Methods in Molecular Biology 1150

Springer Protocols

Benjamin L. Kidder *Editor*

Stem Cell Transcriptional Networks

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

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

Stem Cell Transcriptional Networks

Methods and Protocols

Edited by

Benjamin L. Kidder

Systems Biology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA

Humana Press

Editor
Benjamin L. Kidder
Systems Biology Center
National Heart, Lung, and Blood Institute
National Institutes of Health
Bethesda, MD, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-0511-9 ISBN 978-1-4939-0512-6 (eBook)
DOI 10.1007/978-1-4939-0512-6
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014936042

© Springer Science+Business Media New York 2014
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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
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.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Research in stem cell biology has generated immense interest recently due to the potential of stem cells to treat human diseases such as neurodegenerative disease, cardiovascular disease, and cancer. Advancements in stem cell research aid our understanding of genetics and developmental biology, and, because stem cells have the ability to repopulate tissues, there is an expectation that stem cells research will translate into clinical therapies. However, much work is required to understand the underlying transcriptional programs of stem cells that promote self-renewal vs. differentiation. Investigating how epigenetic and transcriptional landscapes are patterned in stem cells and committed lineages will provide insight into how these features regulate unique cellular expression programs during development, and contribute to the diverse cellular repertoire that exists in mammals. Next-generation sequencing technologies have recently been used to survey global expression and protein–DNA binding interactions at nucleotide resolution in stem cells. Use of these emerging technologies has shed light on stem cell transcriptional networks that define primitive vs. committed epigenetic landscapes. Moreover, recent advancements in reprogramming and transdifferentiation have armed stem cell biologists with additional tools to develop solutions for regenerative medicine purposes. The aim of this volume is to provide a resource for biologists to interrogate stem cell transcriptional networks.

Bethesda, MD, USA *Benjamin L. Kidder*

Contents

Contributors

Frederick Anokye-Danso • *Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

Daniel Blankenberg • *Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA, USA*

Claudia Cavelti-Weder • *Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center Boston, Boston, MA, USA*

James B. Cooper • *Department of Molecular, Cellular, and Developmental Biology, University of California at Santa Barbara, Santa Barbara, CA, USA*

Victor J. Dzau • *Division of Cardiology, Department of Medicine, Mandel Center for Hypertension and Atherosclerosis Research, and the Cardiovascular Research Center, Duke University Medical Center, Durham, NC, USA*

Kai Ge • *Laboratory of Endocrinology and Receptor Biology, Adipocyte Biology and Gene Regulation Section, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA*

Boris Greber • *Max Planck Institute for Molecular Biomedicine, Münster, Germany; Chemical Genomics Centre of the Max Planck Society, Dortmund, Germany*

Sean Grullon • *Department of Physics, The George Washington University, Washington, DC, USA; Laboratory of Endocrinology and Receptor Biology, Adipocyte Biology and Gene Regulation Section, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA*

Jennifer Hillman-Jackson • *Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA, USA*

Gangqing Hu • *Systems Biology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA*

Guang Hu • *Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA*

Tilanthi Jayawardena • *Division of Cardiology, Department of Medicine, Mandel Center for Hypertension and Atherosclerosis Research, and the Cardiovascular Research Center, Duke University Medical Center, Durham, NC, USA*

Benjamin L. Kidder • *Systems Biology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA*

Arnaud R. Krebs • *Friedrich Miescher Institut for Biomedical Research, Basel, Switzerland*

Weida Li • *Department of Stem Cell and Regenerative Biology, Harvard Stem Cell Institute, Harvard University, Cambridge, MA, USA*

Arthur Liberzon • *Broad Institute of MIT and Harvard, Cambridge, MA, USA*

Tao Liu • *Department of Biochemistry, University at Buffalo-COEBLS, Buffalo, NY, USA*

Ben D. MacArthur • *School of Mathematics and Institute for Life Sciences, University of Southampton, Southampton, UK*

Samuele Marro • *Department of Pathology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA*

Maria Mirotsou • *Division of Cardiology, Department of Medicine, Mandel Center for Hypertension and Atherosclerosis Research, and the Cardiovascular Research Center, Duke University Medical Center, Durham, NC, USA*
Aaron M. Newman • *Institute for Stem Cell Biology and Regenerative Medicine, School of Medicine, Stanford University, Stanford, CA, USA*
Olga Nikolayeva • *Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland*
Weiqun Peng • *Department of Physics and Department of Anatomy and Regenerative Biology, The George Washington University, Washington, DC, USA*
Jyoti Rao • *Max Planck Institute for Molecular Biomedicine, Münster, Germany*
Sarina Ravens • *Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CU de Strasbourg, Strasbourg, France*
Mark D. Robinson • *Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland*
Patrick S. Stumpf • *Centre for Human Development, Stem Cells, and Regeneration, Institute of Developmental Sciences, University of Southampton, Southampton, UK*
Làszlò Tora • *Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CU de Strasbourg, Strasbourg, France; School of Biological Sciences, Nanyang Technological University, Singapore, Singapore*
Gordon C. Weir • *Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center Boston, Boston, MA, USA*
Shiliyang Xu • *Department of Physics, The George Washington University, Washington, DC, USA; Laboratory of Endocrinology and Receptor Biology, Adipocyte Biology and Gene Regulation Section, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA*
Nan Yang • *Department of Pathology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA*
Tao Ye • *Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CU de Strasbourg, Strasbourg, France*
Keji Zhao • *Systems Biology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA*
Xiaofeng Zheng • *Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA*
Qiao Zhou • *Department of Stem Cell and Regenerative Biology, Harvard Stem Cell Institute, Harvard University, Cambridge, MA, USA*

Part I

Next-Generation Sequencing Library Preparation and Data Analysis

Chapter 1

Efficient Library Preparation for Next-Generation Sequencing Analysis of Genome-Wide Epigenetic and Transcriptional Landscapes in Embryonic Stem Cells

Benjamin L. Kidder and Keji Zhao

Abstract

Gene expression in embryonic stem (ES) cells is regulated in part by a network of transcription factors, epigenetic regulators, and histone modifications that influence the underlying chromatin in a way that is conducive or repressive for transcription. Advances in next-generation sequencing technology have allowed for the genome-wide analysis of chromatin constituents and protein–DNA interactions at high resolution in ES cells and other stem cells. While many studies have surveyed genome-wide profiles of a few factors and expression changes at a fixed time point in undifferentiated ES cells, few have utilized an integrative approach to simultaneously survey protein–DNA interactions, histone modifications, and expression programs during ES cell self-renewal and differentiation. To identify transcriptional networks that regulate pluripotency and differentiation, it is important to generate high-quality genome-wide maps of transcription factors, chromatin factors, and histone modifications and to survey global gene expression profiles. Here, to interrogate genome-wide profiles of chromatin features and to survey global gene expression programs in ES cells, we describe protocols for efficient library construction for next-generation sequencing of ChIP-Seq and RNA-Seq samples.

Key words Next-generation sequencing, Library construction, Chromatin immunoprecipitation, ChIP-Seq, RNA-Seq, Gene expression, Transcriptome, Embryonic stem cells

1 Introduction

ES cell self-renewal and differentiation are controlled in part by interactions between external signaling pathways, *cis*-acting DNA regulatory elements, and *trans*-acting chromatin binding factors (e.g., transcription factors, epigenetic regulators). Together, these interactions influence the underlying chromatin state to one that is conducive or repressive for transcription. ES cells express the core pluripotency regulators Oct4, Sox2, Nanog, and Tbx3 which are thought to perpetuate self-renewal by promoting expression of a network of ES cell-enriched genes and inhibit expression of development genes [1, 2]. Ablation of these networks leads to the

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_1, © Springer Science+Business Media New York 2014

collapse of self-renewal followed by differentiation [3]. Application of next-generation sequencing technologies has provided unprecedented genome-wide views of transcriptional regulatory networks and gene expression in ES cells and during development. However, while many studies have generated genome-wide maps of transcription factors and histone modifications in ES cells at a fixed time point [1, 2, 4–8], few have addressed the roles that epigenetic regulators play in controlling ES cell function [9–11], and few have utilized an integrative approach to simultaneously survey protein–DNA interactions, histone modifications, and expression programs during ES cell self-renewal and differentiation. To fully understand processes that govern ES cell self-renewal versus differentiation, it will be important to employ a systems-based approach to the study of ES cell biology. Perturbation of transcription factors and epigenetic regulators followed by a combinatorial analysis of genome-wide occupancy of chromatin factors and histone modifications and global expression profiles will be useful to identify transcriptional networks that define cellular states. Integration of computational analyses with phenotypic data from knockout or RNAi knockdown ES cells will be essential to generate a phenome–epigenome network. However, to meet these goals, it will be important to utilize high-quality protocols for chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) and transcriptome analysis (RNA-Seq). Here, we outline detailed protocols for ChIP-Seq and RNA-Seq analysis, including protocols for efficient generation of libraries for sequencing on next-generation sequencing platforms.

