the morula aggregation technology or for in vitro differentiation assays in ESCs.

2 Material

2.1 Vector and Guide RNA

1. The pSpCas9(BB)-2A-Puro (PX459) vector, can be obtained from Addgene (www.addgene.org).
2. Two oligonucleotides containing the sequences designed from the Zhang lab webtool: http://crispr.mit.edu/, as well as cloning overhangs.

3. T4 DNA ligase supplied with 10× Ligase Reaction Buffer.
4. A heat block for 1.5 ml Eppendorf tubes.
5. Adenosine Tri-Phosphate (ATP) 10 μM.
6. BbsI Restriction Enzyme supplied with 10× buffer.
7. Agarose gel electrophoresis apparatus.
8. Gel extraction kit.
9. ColR oligonucleotide (CACGCGCTAAAAACGGACTA).
10. Ampicillin (100 μg/ml) containing Agarose plate.
11. Ampicillin (100 μg/ml) containing LB medium.
12. Plasmid purification kit.
13. Midi plasmid purification kit.
14. Nanodrop device.
15. Thermocycler.
16. Taq DNA Polymerase with Standard Taq Buffer.
17. BigDye® Terminator v1.1 & v3.1 5× Sequencing Buffer from Life Technology (4336697).
18. Competent *E. coli*.
19. SOC medium.

2.2 Transfection and Selection of Murine Embryonic Stem Cells

1. 0.1% gelatin, in cell culture grade water, 0.22 μm filtered.
2. ES cell incubator.
3. Mitomycin C arrested CD-1 feeder cells.
4. Mitomycin C arrested DR-4 feeder cells.
5. Filtered ES medium (ESM). Knockout Dulbecco's Modified Eagle's Medium (DMEM) 4500 mg/ml glucose, with sodium pyruvate, ES cell tested fetal calf serum (FCS), 100× glutamine, 200 mM, 100× penicillin (5000 U/ml)/streptomycin (5000 μg/ml), 100× nonessential amino acids, 100× fresh 10 mM β-mercaptoethanol (2-ME) (10 mM in PBS, see below), 100× nucleosides.
6. ESM + LIF 0.01% LIF Murine Leukemia Inhibitory Factor ESGROTM (10^7 U/ml).
7. ESM + LIF-Strep without 100× penicillin (5000 U/ml)/streptomycin (5000 μg/ml).
8. Early passage (<15) G4 ES cells.
9. ESM + LIF + puromycin.
10. 8 μg of each cloned px459 plasmid.
11. Reduced Serum Medium.
12. Lipid-based Transfection Reagent.
13. Trypsin: 1× Trypsin-EDTA (0.5 g/l).

14. Centrifuge for 15 ml Falcon tube.
15. Cell culture grade PBS.
16. Plate-Freezing Medium "A": 80% bicarb-free DMEM, 10 mM Hepes, 20% FCS.
17. Plate-Freezing Medium "B": 60% bicarb-free DMEM, 10 mM Hepes, 20% FCS, 20% DMSO.

2.3 ES Cell Clone Lysis and Screening

1. Cell culture grade PBS.
2. Cell Lysis Buffer (10 mM Tris pH 7.3, 10 mM EDTA pH 8.0, 10 mM NaCl, 0.3% Sacrosyl).
3. Proteinase K.
4. Incubator at 60 °C.
5. Magnetic beads.
6. 70% EtOH.
7. 80% EtOH.
8. Double distilled water.
9. 2.0 ml Deep Well Plates.
10. 96-Well Polypropylene Plates.
11. DNA extraction robot.
12. 96 tip comb for DW magnets.
13. Taq DNA Polymerase with Standard Taq Buffer.
14. Thermocycler.
15. Gel electrophoresis device.
16. BigDye® Terminator v1.1 & v3.1 5× Sequencing Buffer.
17. qPCR machine.

2.4 Clone Expansion

1. 0.1% gelatin, in cell culture grade water, 0.22 μm filtered.
2. Mitomycin C arrested CD-1 feeder cells.
3. Waterbath.
4. Filtered ES medium (ESM) (*see* Subheading 2.2, **item 5**).
5. 15 ml falcon tube.
6. ESM + LIF 0.01% LIF Murine Leukemia Inhibitory Factor ESGROTM (10^7 U/ml).
7. Freezing Medium A: ESM with 20% FCS.
8. Freezing Medium B: ESM, 20% FCS, 20% DMSO medium.
9. Lysis buffer (*see* Subheading 2.3, **item 2**).
10. Proteinase K.
11. Incubator at 60 °C.
12. RNAse A.
13. NaCL 5 M.

14. Phenol-Chloroform. (125:24:1 phenol:chloroform:isoamyl alcohol) pH 6.6.
15. Chloroform.
16. Isopropanol.
17. 70% EtOH.
18. Tris (8.06 pKa, −0.028 dpKa/°C, Counter ion chloride) 10 mM pH 8.0.

3 Methods

3.1 Producing the CRISPR Vector

1. sgRNA are designed using the Zhang lab web design tool (http://crispr.mit.edu/) [11]. The desired target DNA sequence is given as input and candidate sgRNAs are proposed, with specific scores, in return. A high score (>80) as well as a low amount of exonic secondary targets (<5) is desirable. The off-targets in exons should have a score under one.
2. The sgRNA oligonucleotide should be synthetized with a "CACC" sequence at it's 5′ end when it is starting with a G, and a CACCG sequence at it's 5′ when it is starting with a A, a C or a T. The complementary oligonucleotide is synthetized starting with a "AAAC" sequence and should finish with a "C" nucleotide complementary to the G in the 5th position of the sense oligonucleotide (*see* Fig. 1).
3. Both oligonucleotide are resuspended to a final concentration of 100 μM and should be annealed in the following conditions: 10 μl oligonucleotide one, 10 μl oligonucleotide two, 10 μl 10× Ligase Reaction Buffer and 70 μl bi-distilled water. The 100 μl mix is heated at 95 °C for 15 min and is let to cool down to Room Temperature (RT) during 45 min.
4. 10 μg of px459 vector is digested with 30 U BbsI (3 μl) in 100 μl of 1× NEB buffer 2.1 for 2 h at 37 °C. The digested plasmid is then run on a 1% agarose gel to separate the digested and undigested forms of the plasmid. The digested plasmid is then extracted from the gel using the QIAquick® Gel Extraction Kit from Qiagen and quantified using Nanodrop (*see* **Note 1**).
5. The sgRNA is then ligated into pX459 vector using the following conditions: 100 ng of the digested pX459 plasmid, 2 ng (2 μl) of annealed oligonucleotides, 2 μl of 10× Ligase Reaction Buffer, 400 U (1 μl) of T4 DNA ligase and filled up to 20 μl with bi-distilled water. The reaction is incubated at RT for 2 h.
6. 5 μl of the ligation reaction are mixed to 100 μl of Top10 chemo-competent cells and heat shocked for 40 s at 42°. Bacterial cells are then supplemented with 500 μl of SOC medium and incubated at 37° for 30 min. The cells are then spin at 1000 × *g* for 5 min, resuspended in 200 μl SOC medium,

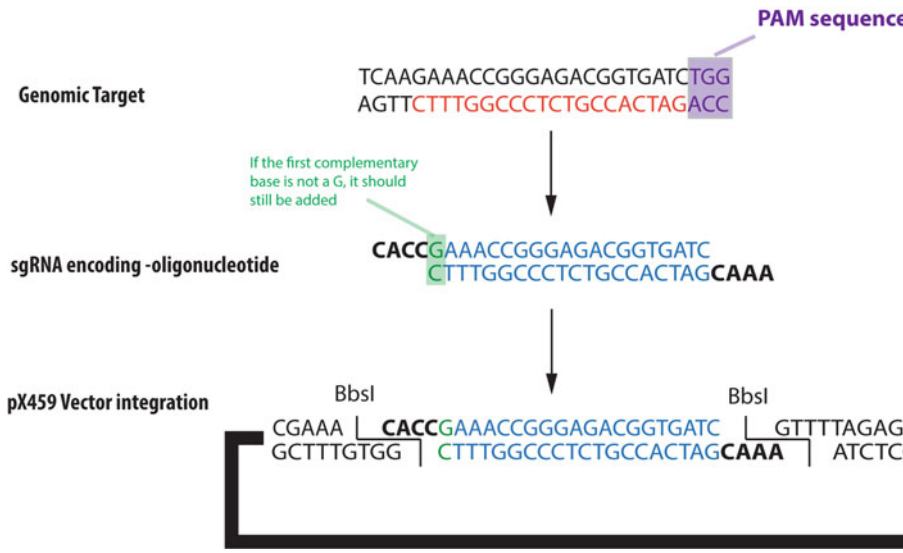


Fig. 1 The genomic target of the sgRNA is a sequence with high specificity that allows a limited amount of off-target sites and contains a "NGG" PAM sequence at its 3′ end. The oligonucleotide encoding the hybridizing sgRNA is 20 nt long and starts with a CACCG (5′ to 3′). The CACC sequence is a sticky overhang for the BbsI cut of the pX459 plasmid. Moreover, a G must follow the CACC, whether it is included in the guide sgRNA hybridization region or not. The complementary oligonucleotide starts with an AAAC (5′ to 3′) also to produce a sticky overhang with the second BbsI cut of the pX459 plasmid

applied to a LB agar plate containing ampicillin and grown overnight at 37 °C.