2 Materials

2.1 Cells and Culture Reagents

1. Cells.
   - Mitotically inactivated mouse embryonic fibroblasts (iMEFs) derived from E13.5 to E14.5 mouse embryos. Store in liquid nitrogen until use (*see* **Note 1**).
   - Mouse embryonic stem cells (ES cells) with a normal karyotype and at low passage. Store in liquid nitrogen until their use (*see* **Note 2**).
2. Growth factor.
   - Mouse leukemia inhibitory factor (LIF, e.g., Millipore, 10^6 or 10^7 units). Store at 4 °C.
3. Small-molecule inhibitors.
   - GSK3β inhibitor CHIR99021 (GSK3i, *see* **Note 3**). Resuspend CHIR99021 in DMSO to a final concentration of 10 mM. Aliquot and store at −20 °C. Store thawed aliquots at 4 °C for several days.

- MEK inhibitor PD0325901 (MEKi, *see* **Note 3**). Resuspend PD0325901 in DMSO to a final concentration of 10 mM. Aliquot and store at −20 °C. Store thawed aliquots at 4 °C for several days.

4. Cell culture.
   - Mouse ES cell culture media: DMEM high glucose, 15 % ES cell-qualified FBS, LIF (10 ng/mL), 1× penicillin streptomycin (pen strep), 1× glutamine, 1× 2-mercaptoethanol (cell culture grade), nonessential amino acids (NEAA) at 37 °C with 5 % CO2.
   - MEF culture media: DMEM high glucose, 10 % FBS, 1× pen strep, 1× glutamine at 37 °C with 5 % CO2.
   - 0.1 % gelatin solution in water.
   - Lab Armor bead bath.
5. Passage of cells.
   - PBS without calcium and magnesium stored at room temperature.
   - 0.25 % Trypsin–EDTA (1×, phenol red) warmed to 37 °C.

2.2 Reagents for Preparation of Chromatin

1. Cross-linking.
   - 37 % formaldehyde.
   - Cell culture media warmed to 37 °C.
   - PBS without calcium and magnesium stored at 4 °C.
   - Glycine (1.25 M, 10×) stored at 4 °C.
   - Liquid nitrogen or dry ice to freeze cell pellet.
2. Sonication.
   - Ice-cold TE (pH 8.0).
   - Fresh PMSF protease inhibitor (1 mM final).
   - Protease inhibitor cocktail (1× final).
   - SDS when preparing chromatin for analysis of histone modifications (0.1 % final).
   - 15 mL conical tubes.
   - Sonifier.
   - 10 % Triton X-100.
   - 10 % sodium deoxycholate.
   - 10 % SDS (*see* **Note 4**).
   - 5 M NaCl (*see* **Note 5**).

2.3 Chromatin Immunoprecipitation (ChIP) Buffers

1. Magnetic separation reagents.
   - Protein A or Protein G Dynabeads® (40 μL).
   - Magnet for molecular separation application (fits 1.5–2 mL tubes).
   - PBS without calcium and magnesium stored at room temperature.
2. RIPA wash buffer.
   - 0.1 % SDS.
   - 0.1 % sodium deoxycholate.
   - 1 % Triton X-100.
   - TE (10 mM Tris–HCl; 1 mM EDTA) pH 8.0.
3. LiCl wash buffer.
   - 0.25 M LiCL.
   - 0.5 % NP40.
   - 0.5 % sodium deoxycholate.
   - H_2O.
4. RIPA buffer + 0.3 M NaCl.
   - 0.3 M NaCl.
   - 47 mL RIPA buffer (for 50 mL).
5. TE + 0.2 % Triton X-100.
   - 0.2 % Triton X-100.
   - 49.9 mL TE pH 8.0 (for 50 mL).
6. TE + 0.5 M NaCl.
   - 0.5 M NaCl.
   - 45 mL TE pH 8.0 (for 50 mL).

2.4 Reagents for Isolation of Total RNA

1. Qiagen miRNeasy or RNeasy Kit.

2.5 Reagents for the Purification of mRNA from Total RNA

1. Dynabeads mRNA DIRECT™ Kit (*see* **Note 6**).
2. Dynal oligo(dT) beads (Invitrogen).
3. Binding buffer (20 mM Tris–HCl, pH 7.5, 1.0 M LiCl, 2 mM EDTA).
4. Wash buffer B (10 mM Tris–HCl, pH 7.5, 0.15 M LiCl, 1 mM EDTA).
5. 10 mM Tris–HCl, pH 7.5.
6. Thermomixers or heat blocks.
7. Magnet for molecular separation application (fits 1.5–2 mL tubes).
8. Rotator for 1.5–2 mL tubes.

2.6 Reagents for Generating Double-Stranded cDNA from RNA

1. First-Strand cDNA Synthesis Reagents.
   - SuperScript® Double-Stranded cDNA Synthesis Kit (*see* **Note 7**).
   - Random hexamer primers (3 μg/μL, Invitrogen; *see* **Note 8**).
   - 5× first-strand buffer.
   - 100 mM DTT.
   - 10 mM dNTP mix.
   - RNaseOUT (40 U/μL, Invitrogen).
   - Superscript II enzyme (200 U/μL, Invitrogen; *see* **Note 9**).
2. Second-Strand Synthesis Reagents.
   - SuperScript Double-Stranded cDNA Synthesis Kit.
   - DEPC-treated water.
   - 5× second-strand reaction buffer.
   - 10 mM dNTP mix.
   - *E. coli* DNA ligase (10 U/μL).
   - *E. coli* DNA polymerase I (10 U/μL).
   - *E. coli* RNase H (2 U/μL).
   - T4 DNA polymerase.
   - QIAquick PCR Purification Kit.

2.7 Reagents for Library Preparation

DNA end-repair reagents

1. 1–34 μL DNA.
2. 5 μL 10× end-repair buffer (Epicentre End-It DNA End-Repair Kit; 330 mM Tris–acetate, pH 7.8, 660 mM potassium acetate, 100 mM magnesium acetate, mM Dithiothreitol (DTT)).
3. 5 μL 2.5 mM each dNTP.
4. 5 μL 10 mM dATP.
5. 1 μL End-Repair enzyme mix (T4 DNA polymerase, T4 polynucleotide kinase).
6. Qiagen MinElute Reaction Cleanup Kit.

Addition of "A" overhang to 3′ end reagents

1. 30 μL of end-repaired DNA.
2. 5 μL 10× NEB buffer #2.
3. 1 μL dATP (10 mM stock).
4. 3 μL 5 U/μL Klenow fragment (3′ → 5′ exo-).
5. Qiagen MinElute Reaction Cleanup Kit.

Linker ligation reagents

1. 23 μL end-repaired DNA with added "A" overhang.
2. 3 μL 10× T4 DNA ligase buffer.

3. 1 μL Index adapter oligo mix (1:10 diluted, *see* **Note 10**).
4. 3 μL T4 DNA ligase (400 U/μL).

Gel purification reagents for size selection

1. 1 kb Plus DNA Ladder.
2. 2 % agarose gel (E-Gel® EX or freshly made) (*see* **Note 11**).
3. Qiagen MinElute Gel Extraction Kit.
4. Razor blades.

PCR amplification reagents

1. 24 μL of gel-purified DNA.
2. 25 μL of PCR master mix (Phusion High-Fidelity PCR Master Mix, *see* **Note 12**).
3. 1 μL PCR primer cocktail (Illumina PE primers 1.0 and 2.0 or custom primers, *see* **Note 13**).
4. 2 % agarose gel (E-Gel® EX or freshly made).
5. PCR 8-well strip tubes.
6. Thermocycler.

3 Methods

3.1 Culture of ES Cells in Feeder-Free Conditions

Preparation of feeder layer

1. Add 2 mL of gelatin solution to 6-well plate and incubate at 37 °C for 10–15 min.
2. Pre-warm MEF media (10 % FBS) at 37 °C (*see* **Note 14**) in a Lab Armor bead bath or water bath.
3. Thaw vial of frozen iMEFs in a Lab Armor bead bath, add cells to several mLs of pre-warmed MEF media in 15 mL conical tube, and centrifuge at 1,500 rpm (500×*g*) for 3 min.
4. Aspirate gelatin from 6-well dish, resuspend iMEFs in MEF medium, and plate at 250 k–300 k cells per well. Before placing the dish in the incubator, move the plate vertically and then horizontally to evenly distribute MEFs on culture dish.

Culture of ES cells on iMEFs

1. The next day, pre-warm MEF media and ES cell media at 37 °C.
2. Thaw a vial of frozen ES cells in a 37 °C Lab Armor bead bath, add cells to several milliliters of pre-warmed MEF media in 15 mL conical tube, and centrifuge at 1,500 rpm (500×*g*) for 3 min.
3. Aspirate MEF media from 6-well culture dish, resuspend ES cells in ES cell media, and plate ES cells on the layer of iMEFs. Move the plate vertically and horizontally to evenly distribute the cells, and then place in the incubator.