7. Colonies are subject to PCR using the following condition: a piece of the bacterial colony, 0.4 μl oligonucleotide one, 0.4 μl ColR primer, 2 μl Standard Taq Buffer from NEB, 0.25 μl of Taq DNA Polymerase from NEB, 0.5 μl of dNTPs (10 mM), 16.45 μl bi-distilled water. The thermocycler is set-up as follow: A. 3′ at 94 °C, B. 32 cycles of 45″ at 94 °C, 45″ at 55 °C, 45″ at 72 °C, and C. a final elongation step of 7′ at 72 °C. PCR are subjected to gel electrophoresis and positive ones indicates positive bacterial colonies.
8. Positive colonies are grown either for mini- or midi-prep depending on the required amount of plasmid for the future steps and extracted with NucleoSpin® Plasmid or NucleoBond® Xtra Midi respectively.
9. Sanger sequencing of the plasmid is performed in order to verify the proper sequence of the integrated sgRNA using the following condition: 100 ng of plasmid, 1 μl ColR primer, 1 μl BigDye, 1.5 μl of 5× seq buffer, and fill up to 10 μl with bi-distilled water. The thermocycler is set-up as follow: A. 3′ at 95 °C, B. 25 cycles of 45″ at 94 °C, 45″ at 55 °C, 4′ at 72 °C, and C. a final elongation step of 7′ at 72 °C. The sequence is then determined after electrophoresis on capillaries.

10. For homologous recombination: a targeting cassette containing an alternative mutated sequence or a specific single stranded DNA segment to integrate (e.g., HA tag, loxP site, etc…) should be flanked by two 80–100 nucleotides homology arms. These homology arms should target both sides of the sgRNA cutting site and be depleted of repeated elements (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker).

3.2 ESC Culture

1. Day 1. Cover a well from a 6-well-dish with 2 ml 0.1 % gelatin solution and incubate for 30′ in an ES cell incubator at 37 °C. Discard the gelatin and add 8×10^5 CD-1 feeder cells in prewarmed 3 ml ESM. Gently mix feeders by doing "eights" with the dish on the bench. Incubate overnight to allow cells to attach to the dish (*see* **Notes 2** and **3**).
2. Day 2. Change the medium of the well with 2 ml prewarmed ESM + LIF and add 1 ml of prewarmed ESM + LIF containing 300,000 G4 ES cells (*see* **Note 4**). Gently mix cells by doing "eights." Incubate the cell overnight.
3. Day 3. Two hours before transfection, change the medium with 1.75 ml prewarmed ESM -streptomycin + LIF. The antibiotic can interfere with the proper transformation. Mix solution "A" (8 μg of each cloned px459 plasmid and 100 μl of Opti-MEM medium) on one side and solution "B" (25 μl FuGENE with 100 μl Opti-MEM on the other). When homologous recombination is performed 8 μg of plasmid and 200 ng of the recombination cassette should be mixed in the solution A. Add solution A drop-by-drop to solution B, and mix by flicking the tube. Leave the solution for 15′ at RT and then add it drop-by-drop to the cells. Gently mix by doing "eights." Incubate the cells overnight (*see* **Note 5**).
4. Day 4. Cover four 6 cm dishes with 3 ml 0.1 % gelatin solution and incubate for 30′ in a ES cell incubator at 37 °C. Discard the gelatin and add 1×10^6 DR-4 puromycin-resistant feeder cells in prewarmed 3 ml ES cell medium to each 6 cm dish. Gently mix cells by doing "eights" with the dish on the bench. Incubate overnight to allow the cells to attach to the dish.
5. Day 4. Change the medium with 3 ml ESM + LIF.
6. Day 5. Two hours before splitting the cells: change the medium with 3 ml ESM + LIF. Split the cells: remove the medium and wash two times with 2 ml PBS. Add 1 ml Trypsin and incubate for 15 min at 37 °C. Resuspend the cells in 5 ml ESM –LIF and spin them at $260 \times g$ for 5 min. Discard the supernatant and resuspend the pellet in 12 ml ESM + LIF + PURO. Add 3 ml of the resuspended cells to each of the four dishes containing DR4-feeder cells after removing the old medium.
7. Day 6. Change the medium with 3 ml ESM + LIF + PURO.

8. Day 7. Change the medium with 3 ml ESM + LIF.
9. Day 8. Change the medium with 3 ml ESM + LIF.
10. Day 9. Change the medium with 3 ml ESM + LIF.
11. Day 10. Change the medium with 3 ml ESM + LIF.
12. Day 10. Cover the corresponding amount of 96-well plates needed for the amount of clone to pick with 100 µl 0.1 % gelatin solution per well. Incubate for 30′ in an ES cell incubator at 37 °C. Discard the gelatin and add 1×10^4 CD-1 feeder cells in prewarmed 150 µl ES cell medium to each well. Gently mix feeders by doing "eights" with the dish on the bench. Incubate overnight to allow the cells to attach to the dish (*see* **Notes 6** and 7).
13. Day 11. Two hours before the clone picking, change the medium in the 6 cm dishes with 3 ml ESM + LIF. Change the medium in each 96-well plate with 100 µl of ESM + LIF. Prepare, round-bottom 96-well plates with 30 µl trypsin in each well. Clone picking: remove the medium from the clone-containing 6 cm dishes and wash two times with 2 ml PBS. Leave 2 ml PBS to pick clones. Transferred each picked clone with a tip in a well containing 30 µl trypsin and incubate every 24 picked-clones for 12′ at 37 °C. Resuspend the clones with 60 µl ESM + LIF and transfer the 90 µl to the 100 µl of the CD-1 containing 96-well plates using at 12-channel pipette. Start again until the aimed number of clones is picked. Let the clones attached overnight.
14. Day 12. Change the medium with 150 µl ESM + LIF in every wells.
15. Day 13. Change the medium with 150 µl ESM + LIF in every wells.
16. Day 14. Change the medium with 150 µl ESM + LIF in each well 2 h before split and freeze. Split and freeze: wash every well two times with 100 µl PBS. Add 30 µl trypsin to every well and incubate for 10′ at 37 °C. Resuspend each well with 100 µl of Plate-Freezing Medium "A." Transfer two times 50 µl in each well of two round-bottom 96-well already containing 50 µl of Plate-Freezing Medium "B" and mix by pipetting. Transfer the round-bottom 96-well plates into a styrofoam box in an −80 °C freezer. Add 200 µl ESM to the remaining 30 µl in each 96-wells of the flat bottom plate. Incubate for 3–4 days to obtain enough cells for genotyping.

3.3 ES Cell Clone Lysis and Genotyping

1. When clones are sufficiently confluent, wash every well two times with 100 µl PBS and add 50 µl Cell Lysis Buffer complemented with Proteinase K (1 mg/ml final). The 96-well plate should be sealed with a sticking plastic foil. The lysis is performed overnight at 60 °C.

2. Every 50 µl lysis mix are added to a 96 deep-well plate containing 8 µl MaggAttract beads and 27.5 µl 70% EtOH. Following a standard protocol, the Kingfisher Flex robot incubates beads with the lysis mix for 15 min and washed them two times for 5 min in 80% EtOH. After a drying step of 5′, the DNA is eluted in 50 µl of bi-distilled water in a 96 well plate.
3. The extracted DNA is then used to PCR-screen the desired mutation. To do so primers should be designed accordingly to the type of engineered mutation using Primer3 (*see* Fig. 2 for detail). Primers should have an annealing temperature of 58 ±2 °C. Primers should be BLAST using Ensembl to avoid multiple hits (>10) in the genome. For the screening, a control PCR at an unrelated locus, should be done in parallel to the PCR detecting the desired mutant allele, to account for the gDNA template quality.
4. The PCR is done as follows: 20–50 ng template genomic DNA, 0.75 µl primer fw (10 mM stock), 0.75 µl primer rev (10 mM stock), 2 µl Standard Taq Buffer from NEB, 0.25 µl Taq DNA Polymerase from NEB, 0.5 µl of dNTPs (10 mM), 16.45 µl bi-distilled water. PCR cycling as follows: A. 3′ at 94 °C, B. 35 cycles of 45″ at 94 °C, 45″ at 58 °C, 1′ at 72 °C, and C. a final elongation step of 7′ at 72 °C. PCRs are subjected to gel electrophoresis and bands indicate positive clones (*see* **Note 8**).
5. Clones bearing the desired mutations should be screened for a variety of unwanted mutations, i.e., accompanying indels, inversions, deletions or duplications. The targeted site(s) should be assessed in both homologous chromosomes. Finally the PCR products spanning the mutated area should be Sanger-sequenced using the forward or reverse primer, following the same protocol as in Subheading 3.1, **step 10** (*see* **Note 9**).
6. As ES cells are grown with CD-1 feeder cells, it is difficult to assess the homozygote loss of a wild-type PCR product. So, to determine if the indels, deletions or inversions are occurring on both allele of the ESC, Quantitative real-time PCR (qPCR) should be used. Two control primer pairs outside the rearranged area in combination with two primer pairs measuring the quantity of wild-type DNA in the deleted, inverted (for inversion, the qPCR product should span the cutting site) or indel (for indels, one of the primers should anneal at the cutting site) region (*see* Fig. 2) are required. A 50% loss is indicative of a heterozygote allele, loss above 90% is indicative of a homozygous allele.