4. Once the colonies reach semi-confluence (cover 30–50 % of the dish, *see* **Note 15**), the cells should be passaged onto a gelatin-coated dish without feeders. Prepare a gelatin-coated dish as described above. Meanwhile, pre-warm 10 % MEF media, ES cell media, and 0.25 % trypsin solution, and prepare to passage the ES cells as described below.
5. Wash 6-well plate with 1 mL of 1× PBS, add 1 mL of pre-warmed 0.25 % trypsin solution, and let stand for 1–2 min or until colonies begin to disassociate.
6. Using a 1 mL pipette, pipette up and down several times to thoroughly dissociate the ES cells. Be careful not to over pipette as this may result in excessive cell death. Check the status of the dissociation under a microscope.
7. Quench the trypsin reaction by adding the cells to 3–4 mL of MEF media in a 15 mL conical tube, and centrifuge at 1,500 rpm ($500 \times g$) for 3 min.

Culture of ES cells without feeders

1. Remove the media and resuspend the cell pellet in ES cell media containing GSK3i (2–3 μM CHIR99021) or GSK3i and MEKi (2–3 μM GSK3i, CHIR99021; 1 μM MEKi, PD0325901), remove the gelatin solution from the 6-well dish prepared above, plate ES cells in the culture dish, and incubate at 37 °C (Fig. 1a). Dual inhibition of GSK3i and MEKi has been shown to promote ES cell self-renewal in feeder-free and serum-free conditions [12].
2. Passage the ES cells at least twice without iMEFs to remove all traces of MEFs (Fig. 1b, c). Before harvesting cells for ChIP-Seq, expand ES cells onto 10 cm dishes. For RNA-Seq studies, harvesting ES cells from a 6-well dish should suffice.

Fixation of ES cells

1. Harvest 10–100^6 ES cells. First, wash 10 cm culture dishes containing ES cells with PBS, add 2 mL trypsin, and incubate at room temperature for 1–3 min. Use a microscope to visualize progress of dissociation. Pipette up and down to dissociate cells, and add cells to several mLs of pre-warmed MEF media in conical tube. Centrifuge for 3 min at 1,500 rpm ($500 \times g$).
2. Resuspend ES cell pellet in pre-warmed MEF media at a density of 2×10^6 cells per mL.
3. Fix ES cells by adding 37 % formaldehyde to a final concentration of 1 %. Incubate at room temperature on a rocker for 8 min.
4. Quench fixation reaction by adding 1.25 M ice-cold glycine to a final concentration of 0.125 M. Incubate at room temperature on a rocker or rotator for 5 min, and then centrifuge at 1,500 rpm ($500 \times g$) for 5 min and remove the supernatant.

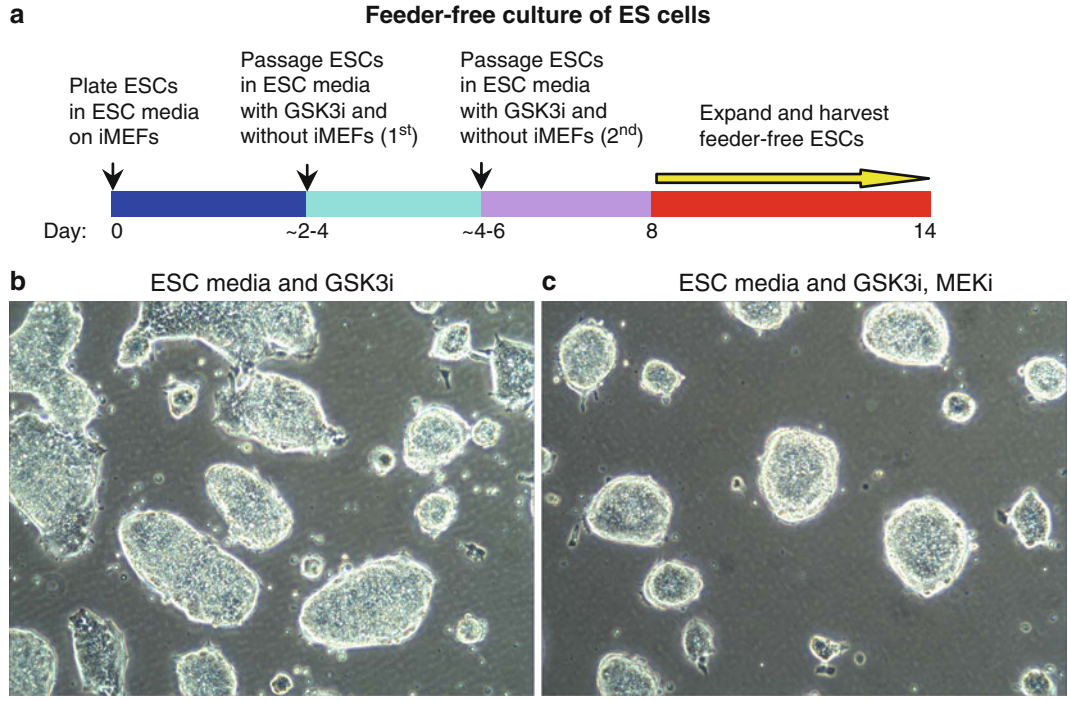


Fig. 1 Feeder-free culture of ES cells. (**a**) Experimental design. ES cells cultured without iMEFs for at least two passages on gelatin-coated dishes in ES cell media containing LIF and small-molecule inhibitors (**b**) GSK3i or (**c**) GSK3i and MEKi

5. Wash cell pellet with ice-cold PBS, centrifuge, and remove PBS. Repeat **step 5** once more. After resuspending in PBS for the second time, aliquot cells into 15 mL conical tubes with 15×10^6 cells per tube (*see* **Note 16**).
6. Flash-freeze cell pellet in liquid nitrogen or dry ice, and store at −80 °C.

3.2 Sonication of Chromatin

1. Thaw frozen pellet of fixed ES cells from above on ice.
2. Resuspend pellet in 2.5 mL of sonication buffer (*see* **Note 17**, for histones, 1× TE, 1 mM PMSF, 1× proteinase inhibitor cocktail, 0.1 % SDS; for protein factors, 1× TE, 1 mM PMSF, 1× proteinase inhibitor cocktail).
3. Shear chromatin by sonicating with 16–18 cycles of 30 s pulses at maximum setting with 30 s intervals on ice (*see* **Note 18**) (Fig. 2).
4. Adjust to RIPA buffer by adding 0.28 mL of 10 % Triton X-100 and 28 μL sodium deoxycholate (and 28 μL of 10 % SDS for protein factors). Mix well to solubilize chromatin and aliquot equal volumes to 1.5 mL tubes.
5. Pre-clear samples by centrifuging at 13,000 rpm (14,000 × *g*) (max speed on tabletop) for 10 min at 4 °C. Transfer supernatant to new tube and store at −80 °C until use.

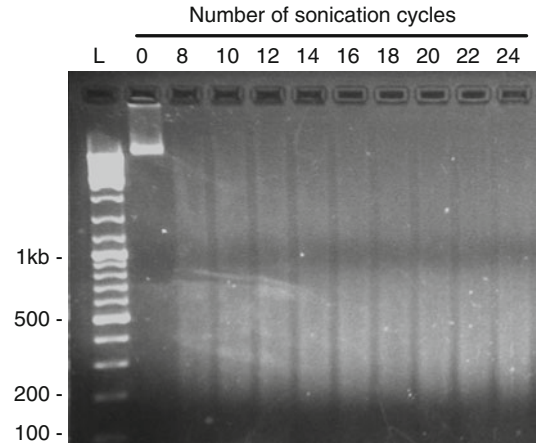


Fig. 2 Relationship between the number of cycles of sonication and fragment size. 12×10^6 cells cross-linked ES cells were sonicated in SDS-containing buffer using a Misonix 3000 sonifier with a microtip according to the following conditions: power setting 5, 8–24 cycles (30 s sonication pulse, 30 s rest). Samples were removed every two cycles and cross-linking was reversed. DNA was purified and run on a 2 % agarose gel. The lane with a molecular weight marker is labeled L and the lanes with sonicated samples are labeled with the number of sonication cycles. The size of the molecular weight ladder is indicated. Note that 8 cycles resulted in under-sonication, while 14 cycles resulted in appropriate sonication. Sonication beyond 16 cycles did not provide much improvement in fragment size

3.3 Chromatin Immunoprecipitation (ChIP)

1. Add 40 μL of Dynabeads Protein A or G to a 1.5 mL tube. Wash beads with 600 μL PBS. Insert 1.5 mL tube into magnetic rack and let stand for 1 min. Aspirate PBS and resuspend beads in 100 μL PBS.
2. Add 4–8 μg antibody to each tube containing beads. Tap gently to mix, and rotate at room temperature for 40 min to allow attachment of antibody to beads.
3. Attach 1.5 mL tube to magnet, let stand for 1 min, and remove PBS.
4. Wash the beads twice with 200 μL PBS to remove free IgGs. Rotate at room temperature for 5 min each time.
5. Use the magnet to remove the supernatant.
6. Add chromatin extract (equivalent to 4×10^6 cells per ChIP) to the beads and rotate overnight at 4 °C.
7. Wash the beads at room temperature with rotation for 10 min each time.
   - 2× with 1 mL RIPA buffer.
   - 2× with 1 mL RIPA buffer with 0.3 M NaCl.
   - 2× with 1 mL LiCl buffer (0.25 M LiCl, 0.5 % NP40, 0.5 % NaDOC).

- 1× with 1 mL TE with 0.2 % Triton X-100.
- 1× with 1 mL TE.

8. Resuspend beads in 100 μL 1× TE and add 3 μL of 10 % SDS and 5 μL 20 mg/mL proteinase K. Incubate overnight at 65 °C.
9. Mix beads by vortexing, and use the magnet to transfer the supernatant to a new tube. Wash once with 100 μL 1× TE with 0.5 M NaCl. Combine two supernatants.
10. Use 200 μL phenol/chloroform (PCI) to extract DNA. Shake by hand for 30 s, spin at max speed for 10 min, and remove supernatant and add to new tube.
11. Add 2 μL GlycoBlue, 20 μL 3 M NaOAc, and 600 μL ethanol. Invert several times to mix. Incubate on dry ice or −80 °C for 2 h.
12. Centrifuge at max speed (tabletop) for 30 min at 4 °C.
13. Wash pellet with 70 % ethanol and centrifuge for 10 min at 4 °C.
14. Air-dry the pellet for 2–3 min and resuspend in 40 μL 1× TE.

3.4 Library Preparation for ChIP-Seq and RNA-Seq

Repair DNA ends

1. Follow the Epicentre End-It DNA End-Repair Kit to generate blunt-ended DNA.
2. Set up the following reaction in a 1.5 mL tube.
   - 34 μL DNA.
   - 5 μL 10× end-repair buffer (300 mM Tris–acetate, pH 7.8, 660 mM potassium acetate, 100 mM magnesium acetate, 5 mM DTT).
   - 5 μL 2.5 mM dNTPs.
   - 5 μL 10 mM ATP.
   - 1 μL End-Repair enzyme mix (T4 DNA polymerase, T4 polynucleotide kinase).
3. Incubate at room temperature for 45 min.
4. Purify DNA using a Qiagen MinElute Reaction Cleanup Kit. Elute in 32 μL EB buffer.

Add "A" overhang to 3′ ends

1. Mix the following reagents in order to a 1.5 mL tube.
   - 30 μL DNA from above.
   - 5 μL 10× NEB buffer #2.
   - 1 μL dATP (10 mM stock).
   - 11 μL water.
   - 3 μL 5 U/μL Klenow fragment (3′ → 5′ exo-).

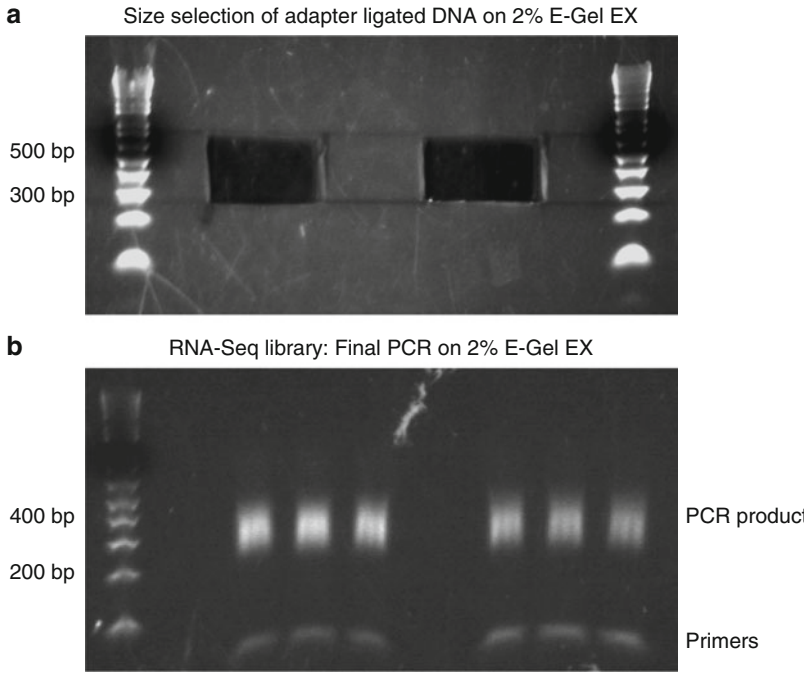


Fig. 3 Size selection and gel purification of libraries. (**a**) Adapter-ligated samples flanked by a 1 kb Plus DNA Ladder were run on a 2 % agarose E-Gel EX for 10 min. The 250–450 bp region was marked on the back of the EX plastic case, excised, and purified. (**b**) PCR products from adapter-ligated samples were run on a 2 % agarose E-Gel EX. Eighteen cycles of PCR were performed. The 1 kb Plus DNA Ladder is shown in the *first lane*

2. Incubate at 37 °C for 30 min.
3. Purify DNA using a Qiagen MinElute Reaction Cleanup Kit. Elute in 25 μL EB buffer.

Linker ligation for individual sample or multiplexed paired-end (PE) sequencing

1. Mix the following reagents in order on ice (*see* **Note 19**).
   - 23 μL DNA from above.
   - 3 μL 10× T4 DNA ligase buffer.
   - 1 μL PE adapter oligo mix or Index PE adapter oligo mix (*see* **Note 20**).
   - 3 μL T4 DNA ligase.
2. Incubate at room temperature for 30 min.
3. Size selection of DNA using a 2 % E-Gel EX gel. Run DNA on precast gel for 10 min. Cut region of gel around 250–450 bp (*see* **Note 21**) (Fig. 3a).
4. Purify the DNA using a Qiagen MinElute Gel Extraction Kit. Elute in 20 μL EB buffer (*see* **Note 22**).

PCR and purification

1. Mix the following reagents in order to an 8-well strip PCR tube.
   (a) *For Illumina PE adapter-ligated samples.*
   - 12 μL DNA from above (use 50 % of DNA).
   - 12 μL molecular grade water.
   - 25 μL master mix (2× Phusion HF, Finnzymes).
   - 1 μL PE PCR primer 1.0 (diluted 1:2; *see* **Note 23**).
   - 1 μL PE PCR primer 2.0 (diluted 1:2; *see* **Note 23**).

   (b) *For Illumina Index PE adapter-ligated samples (multiplexing kit).*
   - 12 μL DNA from above (use 50 % of DNA).
   - 12 μL molecular grade water.
   - 25 μL master mix (2× Phusion HF, Finnzymes).
   - 1 μL PCR primer InPE 1.0 (diluted 1:2; *see* **Note 23**).
   - 1 μL PCR primer InPE 2.0 (diluted 1:2; *see* **Note 23**).
   - 1 μL PCR primer Index #1–12 (diluted 1:2; *see* **Note 23**).
2. Perform PCR using the following cycling conditions.
   (a) Denature at 98 °C for 30 s.
   (b) 18 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s.
   (c) 72 °C for 5 min.
   (d) Hold at 4 °C.
3. Size selection. Use a precast 2 % E-Gel EX gel or 2 % agarose gel. Run DNA on a 2 % gel and excise the region corresponding to 250–450 bp (Fig. 3b).
4. Purify the DNA using a Qiagen MinElute Gel Extraction Kit. Elute in 20 μL EB buffer.
5. Measure the DNA concentration using a Qubit fluorometer.
6. Perform next-generation sequencing on an Illumina HiSeq or GAIIX machine.

3.5 RNA-Seq

Purification of mRNA from total RNA

1. Equilibrate mRNA purification reagents (oligo(dT) beads, binding buffer, wash buffer B, Tris–HCl) to room temperature for 15 min before proceeding to **step 2**.
2. Dilute 0.5–10 μg of total RNA with nuclease-free water to 50 μL in a 1.5 mL RNase-free tube.
3. Incubate at 65 °C for 5 min to denature RNA secondary structure and then place tube on ice for 2 min.

4. Aliquot 100 μL Dynal oligo(dT) beads to a 1.5 mL tube.
5. Wash beads twice with 100 μL binding buffer (20 mM Tris–HCl, pH 7.5, 1.0 M LiCl, 2 mM EDTA). Attach a 1.5 mL tube to the magnetic rack, wait 30 s, then aspirate off supernatant, and add binding buffer. Gently flick the tube to mix beads, attach to magnet, and remove the supernatant.
6. Resuspend the beads in 50 μL binding buffer and add 50 μL of total RNA from **step 3** above. Rotate at room temperature for 5 min, then attach to magnet, and remove supernatant.
7. Wash the beads twice with 100 μL wash buffer B (10 mM Tris–HCl, pH 7.5, 0.15 LiCl, 1 mM EDTA).
8. Prepare for second round of purification by aliquoting 80 μL binding buffer to a new 1.5 mL tube.
9. Remove the supernatant from the beads of **step 7** and add 20 μL 10 mM Tris–HCl and incubate at 80 °C for 2 min. Immediately place the tube with the beads on the magnet rack and transfer the supernatant (mRNA) to the tube containing 80 μL of binding buffer. Quickly add 100 μL of wash buffer B to the tube with beads.
10. Incubate the RNA at 65 °C for 5 min to denature, and place the tube on ice for 2 min.
11. Wash the beads from **step 9** twice with 100 μL wash buffer B by pipetting up and down several times. Remove the supernatant.
12. Add 100 μL RNA from **step 10** to the beads from **step 11**. Gently flick to mix and rotate at room temperature for 5 min.
13. Remove the supernatant and wash the beads twice with 100 μL wash buffer B.
14. Remove the supernatant and add 11 μL of 10 mM Tris–HCl to the beads. Incubate at 80 °C for 2 min. Immediately place the beads on the magnet and transfer the mRNA to a new 1.5 mL tube or 200 μL thin-walled PCR tube.