3.4 Clone Expansion

1. When positive clones have been identified, one needs to amplify them in order to work with them. We recommend here to expand at least three positive clones per desired allele. Cover the corresponding amount of wells of a 96-well plate to the number of clones that should be expanded with 100 µl 0.1%

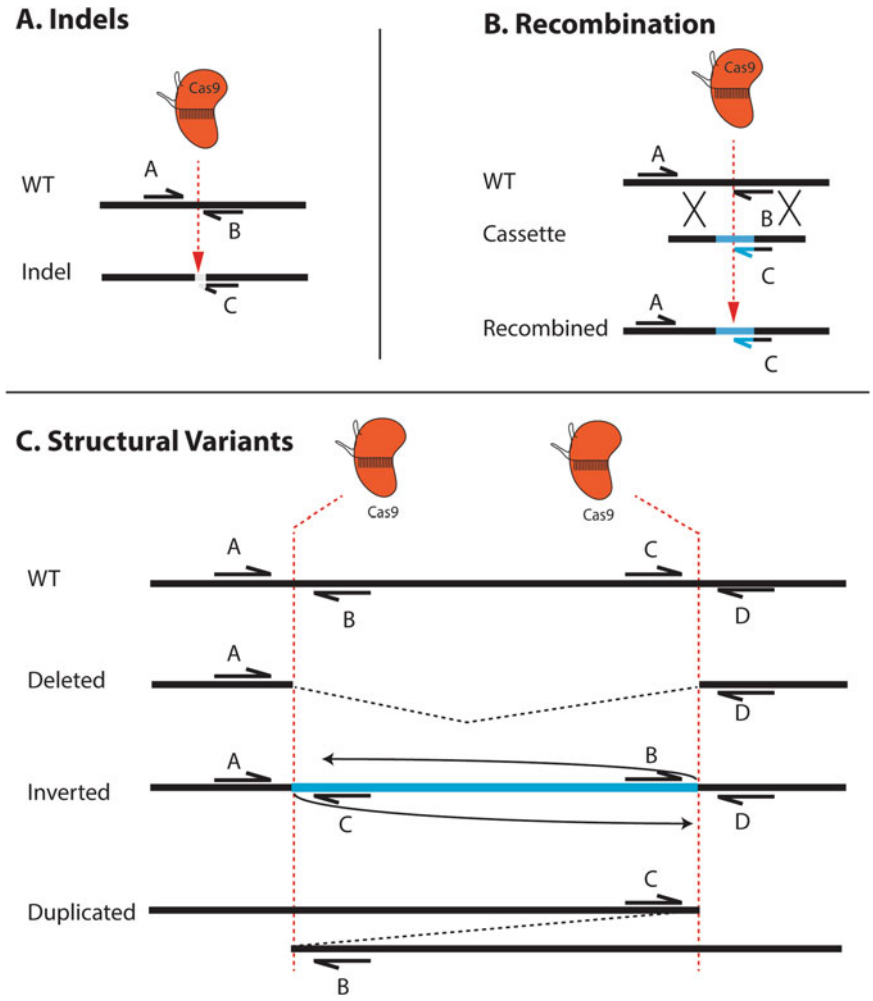


Fig. 2 PCR-genotyping of different types of CRISPR/Cas9 engineered mutations. A control wild-type PCR must be designed to amplify the CRISPR/Cas9 target site, here A + B or C + D for structural variants. (**a**) Indels are produced using a single sgRNA, the directed Cas9 enzyme will introduce a DSB that will ultimately allow the production of small indels at the cutting site. To genotype such kind of alleles, one uses either two primers to amplify the targeted (A + B) site and Sanger-sequence it, either one external primer or a primer at the cutting site with the desired mutated sequence (A + C), in order to only pick clones with a specific mutation. (**b**) To detect a specific sequence introduced by a donor DNA at the cutting site through homologous recombination, one can use a primer outside of the homology region and a specific one in the cassette to generate a specific PCR product (A + C). (**c**) For structural variants, four primers are needed: two facing one another on each side of the first cutting site and two others at the second cutting site. By changing the primer pair's combination, one can produce specific PCRs for deletions (A + D), inversions (A + C or B + D) and duplications (C + D)

gelatin solution and incubate for 30′ in a ES cell incubator at 37 °C. Remove the gelatin and add 1×10^4 CD-1 feeder cells in prewarmed 200 μl ES cell medium per well. Incubate overnight to allow the cells to attach to the dish.

2. Once the feeder cells have attached, unfreeze the positive clones in a 37 °C waterbath, transfer the 100 μl freezing

medium and cells to 2 ml of ESM –Lif in a 15 ml falcon tube. Spin the tube for 5′ at 260 × *g* and remove the supernatant. Resuspend the pellet in 200 µl of ESM + LIF. Replace the 200 µl ESM of the CD-1 feeder cells with the 200 µl ES + LIF + clones (*see* **Note 10**).

3. Change the ESM + LIF every day.
4. When ESC reach 50% of confluence, cover the equivalent number of wells in a 24-well plate with 500 µl 0.1% gelatin solution and incubate for 30′ at in a ES cell incubator at 37 °C (*see* **Note 11**). Remove the gelatin and add 8×10^4 CD-1 feeder cells in prewarmed 500 µl ESM per well. Incubate overnight to allow the cells to attach to the dish.
5. When ESC reach 80% of confluence, change ESM + LIF 2 h before splitting (*see* **Notes 12** and **13**). Replace the ESM medium of the wells of the 24 well plates with 500 µl ESM + LIF. Splitting: wash ESC twice with PBS and add 50 µl trypsin. Incubate for 12′ at 37 °C. Resuspend the cells in 200 µl ESM + LIF and transfer them to the wells of the 24-well plate.
6. Change the ESM + LIF every day.
7. When ESC reach 50% of confluence, cover the equivalent number of wells in a 6-well plate with 1 ml 0.1% gelatin solution and incubate for 30′ in a ES cell incubator at 37 °C. Discard the gelatin and add 4×10^5 CD-1 feeder cells in prewarmed 2 ml ESM per well. Incubate overnight for the cells to attach to the dish.
8. When ESC reach 80% of confluence change ESM + LIF 2 h before splitting. Replace the ESM medium of the wells of the 6 well plates with 2 ml ESM + LIF. Splitting: wash ESC twice with cell culture grade PBS and add 200 µl trypsin. Incubate for 12′ at 37 °C. Resuspend the cells in 1 ml ESM + LIF and transfer them to the wells of the 24-well plate.
9. Change the ESM + LIF every day.
10. When ESC reach 50% of confluence, cover the equivalent number 6 cm dishes with 3 ml 0.1% gelatin solution and incubate for 30′ in a ES cell incubator at 37 °C. Discard the gelatin and add 1×10^6 CD-1 feeder cells in prewarmed 3 ml ESM per well. Incubate overnight for the cells to attach to the dish.
11. When ESC reach 80% of confluence change ESM + LIF 2 h before splitting. Replace the ESM medium of the wells of the 6 cm dishes with 3 ml ESM + LIF. Splitting: wash ESC twice with cell culture grade PBS and add 600 µl trypsin. Incubate for 12′ at 37 °C. Resuspend the cells in 3 ml ESM + LIF and transfer them to the wells of the 24-well plate.
12. When ESC reaches 80% of confluence, change ESM + LIF 2 h before freezing. Freezing: wash with 4 ml cell culture grade PBS and add 1 ml trypsin. Incubate for 12’ at 37 °C. Resuspend the

cells in 3 ml ESM and transfer them to 15 ml falcon tube. Spin cells for 5′ at 260 × *g* and discard the flow through. Cell can be counted using a cell counter at this step to insure an ideal freezing concentration. We recommend to freeze 1×10^6 ESC per vial (*see* **Note 14**). An 80% confluent 6 cm plate should contain 1×10^6 ES cells. Resuspend the pellet in 1.8 ml Freezing Medium A. For each clone prepare three freezing vials with 500 µl Freezing Medium B. To each of the vial add 500 µl of freezing medium A containing the cells. Invert the tube and transfer them to a freezing container. Transfer the container to an −80 °C freezer. After 2 days, transfer the vials to a cryobox and freeze them in liquid nitrogen. Transfer the remaining 300 µl of Freezing Medium A and cells to a gelatin-coated well of a 12 well plate. Leave the cells to attach and once they reach 90% confluence lyse them in 500 µl lysis buffer with protK overnight at 60 °C. Treat with 20 µg RNAseA for 30′ at 37 °C. Add 200 µl of NaCL 5 M and spin at 11,000 × *g* for 10′. Transfer the supernatant to a new eppendorf tube and add 700 µl of Phenol-Chloroform and mix it vigorously. Spin at 11,000 × *g* and transfer the upper phase in 700 µl of chloroform. Spin at 11,000 × *g* and transfer the upper phase to a new eppendorf tube. Precipitate for 20′ at −80 °C the genomic DNA by adding 500 µl of isopropanol. Spin at 11,000 × *g* for 20′ at 4 °C and wash two times the pellet with 70% EtOH. Dry the pellet at 37 °C for 20' and resuspend it in 10 mM Tris pH 8.0.