3.6 First-Strand cDNA Synthesis (Superscript Double-Stranded cDNA Synthesis Kit, Invitrogen)

1. Set up the following reaction in a 200 μL thin-walled PCR tube.
   - 10.5 μL mRNA from above.
   - 1 μL random hexamer primers (3 μg/μL, Invitrogen).
2. Incubate at 65 °C for 5 and 2 min at 4 °C.
3. Mix the following in order.
   - 4 μL first-strand buffer.
   - 2 μL 0.1 M DTT.
   - 1 μL dNTP mix.
   - 0.5 μL RNaseOUT.

4. Add 7.5 μL of the reaction mixture to the PCR tube and incubate at 25 °C for 2 min.
5. Add 1 μL Superscript II (200 U/μL) to the PCR tube and incubate according the following settings.
   - 10 min at 25 °C.
   - 50 min at 42 °C.
   - 15 min at 70 °C.
   - Hold at 4 °C.

3.7 Second-Strand cDNA Synthesis

Synthesize second-strand cDNA from the 20 μL reaction from above.

1. Place the PCR tube from above on ice, and add the following reagents in order to the mixture.
   - 91 μL DEPC-treated water.
   - 30 μL 5× second-strand reaction buffer.
   - 3 μL 10 mM dNTP mix.
   - 1 μL *E. coli* DNA ligase (10 U/μL).
   - 4 μL *E. coli* DNA polymerase I (10 U/μL).
   - 1 μL *E. coli* RNase H (2 U/μL).
2. Vortex gently to mix. Gently flicking the PCR tube may not sufficiently mix the reagents, which may result in inadequate second-strand synthesis.
3. Incubate at 16 °C for 2 h.
4. Add 2 μL of T4 DNA polymerase and incubate at 16 °C for 5 min.
5. Purify the double-stranded cDNA using a Qiagen PCR Purification Kit. Elute the cDNA in 40 μL EB buffer.

3.8 Fragment Double-Stranded cDNA

1. Use Bioruptor sonifier to sonicate the double-stranded cDNA (dscDNA) in a 1.5 mL tube to shortened fragments of dscDNA. Sonicate 3 × 10 min or 4 × 10 min with 30 s pulses on the medium setting on ice (*see* **Note 24**). After each cycle, briefly spin down the 1.5 mL tube, add fresh ice to the water bath, and continue the sonication cycles. Sonication efficiency can be evaluated by running 3–4 μL (~50 ng) of sonicated dscDNA on a 2 % gel.
2. Proceed to library construction (Subheadings 3.4, DNA end-repair, addition of "A" overhang, linker ligation, and PCR) for sequencing on the next-generation Illumina HiSeq or GAIIX platform.

4 Notes

1. MEFs can also be obtained from commercial vendors or prepared from E13.5 to E14.5 MEFs. Methods to harvest MEFs can be found in the literature and are readily available on the Internet.
2. ES cells can be derived by plating E3.5 mouse blastocysts on iMEFs in ES cell derivation media (DMEM high glucose, 20 % ESC-qualified FBS, LIF, pen strep, nonessential amino acids, 1× 2-mercaptoethanol, glutamine) and incubating at 37 °C with 5 % CO_2. After the blastocyst adheres to the layer of MEFs and the trophectoderm layer spreads out, the inner cell mass (ICM) outgrowth can be picked, dissociated in trypsin briefly, and replated in ES cell derivation media. Alternatively, established murine ES cell lines (e.g., R1) can also be used.
3. The GSK3β inhibitor, CHIR99021, and the MEK inhibitor, PD0325901, can be obtained from several commercial vendors. It is important to test the activity of small-molecule inhibitors acquired from various vendors.
4. Sonication in SDS-containing buffers may not be suitable for analysis of transcription factors or epigenetic regulators because SDS may disrupt protein–DNA interactions [13]. However, SDS increases efficiency of sonication and therefore is suitable for analysis of histone modifications such as H3K79 methylation which resides in the nucleosome core [13].
5. The stringency of chromatin immunoprecipitation (ChIP) buffers can be modified by changing the concentration of NaCl. For antibodies that have high specificity, a low concentration of NaCl may be used. However, for antibodies that have nonspecific binding, a high concentration of NaCl may be used to increase the signal and reduce the background noise.
6. Two rounds of poly-A mRNA purification are necessary to ensure enrichment of mRNA from total RNA.
7. It is important to add the reagents in the order described in the kit. Addition of reagents prematurely may reduce the activity of the enzymes which may result in incomplete cDNA synthesis.
8. It is important to use random hexamer primers at a concentration of 3 μg/μL. Use of random hexamers at a concentration of 50 ng/μL will be insufficient to carry out the second-strand synthesis reaction.
9. It may be possible to substitute Superscript III for Superscript II during the first-strand cDNA synthesis reaction. However, because these conditions have not been optimized by the manufacturer, it is recommended to use Superscript II.

10. When using the adapter oligo mix from Illumina, the adapter should be diluted appropriately to avoid adapter-dimer contamination during the PCR step of library construction.
11. When using the E-Gel system, it is important to use the EX model of gels. Other models may be difficult to open and subsequently perform size selection.
12. Other high-fidelity DNA polymerases may be substituted for the Phusion 2× master mix. However, because enzymes from various vendors may perform differently, it is important to test the activity of these enzymes.
13. Custom PCR primers can be substituted for the Illumina PE primers 1.0 and 2.0. Illumina indexing and PCR primer sequences can be found by searching the internet. Although the modifications of the Illumina PCR adapters and primers are not disclosed, it has been shown that addition of a phosphorothioate group between the two bases at the 3′ end prevents nuclease digestion [14].
14. Use of Lab Armor beads reduces contamination often found in water baths (e.g., fungal, bacterial).
15. ES cell colonies should not be cultured to a density that they are touching each other. This will result in a loss of self-renewal or differentiation.
16. It is important to aliquot cells before freezing because fixed cells should not be freeze-thawed.
17. SDS-containing buffer should be used for histone modifications, but may not be suitable for protein factors (e.g., transcription factors, epigenetic regulators).
18. Sonication conditions need to be empirically optimized due to variability between sonifiers, cells, fixation conditions, etc.
19. The adapter ligation reaction should contain a tenfold molar excess of adapter relative to the DNA template.
20. Illumina's PE adapter can be used for non-multiplexed single-end (SE) sequencing or PE sequencing. Illumina's Index PE adapter, or custom in-house indexing adapters, can be used for multiplexed PE sequencing.
21. Cutting the gel slightly above 200 bp (e.g., 250 bp) will minimize contamination of unligated adapters that form dimers during the ligation step. If these dimers are not removed, they will saturate the PCR reaction in the subsequent step.
22. During the gel purification step, it may be useful to melt the agarose gel at room temperature rather than 50 °C to avoid bias of guanosine–cytidine-enriched sequences due to a depletion of adenosine and thymidine [13, 14].

23. If Illumina's PE adapter was used for the ligation step, use Illumina PE primer 1 and PE primer 2 for the PCR step. If Illumina's Index PE adapter was used for the ligation step, use Illumina's PCR primer InPE 1.0, PCR primer InPE 2.0, and PCR primer Index #1–12 for the PCR step.
24. If 0.5 μg of total RNA was used for the mRNA purification step, three cycles of sonication should be sufficient. If 1–5 μg of total RNA was used for the mRNA purification step, four cycles of sonication may be needed. It is important to empirically test sonication conditions.

Acknowledgments

This work was supported by the Division of Intramural Research Program of the NIH, National Heart, Lung, and Blood Institute.