13. The genomic DNA containing solution is used to confirm the molecular characterization of each clone. The same set of PCR, qPCR, and Sanger sequencing described in Subheading 3.3, **steps 6** and **7** should be performed to validate the genetic content of the clones.

4 Notes

1. The px459 vector carries a puromycin resistance and is optimal for this protocol, but other vectors are also available at: http://crispr.mit.edu/.
2. When CD-1, DR-4 or ES cells are thawed one must remove the freezing media containing DMSO before proceeding to subsequent steps. Cells in their freezing media should thus be mixed with 5× their volume of ESM, spin at 260 × *g* for 5′, and resuspended in the desired growth media.
3. It is essential to evenly distribute the feeder cells in the dish to avoid clusters or "holes" in the feeder cell layer. For this purposed we recommend to mix the plates by doing "eights" with them.
4. Check feeder cell density before seeding out G4 ES cells. An optimal feeder cell layer will influence the outcome of the experiment. If the feeder cell layer is not dense enough new feeders can be added as a complement, at least 2 h before seeding the G4 ES cells.

5. It is important to have a correct estimation of the transfection efficiency, as it will considerably influence the clones' density. A too large amount of clones will be hard to pick and too few clones will impact the probability of finding the right one.
6. The growth and amount of clones may also vary from experiment to experiment. As for the growth, clones can be grown up to 2–3 extra days in order to have the perfect size for picking. The amount of clones may vary depending on the transfection efficiency mostly. It is thus very important to rigorously follow the transfection protocols.
7. Targeting efficiencies may vary depending on the size and the type of desired mutation. For deletions, inversions or duplications of small size (below 10 kb) we recommend picking a minimum number of 200 clones. For larger size structural variants as well as for recombined alleles a minimal number of 400 clones is desirable. In contrast, 100 clones are enough if one aims at small indels.
8. The genotyping may show no bands or very weak bands. If a WT PCR did produce a strong band one can assume that the quality of the template DNA is good. It is then possible that the primer combination used to detect the mutation is incompatible. If so, we recommend designing a new primer set. If the PCR fails again it is very likely that the experiment did not produce the desired allele.
9. The genotyping may show mixture of WT and mutant bands. The PCR that is supposed to detect the WT band can be shifted compared to the WT size. This is due to the production of indels at both targeting sites. CRISPR/Cas is an extremely efficient method and one can hardly find WT alleles in ES cells after they have been exposed to the endonuclease.
10. Individual clones might grow at different speed. Thereby, it is important to consider the confluence of clones, as well as their size prior to splitting rather than the absolute time they have spent in culture.
11. The confluence of clones is a good indicator for the amount of cells one can expect from a particular dish. However to evaluate their quality one should observe other aspects. Good and healthy clones are rounded, have clear edges and do not bear small dotes.
12. Clones may grow very slowly. Many factors can influence the growth of clones. The first one is the density of CD-1 feeder cells. A too dense or to scattered CD-1 layer will decrease the speed of growth. It is thus important to seed the right amount of cells. The second possibility is that freshly thawed ESC clones might need time to start growing. In such a case we recommend to split the cells on the same surface (96-well plate) until a good density of clones is reached.

13. If one has very few colonies growing on the plate, it is important to include a step where clones are trypsinized and reseeded on the same well size.
14. To obtain the correct freezing density one need to estimate the amount of clones that are present on the well. ESC clones in the final 6 cm dish should reach 80% confluence and have an intermediate size. It is vital that approximately 1,000,000 cells are frozen in each vial. To insure this optimal concentration we recommend counting cells using a cell counter.

Acknowledgments

We thank all members of the Mundlos lab in particular Katerina Kraft for sharing their experiences and fruitful discussions about the protocol. We also thank Heiner Schrewe and Lars Wittler for helping us to establish the ESC culture.

References

1. Sternberg SH, Doudna JA (2015) Expanding the Biologist's Toolkit with CRISPR-Cas9. Mol Cell 58:568–574
2. Marraffini LA (2015) CRISPR-Cas immunity in prokaryotes. Nature 526:55–61
3. Mali P, Esvelt KM, Church GM (2013) Cas9 as a versatile tool for engineering biology. Nat Methods 10:957–963
4. Nishimasu H, Ran FA, Hsu PD et al (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell 156:935–949
5. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 32:347–355
6. Lieber MR (2010) The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem 79:181–211
7. Wang H, Yang H, Shivalila CS et al (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153:910–918
8. Yang H, Wang H, Jaenisch R (2014) Generating genetically modified mice using CRISPR/Cas-mediated genome engineering. Nat Protoc 9:1956–1968
9. Yang H, Wang H, Shivalila CS et al (2013) One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. Cell 154:1370–1379
10. Kraft K, Geuer S, Will AJ, Chan WL, Paliou C, Borschiwer M, Harabula I, Wittler L, Franke M, Ibrahim DM, Kragesteen BK, Spielmann M, Mundlos S, Lupiáñez DG, Andrey G.Cell Rep. 2015 Feb 4. pii: S2211–1247(15)00029-7. doi: 10.1016/j.celrep.2015.01.016
11. Hsu PD, Scott DA, Weinstein JA et al (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 31:827–832

Chapter 16

Targeted Gene Activation Using RNA-Guided Nucleases

Alexander Brown, Wendy S. Woods, and Pablo Perez-Pinera

Abstract

The discovery of the prokaryotic CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR-associated) system and its adaptation for targeted manipulation of DNA in diverse species has revolutionized the field of genome engineering. In particular, the fusion of catalytically inactive Cas9 to any number of transcriptional activator domains has resulted in an array of easily customizable synthetic transcription factors that are capable of achieving robust, specific, and tunable activation of target gene expression within a wide variety of tissues and cells. This chapter describes key experimental design considerations, methods for plasmid construction, gene delivery protocols, and procedures for analysis of targeted gene activation in mammalian cell lines using CRISPR-Cas transcription factors.

Key words Genome engineering, Synthetic biology, RNA-guided nucleases, CRISPR-Cas9, Gene expression, Gene activation, Transcription

1 Introduction

Robust and controllable systems for activation of native gene expression have been pursued for multiple applications in gene therapy, regenerative medicine and synthetic biology. These systems, rather than introducing heterologous genes that are expressed from constitutive or tunable promoters, use proteins that regulate transcription of genes in their natural chromosomal context. There are several advantages to activating native gene expression compared with overexpressing exogenous genes including ease of cloning, simple delivery, tunability and potential for simultaneous regulation of multiple gene splicing isoforms.

There are multiple approaches to controlling native gene expression [1–4], however recent advances in genetic engineering have made it possible to rapidly design and assemble artificial transcription factors (ATFs) that are both efficient and highly specific. One key feature of ATFs is that they typically have a modular structure, with two distinct and independent domains: (1) a DNA-binding domain, and (2) a transcriptional activation domain. Through customization of the DNA binding and transcriptional activation domains, it is possible to

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_16, © Springer Science+Business Media New York 2017

select a genomic target and activate gene expression exclusively at that locus [5–7]. Whereas the exact mechanism of transcriptional activation by ATFs has not been clearly established, it is widely accepted that the DNA-binding domain localizes the ATF to the target promoter, where the activation domain recruits transcription preinitiation complexes and activates gene expression.

First generation transcriptional activation domains are relatively weak and require binding of multiple ATFs in close proximity, within the promoter, in order to function synergistically and efficiently initiate transcription [8, 9]. However, there is now a wide range of second-generation transcriptional activation domains that can facilitate high levels of gene activation, even when using a single ATF [6, 10–12] (Table 1).

Artificial transcription factors are classified according to the nature of the DNA-binding domain in three main groups: Zinc Finger Proteins (ZFP), Transcriptional Activator-Like Effectors (TALEs), and RNA-guided nucleases (RGNs) [13]. Each group differs significantly, not only structurally, but also in development cost, and assembly difficulty. While each of these ATFs is effective at activating native gene expression, RGNs have begun to dominate the field of synthetic gene regulation. Arguably, the feature that most contributes to the rising popularity of RGNs is the simplicity of their assembly procedures. By avoiding the numerous challenges of protein engineering required for generating the DNA-binding domains

Table 1
Summary of transcriptional activators commonly used in artificial transcription factors to stimulate gene expression

Transcriptional activation system	Notes	References
NFkB/p65	Transcriptional activator	[28]
VP16	Transcriptional activator	[29]
VP64	Four Tandem repeats of the minimal activation domain of VP16	[30]
CIB1-Cry2	Light inducible system. ATF-CIB1 is used with CRY2-VP64	[31, 32]
GI-LOV	Light inducible system. ATF-GI is used with LOV-VP16	[33]
GCN4 peptide (10× or 24×)	SunTag System[a]	[34]
p300 HAT core	Epigenetic modifier[a]	[11]
VPR	Tripartite VP64, p65, and Rta[a]	[12]
SAM	Modified sgRNA used to recruit multiple effector domains[a]	[6]

[a]These transcriptional activation systems have been shown to efficiently activate native gene expression even when using a single ATF

of ZFPs and TALEs, RGNs both hasten and simplify the construction of ATFs with novel sequence specificity. Furthermore, RGNs offer robust transcriptional activation that is flexible with regard to the fine-tuning of gene expression [7, 14–18].

This review will describe a protocol for activation of *ASCL1* expression using RGNs consisting of *S. pyogenes* Cas9 and single-guide RNAs [7] (Fig. 1). In *Streptococcus pyogenes*, clustered regularly interspaced short palindromic repeats (CRISPR) RNAs (crRNAs) are expressed in conjunction with a scaffold RNA, known as the trans-activating-crRNA (tracrRNA), and guide Cas9

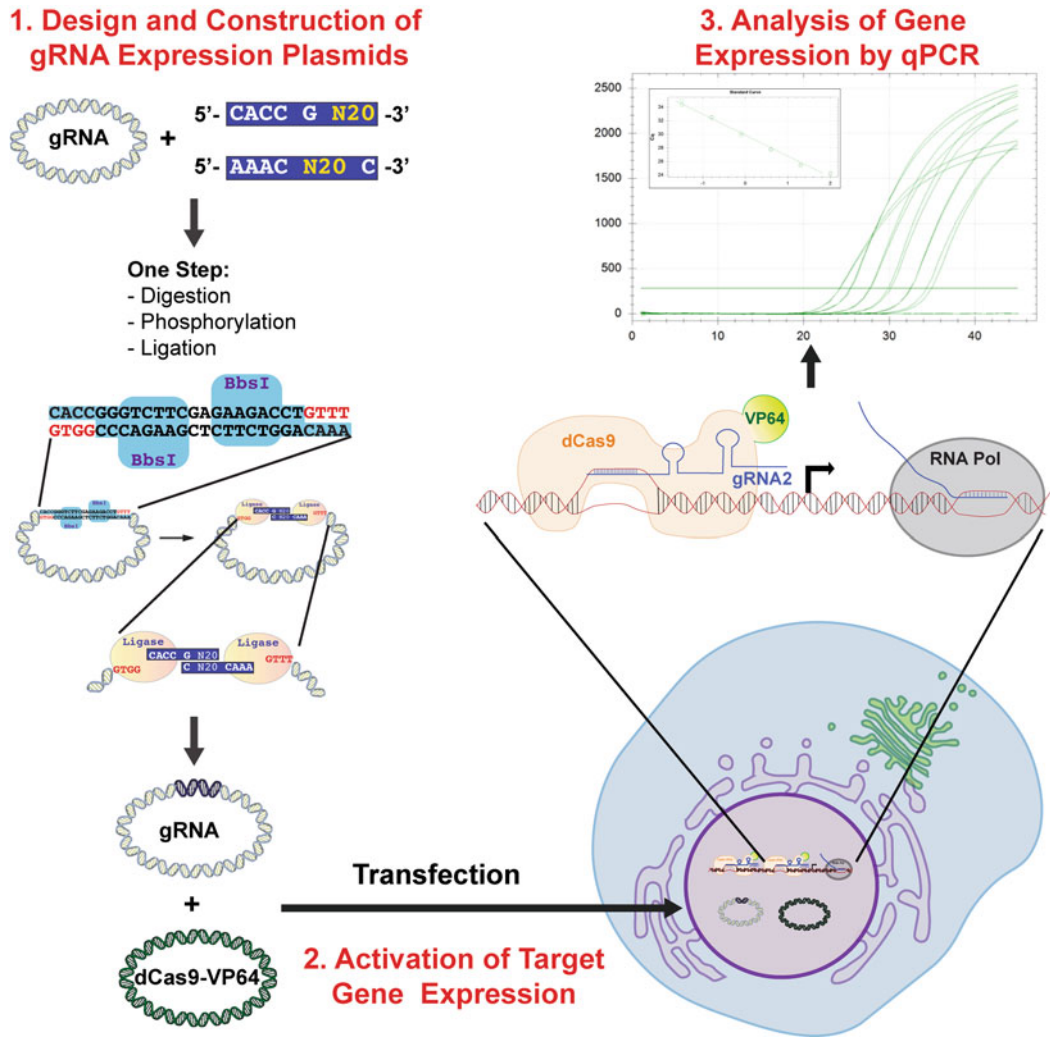


Fig. 1 Schematic representation of the overall procedure for gene activation using RGNs described in this manuscript. This method consists of three stages: (1) sgRNA expression vectors are designed and generated using a single-step digestion, phosphorylation, and ligation reaction, (2) native gene expression is activated by co-delivery of sgRNA and dCas9-transcriptional activator expression plasmids into the target cells, and (3) RNA is isolated and analyzed using qPCR to quantify relative changes in gene expression

to the target DNA. The only constraint for target sequences is that they must immediately precede a suitable protospacer adjacent motif (PAM) of the form NGG [19]. The bacterial CRISPR system has been further simplified to utilize a single-guide RNA molecule (sgRNA), which functions as a chimeric RNA to replace both the crRNA and tracrRNA elements [20, 21]. Furthermore, the native *S. pyogenes* Cas9 has been engineered to work within many eukaryotic systems, including mammalian cells, by delivering expression plasmids of codon-optimized Cas9 cDNA containing one, or more, nuclear localization signals (NLS). Point mutations in amino acids D10 and H840 of Cas9 render the enzyme catalytically inactive (dCas9) [22], providing a programmable DNA-binding protein without nuclease activity. Several groups have demonstrated that dCas9 can function as an effective ATF by fusion with transcriptional activation domains [7, 14–18].

The following protocol for designing, assembling and testing RGN transcription factors assumes that a dCas9-transcriptional activator has already been obtained. Many systems for gene activation have been described recently and several of them are available through the Addgene plasmid repository (Addgene.org). To aid the identification of a suitable activation system, we summarize in Table 2 different dCas9-transcriptional activators compatible with the gene activation system described here.

2 Materials

2.1 Construction of sgRNA Expression Plasmids

1. An appropriate sgRNA vector should be chosen prior to guide design (*see* **Note 1** and Table 3 for review of suitable plasmids available from Addgene.com). This protocol assumes the use of pSPgRNA (Addgene #47108), which includes two *BbsI/BpiI* sites interspaced between a human U6 promoter and the sgRNA loop for cloning of oligonucleotides (Fig. 1).
2. Oligonucleotides for sgRNA construction.

   Target selection: The identification of optimal target sites for activation of gene expression remains, essentially, an empirical process. It has been shown in multiple studies that the region comprising −400 to −50 bp at the 5′ end of the transcriptional start site (TSS) is optimal [10]. Since the TSS is clearly annotated in most genome browsers, the sequence of the gene of interest is imported into DNA analysis software and used to identify potential target sites. We typically use Benchling [23], a freely available web-based DNA analysis platform that incorporates a "Genome Engineering" tool to identify all possible sgRNAs within any sequence specified by the user (*see* **Note 2**).

   The target sequences chosen to activate *ASCL1* gene expression are:

Table 2
Summary of constructs encoding dCas9-transcriptional activators that are publicly available through Addgene for stimulation of gene expression in mammalian cells

Plasmid name	Addgene #	Promoter	Transcriptional activation domain	Reference
SP-dCas9-VPR	63798	CMV	VPR (VP64-p65-Rta)	[12]
pcDNA-dCas9-p300 Core	61357	CMV	p300 Core (human, aa 1048-1664)	[11]
pcDNA-dCas9-VP64	47107	CMV	VP64	[7]
pAC93-pmax-dCas9VP160	48225	CAGGS	VP160	[16]
pAC91-pmax-dCas9VP64	48223	CAGGS	VP64	[16]
pAC92-pmax-dCas9VP96	48224	CAGGS	VP96	[16]
pSL690	47753	CMV	VP64	[15]
pCMV_dCas9_VP64	49015	CMV	VP64	[18]
CMVp-dCas9-3xNLS-VP64 Construct 1	55195	UBC	VP64	[35]
pMSCV-LTR-dCas9-p65AD-BFP	46913	MSCV LTR	p65AD	[14]
pMSCV-LTR-dCas9-VP64-BFP	46912	MSCV LTR	VP64	[14]
EF_dCas9-VP64	68417	EF1a	VP64	[36]
pHAGE TRE dCas9-VP64	50916	TRE	VP64	[37]
pHAGE EF1α dCas9-VP64	50918	EF1a	VP64	[37]
dCAS9-VP64_GFP	61422	EF1a	VP64	[6]
lenti dCAS-VP64_Blast	61425	EF1a	VP64	[6]
pHRdSV40-NLS-dCas9-24xGCN4_v4-NLS-P2A-BFP-dWPRE	60910	SV40	GCN4/SunTag system	[34]

5′-GCTGGGTGTCCCATTGAAA-3′.
5′-CAGCCGCTCGCTGCAGCAG-3′.
5′-TGGAGAGTTTGCAAGGAGC-3′.
5′-GTTTATTCAGCCGGGAGTC-3′.