References

1. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133(6):1106–1117
2. Kim J, Chu J, Shen X, Wang J, Orkin SH (2008) An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132(6):1049–1061
3. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR (2006) Dissecting self-renewal in stem cells with RNA interference. Nature 442(7102):533–538
4. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153):553–560
5. Kidder BL, Palmer S (2010) Examination of transcriptional networks reveals an important role for TCFAP2C, SMARCA4, and EOMES in trophoblast stem cell maintenance. Genome Res 20(4):458–472. doi:10.1101/gr.101469.109, gr.101469.109 [pii]
6. Kidder BL, Palmer S (2012) HDAC1 regulates pluripotency and lineage specific transcriptional networks in embryonic and trophoblast stem cells. Nucleic Acids Res 40(7):2925–2939. doi:10.1093/nar/gkr1151, gkr1151 [pii]
7. Kidder BL, Palmer S, Knott JG (2009) SWI/SNF-Brg1 regulates self-renewal and occupies core pluripotency-related genes in embryonic stem cells. Stem Cells 27(2):317–328. doi:10.1634/stemcells.2008-0710, stemcells.2008-0710 [pii]
8. Kidder BL, Yang J, Palmer S (2008) Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells. PLoS One 3(12):e3932. doi:10.1371/journal.pone.0003932
9. Ang YS, Tsai SY, Lee DF, Monk J, Su J, Ratnakumar K, Ding J, Ge Y, Darr H, Chang B, Wang J, Rendl M, Bernstein E, Schaniel C, Lemischka IR (2011) Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell 145(2):183–197. doi:10.1016/j.cell.2011.03.003, S0092-8674(11)00240-6 [pii]
10. Pasini D, Cloos PA, Walfridsson J, Olsson L, Bukowski JP, Johansen JV, Bak M, Tommerup N, Rappsilber J, Helin K (2010) JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature 464(7286):306–310. doi:10.1038/nature08788, nature08788 [pii]

11. Kidder BL, Hu G, Yu ZX, Liu C, Zhao K (2013) Extended self-renewal and accelerated reprogramming in the absence of Kdm5b. Mol Cell Biol 33:4793–4810. doi:10.1128/MCB.00692-13, MCB.00692-13 [pii]
12. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A (2008) The ground state of embryonic stem cell self-renewal. Nature 453(7194):519–523. doi:10.1038/nature06968, nature06968 [pii]
13. Kidder BL, Hu G, Zhao K (2011) ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol 12(10):918–922. doi:10.1038/ni.2117, ni.2117 [pii]
14. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5(12):1005–1010. doi:10.1038/nmeth.1270, nmeth.1270 [pii]

Chapter 2

Analysis of Next-Generation Sequencing Data Using Galaxy

Daniel Blankenberg and Jennifer Hillman-Jackson

Abstract

The extraordinary throughput of next-generation sequencing (NGS) technology is outpacing our ability to analyze and interpret the data. This chapter will focus on practical informatics methods, strategies, and software tools for transforming NGS data into usable information through the use of a web-based platform, Galaxy. The Galaxy interface is explored through several different types of example analyses. Instructions for running one's own Galaxy server on local hardware or on cloud computing resources are provided. Installing new tools into a personal Galaxy instance is also demonstrated.

Key words NGS, Genomics, Informatics, RNA-seq, ChIP-seq, Workflows, Reproducibility, Open source, Web-based workbench, Big data analysis

1 Introduction

Recent advances in next-generation sequencing (NGS) technology have created a situation where raw data generation is no longer a rate-limiting factor in many genome-scale studies. However, the scale of the data presents not only difficulties for individual researchers attempting to analyze the data but also significant informatics issues for collaboration and reproducibility. Furthermore, simply generating data does not, in itself, lead to an increase in knowledge. Any technological advancement is limited in its ability to uncover new biological meaning if the ability of researchers to interact seamlessly with the data at any step is lacking. An important aspect of this interaction is providing researchers access to software tools.

Galaxy [1–3] is an open source, web-based platform for accessible, reproducible, and transparent computational biomedical research. Galaxy makes bioinformatics analyses accessible to users lacking programming knowledge by enabling them to interactively specify parameters for running tools and Workflows through a point-and-click interface. Every computational analysis is made reproducible by automatically capturing tool parameters and other

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_2, © Springer Science+Business Media New York 2014

Table 1
Common galaxy terminology

Term	Description
Dataset	These are the inputs and outputs from each step in an analysis. Each time a tool is executed, a new Dataset is created that contains the results
Tool	An operation within Galaxy that acts upon Datasets as an analysis step. The underlying function may be developed by the Galaxy team or may be a third party program
History	A persistent container for an analysis. Each Dataset belongs to at least one History. As tools are run, Datasets appear chronologically within the active History. Each step of an analysis is recorded as the Datasets within a History
Workflow	A reusable analysis pipeline that allows any number of analysis steps to be performed automatically. Workflows that group tools into a single functional unit can be created de novo or by extracting directly from a History
Instance	A Galaxy instance is a single occurrence of a Galaxy server. There can be any number of Galaxy instances in existence at any particular time. Running a local Galaxy server would be an example of a Galaxy instance. Every Galaxy instance is independent and has its own set of Users, Datasets, and other objects
Main	The primary public Galaxy instance located at http://usegalaxy.org

information so that any user can repeat and understand the complete analysis. Transparency is maintained by allowing users simple access to share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis. A free public instance of Galaxy is available at http://usegalaxy.org. A local instance of Galaxy can be installed by following the directions at http://getgalaxy.org or run in the cloud by accessing http://usegalaxy.org/cloudlaunch. New tools can be easily installed into a Galaxy instance from Galaxy's application store, known as the ToolShed, available at http://usegalaxy.org/toolshed.

In addition to this chapter, users that are not familiar with Galaxy are directed to follow the tutorial available at http://usegalaxy.org/galaxy101 and to explore the resources at http://galaxyproject.org and https://vimeo.com/galaxyproject. *See* Table 1 for a list of common Galaxy terminology.

The live supplemental is located at https://usegalaxy.org/u/galaxyproject/p/ngs-analysis-2013.

2 Materials

2.1 Requirements for Using the Galaxy Interface

A computer with a modern web browser that supports JavaScript and HTML5 is required to use the interactive Galaxy interface. Most current Internet browsers are supported, such as Firefox,

Chrome, Safari, and Opera. JavaScript must be enabled and any plug-ins that the user has installed that block JavaScript should be disabled (e.g., NoScript).

2.2 Requirements for Analyzing Next-Generation Sequencing Data

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and a user account at the public Galaxy server (http://usegalaxy.org) or a local Galaxy installation or Galaxy cloud instance are needed.

NGS data in the FASTQ format is needed (*see* **Note 1**). You may supply your own Datasets or use the example data listed at the start of the methods.

2.3 RNA-Seq Analysis with Galaxy

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and a user account at the public Galaxy server (http://usegalaxy.org) or a local Galaxy installation or Galaxy cloud instance are needed.

RNA sequencing data in the FASTQ format is needed. You may supply your own Datasets or use the example data listed at the start of the methods.

2.4 ChIP-Seq Analysis with Galaxy

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and a user account at the public Galaxy server (http://usegalaxy.org) or a local Galaxy installation or Galaxy cloud instance are needed.

Chromatin immunoprecipitation (ChIP) sequencing data in the FASTQ format is needed. You may supply your own Datasets or use the example data listed at the start of the methods.

2.5 Creating and Running Galaxy Workflows

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and a user account at the public Galaxy server (http://usegalaxy.org) or a local Galaxy installation or Galaxy cloud instance are needed.

A Galaxy History used to create Workflows is also needed; these can be created in the previous subsections or can be other Histories created independently of this chapter.

2.6 Sharing and Publishing with Galaxy

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and a user account at the public Galaxy server (http://usegalaxy.org) or a local Galaxy installation or Galaxy cloud instance are needed.

Galaxy objects to share are also needed; these can be created in the previous subsections or can be other items created independently of this chapter.

2.7 Installing a Local Galaxy Instance

Installing Galaxy requires command-line access to a computer running a POSIX compliant operating system (OS) such as a Linux distribution or Mac OS X. Python2.6 or Python2.7 (http://www.python.org/getit) and Mercurial (http://mercurial.selenic.com) are

required to run and download Galaxy, respectively. Internet access is required during installation, but the computer does not need to have a fully public IP.

When a POSIX noncompliant OS, e.g., Windows, is on the computer, it is possible to use virtualization software, such as VirtualBox (https://www.virtualbox.org), recommended, or VMware Player (http://www.vmware.com), to install a compatible guest OS, such as Ubuntu (see, e.g., https://help.ubuntu.com/community/VirtualBox), in order to run Galaxy inside of a virtual machine. A virtual machine can also be used even if the host OS is POSIX compliant.

2.8 Running Galaxy in the Cloud

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and an Amazon Web Services (AWS) account with a valid payment method are required to use Galaxy CloudLaunch. Access to a Secure Shell (SSH) client is needed for advanced configuration that is not covered in this chapter. Currently, Amazon EC2 is supported. OpenStack-based cloud services have also been used, such as within the Australian NeCTAR cloud, but will not be covered here.

2.9 Installing New Tools via the Galaxy ToolShed

A computer with a modern web browser that supports JavaScript and HTML5 is required. Access to the Internet and Admin access to a Galaxy instance are needed to install tools (*see* Subheadings 2.7 and 2.8).

3 Methods

3.1 Using the Galaxy Interface

Galaxy is divided into several different operational interfaces. The most commonly used interface is the Analysis interface and is the first one encountered upon loading a Galaxy instance. The Analysis interface is accessed by clicking on "Analysis" within the Galaxy masthead or by clicking on "Galaxy" in the top left. Additional interfaces include the Workflow interface, the Data Library and Shared Data interfaces, the Visualization interface, and the Admin interface. Access to Galaxy can also occur via a RESTful application programming interface (API; *see* **Note 2**) but is beyond the scope of this chapter.