For each target sequence, a sense oligonucleotide is generated in the format: 5′-CACC G X_{20}-3′, where X_{20} represents the 20 bases of the genomic DNA at the 5′ end of the PAM (*see* **Note 3**). The first four bases are complementary to the sgRNA vector overhangs, while the fifth base is G in order to initiate transcription of RNA from the upstream U6

Table 3

Summary of vectors that are available through Addgene for cloning and expression of custom sgRNAs using methods similar to that described in this manuscript

Plasmid name	Addgene #	Promoter	Cloning enzyme(s)	Reference
gRNA_Cloning Vector	41824	Human U6	AflII	[21]
pLKO5.sgRNA.EFS.GFP	57822	U6	BsmBI	[38]
pLKO5.sgRNA.EFS.tRFP	57823	U6	BsmBI	[38]
pLKO5.sgRNA.EFS.tRFP657	57824	U6	BsmBI	[38]
pLKO5.sgRNA.EFS.PAC	57825	U6	BsmBI	[38]
pSPgRNA	47108	Human U6	BbsI	[7]
phH1-gRNA	53186	Human H1	BbsI	[39]
pmU6-gRNA	53187	Mouse U6	BbsI	[39]
phU6-gRNA	53188	Human U6	BbsI	[39]
ph7SK-gRNA	53189	Human 7SK	BbsI	[39]
pHL-H1-ccdB-mEF1a-RiH	60601	H1	BamHI/EcoRI	[40]
pUC57-sgRNA expression vector	51132	T7	BsaI	[41]
pGL3-U6-sgRNA-PGK-puromycin	51133	Human U6	BsaI	[41]
pUC-H1-gRNA	61089	H1	BsaI	[42]
pAC155-pCR8-sgExpression	49045	Human U6	BbsI	[16]
pSQT1313	53370	Human U6	BsmBI	[43]
BPK1520	65777	Human U6	BsmBI	[44]
pU6_gRNA_handle_U6t	49016	U6	SacI	[18]
pGuide	64711	Human U6	BbsI	[45]
pgRNA-humanized	44248	Mouse U6	BstXI + XhoI	[46]

(continued)

Table 3
(continued)

Plasmid name	Addgene #	Promoter	Cloning enzyme(s)	Reference
pLX-sgRNA	50662	Human U6	OE-PCR	[47]
pLenti-sgRNA-Lib	53121	Human U6	BsmBI	[48]
pU6-sgRNA EF1Alpha-puro-T2A-BFP	60955	Mouse U6	BstXI + BlpI	[10]
pLKO.1-puro U6 sgRNA BfuAI stuffer	50920	Human U6	BfuAI	[37]
+pKLV-U6gRNA(BbsI)-PGKpuro2ABFP	50946	Human U6	BbsI	[49]
pH1v1	60244	H1	Gibson	[50]
lentiGuide-Puro	52963	Human U6	BsmBI	[51]
AAV:ITR-U6-sgRNA(backbone)-pEFS-Rluc-2A-Cre-WPRE-hGHpA-ITR	60226	U6	SapI	[52]
AAV:ITR-U6-sgRNA(backbone)-pCBh-Cre-WPRE-hGHpA-ITR	60229	U6	SapI	[52]
AAV:ITR-U6-sgRNA(backbone)-hSyn-Cre-2A-EGFP-KASH-WPRE-shortPA-ITR	60231	U6	SapI	[52]
PX552	60958	Human U6	SapI	[53]
sgRNA(MS2) cloning backbone	61424	U6	BbsI	[6]
lenti sgRNA(MS2)_zeo backbone	61427	U6	BsmBI	[6]
pAC2-dual-dCas9VP48-sgExpression	48236	Human U6	BbsI	[16]
pAC5-dual-dCas9VP48-sgTetO	48237	Human U6	BbsI	[16]
pAC152-dual-dCas9VP64-sgExpression	48238	Human U6	BbsI	[16]
pAC153-dual-dCas9VP96-sgExpression	48239	Human U6	BbsI	[16]
pAC154-dual-dCas9VP160-sgExpression	48240	Human U6	BbsI	[16]

promoter. A second oligonucleotide, representing the antisense target sequence, is generated in the format: 5′-AAAC Y_{20} C-3′. Here, AAAC are vector-complementing overhangs, Y_{20} represents the reverse complement of the target sequence,

and the last C complements the leading G of the sense oligonucleotide (Fig. 1).

The sequences of the oligonucleotides for assembly of sgRNAs that target the *ASCL1* promoter are:

TARGET1S: 5′-**CACC G** GCTGGGTGTCCCATTGAAA-3′.

TARGET1AS: 5′-**AAAC** TTTCAATGGGACACCCAGC **C**-3′.

TARGET2S: 5′-**CACC G** CAGCCGCTCGCTGCAGCAG-3′.

TARGET2AS: 5′-**AAAC** CTGCTGCAGCGAGCGGCTG **C**-3′.

TARGET3S: 5′-**CACC G** TGGAGAGTTTGCAAGGAGC-3′.

TARGET3AS: 5′-**AAAC** GCTCCTTGCAAACTCTCCA **C**-3′.

TARGET4S: 5′-**CACC G** GTTTATTCAGCCGGGAGTC-3′.

TARGET4AS: 5′-**AAAC** GACTCCCGGCTGAATAAAC **C**-3′.

3. Nuclease-free Molecular biology grade (MBG) water.
4. Tris Buffered Saline (TBS), 50 mM Tris pH 7.4 and 150 mM NaCl.
5. Restriction endonuclease *BbsI/BpiI* (*see* **Note 4**).
6. T4 Polynucleotide Kinase (PNK).
7. T4 DNA ligase and T4 DNA Ligase Buffer with ATP (*see* **Note 5**).
8. Transformation-competent *E. coli* (*see* **Note 6**).
9. LB-Agar plates containing 100 μg/mL carbenicillin for bacterial culture.
10. KAPA2G Robust PCR Kit (KAPA Biosystems) and 10 mM dNTP mix.
11. Sequencing and colony PCR primer, M13 Forward: 5′-TGTAAAACGACGGCCAGT-3′.
12. Ethidium bromide, 10 mg/mL.
13. Electrophoresis Buffer (TAE) 40 mM Tris pH 7.2, 20 mM Acetate, and 1 mM EDTA.
14. Agarose.
15. LB broth containing 100 μg/mL carbenicillin.
16. Qiagen Spin Miniprep Kit.

2.2 Activation of Target Gene Expression

1. Mammalian cell line, such as HEK293T.
2. Phosphate-buffered saline (PBS), 8 mM Na_2HPO_4, 2 mM KH_2PO_4 pH 7.4, 137 mM NaCl and 2.7 mM KCl.
3. 0.25% Trypsin-EDTA.
4. Complete mammalian cell culture medium appropriate for the chosen cell line, such as DMEM supplemented with 10% Fetal Bovine Serum (FBS) and 1% penicillin/streptomycin.
5. Lipofectamine 2000 (Thermo Fisher Scientific) or other suitable transfection reagent(s).

6. Opti-MEM (Thermo Fisher Scientific) reduced serum media.
7. Twenty-four well tissue culture-treated plates.
8. Transfection plasmids:

   pSPgRNA(s) with target sequence.

   pcDNA-dCas9-VP64 (Addgene#47107) or other suitable dCas9 transcriptional activator expression vector (*see* Tables 2 and 3).

   pMAX-GFP (Amaxa) or other suitable reporter plasmid for measuring transfection efficiency.

2.3 Analysis of mRNA Expression

1. 0.25% Trypsin-EDTA.
2. PBS.
3. QIAshredder (Qiagen).
4. RNeasy Plus RNA isolation kit (Qiagen).
5. qScript cDNA SuperMix (Quanta Biosciences).
6. RNase/DNase-free water.
7. PerfeCTa® SYBR® Green FastMix (Quanta Biosciences).
8. Oligonucleotides for qPCR (*see* **Notes 7** and **8**).

   ASCL FW: 5′-GGAGCTTCTCGACTTCACCA-3′.

   ASCL REV: 5′-AACGCCACTGACAAGAAAGC-3′.

   GAPDH FW: 5′-CAATGACCCCTTCATTGACC-3′.

   GAPDH REV: 5′-TTGATTTTGGAGGGATCTCG-3′.
9. CFX96 Real-Time PCR Detection System (Bio-Rad).

3 Methods

3.1 Design and Construction of sgRNA Expression Plasmids

The procedure we use for generating sgRNA vectors accomplishes plasmid digestion, oligonucleotide phosphorylation and ligation in a single reaction without DNA purification steps. This is a low cost and highly efficient procedure that can be completed by users with no molecular biology expertise in less than two hours from annealing to transformation.

1. Design and synthesize/order oligonucleotides (*see* Subheading 2.1) to target the regions of the promoter proximal to the TSS of the target transcript (*see* **Note 2**). Stocks of each oligonucleotide prepared at 100 μM in nuclease-free molecular biology grade water, can be stored frozen for extended periods (*see* **Note 7**).
2. Combine 1 μL of each sense and antisense oligonucleotide with 98 μL of TBS in a PCR tube. Anneal the oligonucleotide mix by incubation at 95 °C for 5 min, followed by 25 °C for 3 min.

3. Mix 1 μL of annealed and diluted oligonucleotides with 170 ng sgRNA vector, 2 μL 10× T4 ligase buffer, 1 μL of T4 ligase, 1 μL *BbsI/BpiI*, 1 μL T4 polynucleotide kinase (PNK), and MBG water to a final reaction volume of 20 μL. The sgRNA vector backbone is simultaneously digested and ligated with the annealed, phosphorylated oligonucleotides in a single reaction with the following thermocycling program:

   37 °C, 5 min.

   16 °C, 10 min.

   Repeat a and b for a total of three cycles.

4. Transform ligated plasmid by mixing 1.5 μL of the reaction product with 30 μL of competent *E. coli*, spread onto pre-warmed LB agar (*see* **Note 6**) containing 100 μg/mL carbenicillin, and incubate overnight at 37 °C.

5. We typically ensure correct ligation by analyzing four transformants per plate using colony PCR with KAPA2G Robust PCR Kits (*see* **Note 9**). We use 25 μL reactions containing MBG water (11.9 μL), 5× KAPA2G Buffer (5.0 μL), 5× Enhancer (5.0 μL), 10 mM dNTP mix (0.50 μL), 10 μM M13 Forward primer (1.25 μL), 10 μM Reverse primer (antisense cloning oligonucleotide) (1.25 μL), and 5 U/μL KAPA2G Robust (0.10 μL). With a pipette tip, scrape one colony from the plate, transfer to the PCR reaction and, immediately, to a second PCR tube containing LB broth. The PCR reactions are performed in a thermocycler according to manufacturer's instructions and the PCR products analyzed in 2% agarose gels containing 0.1–0.2 μg/mL ethidium bromide. The expected size of the correct PCR product is ~330 bp.

6. One colony, verified by PCR, is grown overnight in 5 mL of LB broth with 100 μg/mL carbenicillin.

7. The plasmid DNA from the bacterial culture is purified using a plasmid purification kit such as the Qiagen Spin Miniprep Kit (*see* **Note 9**) and the construct is verified by DNA sequencing with M13 Forward primer.

3.2 Activation of Target Gene Expression in Mammalian Cells

1. A typical experimental setup includes reactions containing plasmid mixtures such as the following (*see* **Note 10**):

   GFP (1 μg) (*see* **Note 11**).

   sgRNA 1 and dCas9 (0.5 μg each).

   sgRNA 2 and dCas9 (0.5 μg each).

   sgRNA 3 and dCas9 (0.5 μg each).

   sgRNA 4 and dCas9 (0.5 μg each).

   sgRNA 1+ sgRNA 2 + sgRNA 3 + sgRNA 4 (0.125 μg of each) and dCas9 (0.5 μg).

2. For optimal transfection efficiency, low passage 293T cells in logarithmic growth are trypsinized, harvested, and resuspended at 10^6 cells/mL in DMEM.
3. As per manufacturer's instructions, the DNA is mixed with 50 μL of Opti-MEM in a microfuge tube and, in a separate tube, 2 μL of Lipofectamine 2000 are mixed with 50 μL of Opti-MEM. After 5 min, the contents of both tubes are combined and incubated for an additional 20 min. The 100 μL DNA-lipofectamine reagent mixture is pipetted into one well of a 24-well treated tissue culture dish and promptly mixed with 400 μL of freshly harvested and properly diluted cells (*see* **Note 12**).
4. Incubate the cells for 48–72 h before analyzing gene expression.

3.3 Analysis of Gene Expression by qPCR

1. The cells are trypsinized and washed with PBS once (*see* **Note 13**).
2. Total RNA is isolated using the RNeasy Plus RNA isolation kit (Qiagen) (*see* **Note 14**). The cells are lysed by adding an appropriate volume of RLT Plus with 10 μL/mL of β-mercaptoethanol and homogenized with QIAshredder columns. All other steps are performed according to manufacturer's instructions (*see* **Note 15**).
3. cDNA synthesis is performed using the qScript cDNA SuperMix (Quanta Biosciences) (*see* **Note 16**) by incubation of 1 μg of RNA with 4 μL of qScript cDNA SuperMix and RNase/DNase-free water up to 20 μL. The thermocycling parameters are:
   (a) 5 min at 25 °C.
   (b) 30 min at 42 °C.
   (c) 5 min at 85 °C.
4. Real-time PCR is performed using PerfeCTa® SYBR® Green FastMix (Quanta Biosciences) with the CFX96 Real-Time PCR Detection System (Bio-Rad). The primers are designed using Primer3Plus (*see* **Notes 7** and **8**), purchased from IDT and validated by agarose gel electrophoresis and melting curve analysis. For each sample, quantification of a housekeeping gene (such as GAPDH) must be performed in addition to analysis of the target gene. The qPCR reactions contain 10 μL PerfeCTa® SYBR® Green FastMix (2×), 2 μL forward primer (5 μM), 2 μL reverse primer (5 μM), cDNA and RNase/DNase-free water up to 20 μL. The optimal cycling parameters for each gene must be determined experimentally to ensure efficient amplification over an appropriate dynamic range (*see* **Note 17**).
5. Calculate fold-increase mRNA expression of the gene of interest normalized to GAPDH expression using the ddCt method [24].

4 Notes

1. Dual expression of Cas9 and sgRNA from a single plasmid is an attractive alternative to a two plasmid system. However, since plasmid size correlates negatively with transfection efficiency, further optimization of gene delivery protocols may be needed.
2. Benchling provides on-target [25] and off-target [26] scores associated with each target site. Off-target changes in gene expression are uncommon when using multiple sgRNAs to activate gene expression, since all target sites must be found simultaneously near the TSS of the off-target gene. However, since second-generation systems for gene activation require one single sgRNA, it is important to identify high quality sgRNAs with favorable off-target scores. For each sgRNA, Benchling provides a detailed list of potential off-target sites that can be used for biased detection of off-target gene activation.
3. The number of nucleotides in the sgRNA complementary with the target site can range between 17 and 20 bp. In fact, it has been demonstrated that sgRNAs with 17 or 18 complementary nucleotides efficiently guide *S. pyogenes* Cas9 to the target site where it introduces double strand breaks with improved specificity [27].
4. There are multiple commercial sources for *BbsI/BpiI*. Some formulations of *BbsI/BpiI* require storage at −80 °C and, repeated cycles of freeze-thaw that occur when used frequently, result in decreased enzymatic activity and undesired background during cloning. We prefer formulations of *BbsI/BpiI* that can be stored at −20 °C.
5. T4 DNA ligase buffer typically contains 10 mM dithiothreitol, which is not stable through repeated freeze-thaw cycles. We typically prepare single use aliquots of T4 buffer.
6. Any chemically competent cells or electro-competent cells can be used. We prefer HIT Competent Cells-DH5α. These chemically competent cells can be transformed very efficiently without heat-shock by mixing 1.5 μL of the ligation reaction with 30 μL of competent cells followed by incubation at 4 °C for 1–10 min and plating. When using this short protocol, it is essential to use plates prewarmed at 37 °C, otherwise the transformation efficiency decreases significantly. If the transformation efficiency is too low, addition of 100 μL of SOC broth and incubation at 37 °C with shaking for 10 min should yield hundreds to thousands of colonies.
7. It is essential to use high quality primers for reproducible qPCR results. Repeated freeze-thaw cycles can significantly alter primer binding to the template. Upon receipt, we resuspend the primers in MBG water and prepare single use aliquots

that are stored at −80 °C. Multiple oligonucleotides are often designed and tested for finding a suitable primer combination that is specific and amplifies the target transcript with 90–110% efficiency. Many excellent design tools, such as Primer3Plus, are freely available as stand-alone or web-based applications.