1. Start the web browser and load http://usegalaxy.org. This is the Main public Galaxy instance run by the Galaxy Project team. The Galaxy interface is divided into four main parts: the masthead, the tools menu, the tool interface, and the user History (*see* Fig. 1):
   (a) At the top of the page is the masthead. This allows the user to access help, user account settings, and shared data and to change the interface view. Also visible are the user's current data usage and quota status.

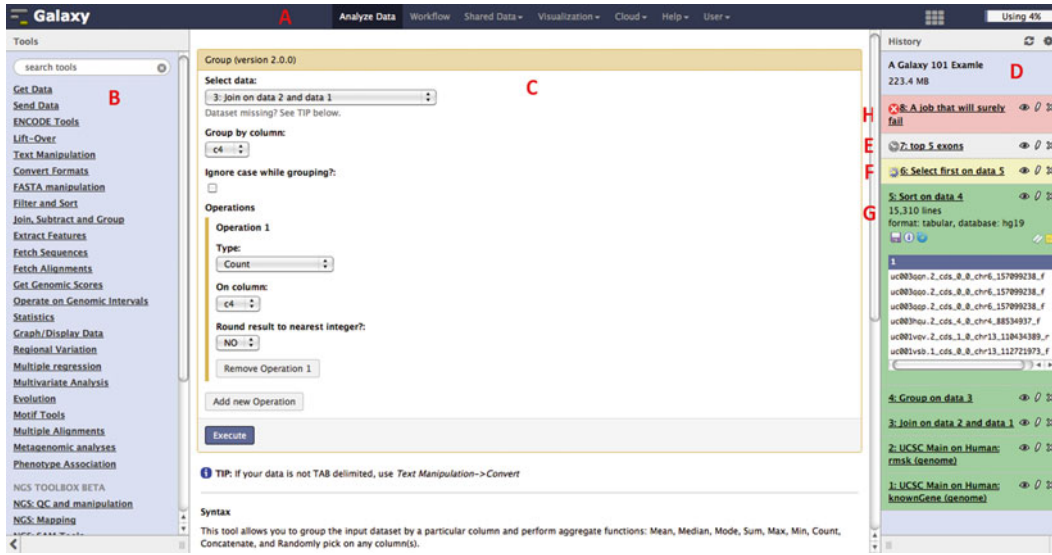


Fig. 1 The Galaxy Analysis interface. The Galaxy Analysis interface is constructed of four main parts: (*A*) the masthead at the *top*, (*B*) the tools menu on the *left*, (*C*) the tool interface in the *center*, and (*D*) the analysis History located on the *right*. The current status of a Dataset is indicated by its color and associated icon. Datasets that are queued (*E*) for execution, either due to limited compute resources or due to waiting for an input Dataset to become available, are *gray*. Datasets that are currently running (*F*) are *yellow*. A *green* Dataset indicates that the Dataset is ready to be used (*G*) and that the creating tool has finished executing successfully. A Dataset will be *red* in color when there has been an error with the analysis step (*H*)

(b) On the left-hand side is the Tools menu. Here, tools are organized into sections based upon function; clicking on a section will expand the available tools. Clicking on the underlined name of a tool will cause that particular tool interface to appear in the middle pane.

(c) The tool interface in the middle pane allows users to select input Datasets and to configure tool parameter settings. As a result of Galaxy's datatype system, only input Datasets that are valid for a particular tool input will be selectable. Clicking the "Execute" button here will cause an analysis job to run in the background and create one or more output Datasets in the user History.

(d) The user History is located on the right-hand side and contains the output Datasets from every step performed during a particular analysis. Every time a Dataset is uploaded or an analysis step is performed, one or more output Datasets are created within the History, with the newest steps appearing at the top. Clicking on the name of a Dataset within the History will expand to show a peek of the Dataset content and allow the user to access additional information and actions that can be performed on the Datasets, such as downloading, viewing the tools and

parameters that generated the Dataset, and allowing the job to be rerun.

2. In the masthead, click "User" and then click "Register." Enter the information requested and click "Submit" to register an account. A confirmation email will be sent for verification to activate the request. While user registration is optional, accessing advanced Galaxy features, such as multiple analysis Histories, Sharing, and Workflows, requires the user to be logged in. Additionally, at the Main public Galaxy server, registered users have a larger Dataset storage quota, allowing more data to be uploaded and more analyses to be performed.
3. In the History pane, click the gear icon to show History options, click "Create New," to create a new empty analysis History.
4. In the History pane, click "Unnamed history" and enter a new name for the History, such as "Learning Galaxy" (*see* **Note 3**).
5. In the Tools menu, click "Get Data" to expand the tool section.
6. Under the "Get Data" tool section, click "Upload File" to display the tool form interface. The Upload tool allows users to get external data into Galaxy by four methods: uploading from the user's computer, entering free text, copying and pasting one or more URLs for Datasets, or through FTP. Video examples can be found at http://vimeo.com/galaxyproject/upload.
7. In the URL/text box, enter "http://goo.gl/8Y3K8r" (without the quotes) and click "Execute." A new Dataset containing sequencing reads in the FASTQ format will appear in the History. In this case it is not necessary to change the file format parameter from "Auto-detect," as Galaxy will determine that the file is in the FASTQ format (*see* **Note 4**).
8. Datasets can also be loaded from a Data Library. Data Libraries are available by clicking on "Shared Data" in the masthead and selecting the "Data Library" option. This will present you with a list of the available Data Libraries on your current Galaxy instance.
9. Open a Data Library by clicking on the name of the Library. The Library will load into the center pane.
10. Datasets within Data Libraries are contained within folders. To expand folders, click on the folder icon next to the name of the folder.
11. Datasets can be loaded into a History by selecting the checkbox next to the name of the desired Dataset and then clicking the "Go" next to the "Import into current History" action at the bottom of the Library.
12. Multiple Datasets can be imported at one time and all of the folder's Datasets can be imported by checking the box next to the desired folder.

13. Datasets can also be loaded through Shared or Published Histories.
14. Shared History links may be entered as URLs into the browser and submitted. At the top right corner, the smaller green plus icon will read "Import" when moused over. Click on this icon to import the History and contained Datasets into the current History. Creating a new History as in **step 3** is advised.
15. Shared Histories can also be found by selecting "Histories Shared with Me" from the History menu. This menu is represented by a small gear icon at the top of the History pane, far right of the History name. From here, Shared Histories can be copied and worked with or unshared if no longer needed.
16. Published Histories are located under the masthead menu "Shared Data" and then "Published Histories." They are also often included in "Published Pages" as embedded data. From either location, the same icon will be present and the same process can be used as in **step 14** to import the History and/or Datasets.
17. Video walkthrough of Dataset attributes is at http://vimeo.com/galaxyproject/datasets1.

3.2 Analyzing Next-Generation Sequencing Data

Typically, NGS analysis in Galaxy begins with raw sequencing data in the FASTQ format. To facilitate the use of supplemental data required for certain tools as well as to support alternate analysis paths, Galaxy permits users to upload or import SAM/BAM, BED, GTF/GFF3, VCF, and other common bioinformatics file formats from local file systems or from integrated external sources including BioMart [4], UCSC Table Browser [5], GenomeSpace (http://genomespace.org), and others:

1. We will start with the History created in the previous section (Subheading 3.1), which now contains a set of sequencing reads as Datasets (*see* **Note 3**).
2. In the Tools menu, click "NGS: QC and manipulation" to expand the tool section.
3. Access the FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) tool by clicking the name "FastQC: Read QC" (*see* **Note 5**).
4. Make sure that the newly uploaded FASTQ Dataset is selected under "Short read data from your current History."
5. Click "Execute" to run the FastQC tool and to create a new Dataset containing an HTML report containing various metrics about the FASTQ Dataset.
6. Once the tool has finished (Dataset has turned green; *see* Fig. 1), click on the Eye icon to display the FastQC report in the middle pane. Pay particular attention to the reported FASTQ variant type.

7. In the Tools menu, click "FASTQ Groomer" to access the FASTQ conversion and validation tool (*see* **Notes 6** and **7**).
8. Make sure that the uploaded FASTQ Dataset is selected under "File to groom" and click "Execute." The default settings are acceptable in this case and a fastqsanger Dataset will be created (*see* **Note 8**). If the Dataset input is different than the Sanger variant, select the variant type reported by the FastQC tool under "Input FASTQ quality scores type." If an output variant other than fastqsanger is desired, change "Advanced Options" to "Show Advanced Options" and select the appropriate settings.
9. The quality scores for sequenced bases generally decrease along the length of the read. Low-quality bases can have a significant impact upon alignment and other downstream steps. Based upon the FastQC report, it may be desirable to trim the ends of reads or to remove low-quality reads altogether by filtering.
10. This video on FASTQ data demonstrates how to groom or assign the proper Galaxy datatype for Illumina data (http://vimeo.com/galaxyproject/fastqprep).

3.3 RNA-Seq Analysis with Galaxy

RNA-seq is the practice of sequencing RNA. Generally, RNA is harvested from an organism and converted to DNA using reverse transcriptase. The resultant cDNA is then sequenced. While methods for sequencing each type of RNA have been developed, we will discuss the most commonly studied, mRNA. mRNA is usually purified by taking advantage of the poly-A tail. Two different types of information are usually investigated during an RNA-seq experiment: (1) alternative splicing and (2) differential expression of genes.