8. We prefer to perform qPCR using fast cycling two-step protocols with amplicons between 100 and 150 bp long. One important consideration for primer design is to use primers that bind different exons separated, if possible, by several kilobases. This will ensure that any residual genomic DNA that might be present in the RNA sample will not be amplified during the PCR reaction.
9. Since this cloning system is both reliable and reproducible, the user may choose to skip colony PCR and proceed directly to construct verification by sequencing.
10. Plasmid DNA purified using Qiagen Spin Miniprep Kit is suitable for transfection of a variety of cell lines, however, the resulting plasmid prep contains significant levels of endotoxins from *E. coli* that can result in decreased viability in some cell types. DNA precipitation with ethanol is usually sufficient to obtain transfection grade DNA suitable for use in most cell types.
11. A control transfection reaction containing a GFP or similar expression plasmid should be used to ensure adequate transfection efficiency is achieved under identical experimental conditions and to serve as a negative control for qPCR.
12. It is typically recommended to perform transfections in antibiotic-free medium. We have not observed decreased transfection efficiency or viability by using antibiotics in 293T cells. However, other cell lines may be affected differently by the presence of antibiotic. Similarly, while transfection in suspension works well in 293T cells, the protocol must be optimized by the user for other cell types since we have observed increased toxicity when the cells are plated at densities below those recommended.
13. We typically analyze gene expression in three independent experiments that are performed on three different days using biological duplicates in each experiment. Since RNA is unstable and degrades rapidly over time, we prefer to harvest the cells and freeze cell pellets until all three experiments have been completed. At that point we perform RNA extraction from all samples simultaneously to minimize variability due to sample handling.
14. When using the RNeasy Plus RNA isolation kit, DNA digestion is performed during a centrifugation step. We have compared this DNA removal system with standard enzymatic removal of genomic DNA after RNA isolation and we have determined that the efficiency of both systems is comparable.

15. It is recommended to prepare 70% ethanol and RPE buffer fresh before use since ethanol evaporation over time can decrease the efficiency of the different wash steps.
16. For the cDNA synthesis reaction to occur identically in all samples, it is important to use equal amounts of RNA from all samples. We typically prepare cDNA from 1 μg of RNA.
17. We prefer to generate standard curves using tenfold dilutions with cDNA obtained from the sample presumed to have the highest transcript concentration. The use of plasmid DNA or other synthetic templates can lead to errors in determining the linear range of the PCR.

References

1. Uil TG, Haisma HJ, Rots MG (2003) Therapeutic modulation of endogenous gene function by agents with designed DNA-sequence specificities. Nucleic Acids Res 31(21):6064–6078
2. Knauert MP, Glazer PM (2001) Triplex forming oligonucleotides: sequence-specific tools for gene targeting. Hum Mol Genet 10(20):2243–2251
3. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. Curr Opin Struct Biol 13(3):284–299
4. Eguchi A, Lee GO, Wan F et al (2014) Controlling gene networks and cell fate with precision-targeted DNA-binding proteins and small-molecule-based genome readers. Biochem J 462(3):397–413
5. Polstein LR, Perez-Pinera P, Kocak DD et al (2015) Genome-wide specificity of DNA binding, gene regulation, and chromatin remodeling by TALE- and CRISPR/Cas9-based transcriptional activators. Genome Res 25(8):1158–1169
6. Konermann S, Brigham MD, Trevino AE et al (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517(7536):583–588
7. Perez-Pinera P, Kocak DD, Vockley CM et al (2013) RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods 10(10):973–976
8. Perez-Pinera P, Ousterout DG, Brunger JM et al (2013) Synergistic and tunable human gene activation by combinations of synthetic transcription factors. Nat Methods 10(3):239–242
9. Maeder ML, Linder SJ, Reyon D et al (2013) Robust, synergistic regulation of human gene expression using TALE activators. Nat Methods 10(3):243–245
10. Gilbert LA, Horlbeck MA, Adamson B et al (2014) Genome-scale CRISPR-mediated control of gene repression and activation. Cell 159(3):647–661
11. Hilton IB, D'Ippolito AM, Vockley CM et al (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol 33(5):510–517
12. Chavez A, Scheiman J, Vora S et al (2015) Highly efficient Cas9-mediated transcriptional programming. Nat Methods 12(4):326–328
13. Gersbach CA, Perez-Pinera P (2014) Activating human genes with zinc finger proteins, transcription activator-like effectors and CRISPR/Cas9 for gene therapy and regenerative medicine. Expert Opin Ther Targets 18(8):835–839
14. Gilbert LA, Larson MH, Morsut L et al (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154(2):442–451
15. Maeder ML, Linder SJ, Cascio VM et al (2013) CRISPR RNA-guided activation of endogenous human genes. Nat Methods 10(10):977–979
16. Cheng AW, Wang H, Yang H et al (2013) Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res 23(10):1163–1171
17. Mali P, Aach J, Stranges PB et al (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol 31(9):833–838
18. Farzadfard F, Perli SD, Lu TK (2013) Tunable and multifunctional eukaryotic transcription factors based on CRISPR/Cas. ACS Synth Biol 2(10):604–613
19. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346(6213):1258096

20. Cong L, Ran FA, Cox D et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121):819–823
21. Mali P, Yang L, Esvelt KM et al (2013) RNA-guided human genome engineering via Cas9. Science 339(6121):823–826
22. Jinek M, Chylinski K, Fonfara I et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337(6096):816–821. doi:10.1126/science
23. Benchling (2015) Biology Software
24. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29(9):e45
25. Doench JG, Hartenian E, Graham DB et al (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol 32(12):1262–1267
26. Hsu PD, Scott DA, Weinstein JA et al (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 31(9):827–832
27. Fu Y, Sander JD, Reyon D et al (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol 32(3):279–284
28. Seipel K, Georgiev O, Schaffner W (1992) Different activation domains stimulate transcription from remote ("enhancer") and proximal ("promoter") positions. EMBO J 11(13):4961–4968
29. Sadowski I, Ma J, Triezenberg S, Ptashne M (1988) GAL4-VP16 is an unusually potent transcriptional activator. Nature 335(6190):563–564
30. Beerli RR, Segal DJ, Dreier B, Barbas CF 3rd (1998) Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. Proc Natl Acad Sci U S A 95(25):14628–14633
31. Konermann S, Brigham MD, Trevino AE et al (2013) Optical control of mammalian endogenous transcription and epigenetic states. Nature 500(7463):472–476
32. Polstein LR, Gersbach CA (2015) A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. Nat Chem Biol 11(3):198–200
33. Polstein LR, Gersbach CA (2012) Light-inducible spatiotemporal control of gene activation by customizable zinc finger transcription factors. J Am Chem Soc 134(40):16480–16483
34. Tanenbaum ME, Gilbert LA, Qi LS et al (2014) A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell 159(3):635–646
35. Nissim L, Perli SD, Fridkin A et al (2014) Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. Mol Cell 54(4):698–710
36. Shechner DM, Hacisuleyman E, Younger ST, Rinn JL (2015) Multiplexable, locus-specific targeting of long RNAs with CRISPR-display. Nat Methods 12(7):664–670
37. Kearns NA, Genga RM, Enuameh MS et al (2014) Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. Development 141(1):219–223
38. Heckl D, Kowalczyk MS, Yudovich D et al (2014) Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. Nat Biotechnol 32(9):941–946
39. Kabadi AM, Ousterout DG, Hilton IB, Gersbach CA (2014) Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Res 42(19):e147
40. Li HL, Fujimoto N, Sasakawa N et al (2015) Precise correction of the dystrophin gene in duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPR-Cas9. Stem Cell Rep 4(1):143–154
41. Shen B, Zhang W, Zhang J et al (2014) Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. Nat Methods 11(4):399–402
42. Zheng Q, Cai X, Tan MH et al (2014) Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. Biotechniques 57(3):115–124
43. Tsai SQ, Wyvekens N, Khayter C et al (2014) Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat Biotechnol 32(6):569–576
44. Kleinstiver BP, Prew MS, Tsai SQ et al (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature 523(7561):481–485
45. Ding Q, Regan SN, Xia Y et al (2013) Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. Cell Stem Cell 12(4):393–394
46. Qi LS, Larson MH, Gilbert LA et al (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152(5):1173–1183

47. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science 343(6166):80–84
48. Zhou Y, Zhu S, Cai C et al (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature 509(7501):487–491
49. Koike-Yusa H, Li Y, Tan EP et al (2014) Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol 32(3):267–273
50. Ranganathan V, Wahlin K, Maruotti J, Zack DJ (2014) Expansion of the CRISPR-Cas9 genome targeting space through the use of H1 promoter-expressed guide RNAs. Nat Commun 5:4516
51. Sanjana NE, Shalem O, Zhang F (2014) Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods 11(8):783–784
52. Platt RJ, Chen S, Zhou Y et al (2014) CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell 159(2):440–455
53. Swiech L, Heidenreich M, Banerjee A et al (2015) In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. Nat Biotechnol 33(1):102–106

Index

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6, © Springer Science+Business Media New York 2017

N

O

P

Q

R

S

T

U