1. Load RNA-seq Datasets into a new History. Some sample FASTQ sequencing reads are available within a Data Library called "2013—MiMB—Stem Cell Transcriptional Networks" under the folder "RNA-seq."
2. Prepare the paired-end FASTQ sequencing reads using the techniques utilized in Subheading 3.2 on each Dataset.
3. Map the sequencing reads to a reference genome (hg19) using TopHat. Open the "Tophat2 Gapped-read mapper for RNA-seq data" tool [6], found under the "NGS: RNA-seq" section.
4. Set "Is this Library mate-paired?:" to "Paired-end."
5. Set "RNA-Seq FASTQ file, forward reads:" to the FASTQ Dataset that has been imported and prepared in **step 2** that corresponds to the forward reads (names end in "/1").
6. Set "RNA-Seq FASTQ file, reverse reads:" to the FASTQ Dataset that has been imported and prepared in **step 2** that corresponds to the reverse reads (names end in "/2").
7. Select the proper reference genome ("hg19").
8. Leave "TopHat settings to use" set to "Commonly Used."

9. Click "Execute" to start the job.
10. Repeat **steps 1–9** for each set of paired-end data.
11. Access the tool "Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data" [7].
12. For each mapped set of data, individually, set "SAM or BAM file of aligned RNA-Seq reads:" to the output of TopHat.
13. Click "Execute."
14. The output of Cufflinks provides a set of assembled transcripts along with estimates of isoform-level relative abundance known as FPKM.
15. Access the "Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments" tool.
16. Set "GTF file produced by Cufflinks:" to the GTF output from one of the Cufflinks results.
17. Click "Add Additional GTF Input Files" for each additional Cufflinks output that was created previously.
18. Set the additional "GTF file produced by Cufflinks:" inputs to the remaining Cufflinks outputs.
19. Set "Use Reference Annotation:" to "Yes" and select the GTF reference annotation file that was imported from the Library.
20. Click "Execute."
21. Examine the output Dataset to locate transcripts that were assembled in each of the input Datasets.
22. Access the "Cuffdiff find significant changes in transcript expression, splicing, and promoter use" tool to assess differential expression.
23. Under "Transcripts:" select the combined transcripts Dataset created by Cuffcompare.
24. Set the "Replicates" input to the TopHat accepted hits output for each of the RNA-seq conditions. To include additional conditions, click "Add new Conditions."
25. Click "Execute." Several outputs will be created that assess any significant changes in transcript expression, splicing, and promoter use between the RNA samples.

3.4 ChIP-Seq Analysis with Galaxy

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an essential technique for genome-wide profiling of protein binding, histone modification, and nucleosome positioning. Generally, within a ChIP-seq experiment, to determine protein-binding locations, bound proteins are cross-linked to their bound DNA locations and then sheered. The cross-linked protein–DNA molecules are passed over a column where the molecules bind to a treated substrate and the non-bound DNA is washed away.

The protein-bound DNA is then eluted from the column and the cross-linking is reversed. The precipitated DNA is purified and then sequenced to determine the DNA sequence around the previously bound loci. The sequencing reads are then mapped to a reference genome and enriched regions are determined (a process commonly called peak calling):

1. Load ChIP-seq Datasets into a new History. Some sample FASTQ sequencing reads are available within a Data Library called "2013—MiMB—Stem Cell Transcriptional Networks" under the folder "ChIP-seq."
2. Prepare the FASTQ sequencing reads labeled as "Enriched" using the techniques utilized in Subheading 3.2.
3. Map the sequencing reads to a reference genome (mm9) using BWA [8] or Bowtie [9] (not shown). Open the "Map with BWA for Illumina" tool, found under the "NGS: Mapping" section.
4. Select the proper reference genome ("mm9").
5. Set "FASTQ file:" to the FASTQ Dataset that has been imported and prepared in **step 2**.
6. Leave "BWA settings to use" set to "Commonly Used."
7. Click "Execute" to start the job.
8. Repeat **steps 2–6** on the Dataset labeled "Control."
9. Within the "NGS: Peak Calling" section, access the tool named "MACS Model-based Analysis of ChIP-Seq" [10].
10. Set "ChIP-Seq Tag File:" to the prepared and mapped Enriched Dataset.
11. Set "ChIP-Seq Control File:" to the prepared and mapped Control Dataset.
12. Set the reference genome to match the previously used reference genome (mm9).
13. Optionally check "Parse xls files into distinct interval files:," set "Save shifted raw tag count at every bp into a wiggle file:" to be "Save," and set "Resolution for saving wiggle files:" to be "1."
14. Click "Execute" to start the peak calling job.
15. The output Datasets consist of one or more result files (a–e) and an HTML summary report (f). These outputs take the form of the following:
    (a) Standard output—peaks: bed.
    (b) Optional output—peaks: interval.
    (c) Optional output—negative peaks: interval.
    (d) Optional output—treatment: wig.
    (e) Optional output—control: wig.
    (f) Standard output—html report.

16. Click the name of the Dataset labeled as "peaks: bed" to expand its contents.
17. Click on the link labeled "display at UCSC main" to view called peaks in a genomic context at the UCSC Genome Browser [11]. Are there any visual correlations between the called peaks and other annotation tracks, such as genes or conservation?

3.5 Creating and Running Galaxy Workflows

Workflows are groups of tools linked together to form an analysis pipeline that can be launched in batch, reused, edited, and started through the user interface or API and are an invaluable aid in ensuring exact replication of experimental conditions when used on a series of input data. Galaxy Workflows can be extracted from existing Histories or created de novo using the Workflow Editor. Once created, any Workflow can be viewed, copied, downloaded, edited again, shared, published, renamed, or deleted at any time.

Galaxy Workflows can be exported and imported across Galaxy instances and many from the Galaxy team and community are available both on the Main public Galaxy (http://usegalaxy.org) at "Shared Data" under "Published Workflows" and in the Main ToolShed (http://usegalaxy.org/toolshed) under "Search for Workflows" (leave search box empty and click on button "search repositories" to view all).

If it has been some time since you have run the original Workflow and the "Extract Workflow" function is invoked or you are uploading or importing a Workflow from another Galaxy server (click on masthead's "Workflows" menu to bring up the "Your Workflows" home page, and find the button in the upper right corner named "Upload or import workflow"), updates or notices about missing tools may be presented. Accept these changes when saving the Workflow on the current server, or, if using a local or cloud instance, adjust tool content on the server as needed (*see* Subheadings 3.7, 3.8, and 3.9). More Workflow help is available at http://wiki.galaxyproject.org/Learn/AdvancedWorkflow.

The first method describes how to extract a Workflow from an existing History, edit it, and then run it:

1. We will start with the History created by Subheading 3.3, which now contains a populated analysis History (*see* **Note 3**).
2. In the History pane, click on the top right gear icon to expand the "Histories List" menu.
3. Select the option "Extract Workflow" to bring up the associated form in the center pane of the Galaxy interface.
4. The form contains the following informative text:

   "The following list contains each tool that was run to create the Datasets in your current History. Please select those that you wish to include in the Workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a Workflow will be shown in gray."

5. In the box under the text "Workflow name," modify the name to be something more meaningful than the default, if desired.
6. Review the tools on the left side of the form and the Datasets on the right side of the form for accuracy. Uncheck any that are not needed in the final analysis path.
7. Input Datasets do not have automated tools to retrieve the data and must be loaded, imported, or created by the user.
8. It is important to make a careful note of the input Dataset's expected content and format, which tools they are used in first, and even the ordering in the History (while scientifically arbitrary, the Workflow Editor will by default organize inputs during runtime in the same order as for display). This will help with labeling the Workflow's data inputs during editing.
9. Alternatively, a second browser window opened to this same page/view can be opened and retained for reference, while the first window moves forward with the next steps.
10. Click on the box "Create Workflow" when finished.
11. Click on "Workflows" in the top masthead menu choices.
12. On the Workflow home page, under "My Workflows," the Workflow just created will be listed first by default if sort by creation date is preserved.
13. Click on the end of the button named for your Workflow, near the down arrow on the right side. A menu will open.
14. Click on "Edit" from the "Workflow List" to open the Workflow Editor.
15. Orientation:
    (a) *Tool icons* are on the left pane.
    (b) *Canvas model* is in the middle pane.
    (c) *Tool options* are on the right pane.
    (d) *Navigation* is in the bottom right corner of the middle pane.
    (e) *Editor List* is in the small gear icon found right above the middle pane.
16. Best practices:
    (a) Make a second copy (backup) of anything important before you begin to edit it. Use the "Copy" function from the "Workflow List."
    (b) Save your Workflow periodically while editing, especially before or after significant edits. Using "Editor List" choice "Save."

(c) Save your Workflow before navigating away from the Workflow Editor or changes will be lost, same method as b. above.

17. Assign a name to each input of the extracted Workflow loaded into the editor:
    (a) Click on each box named "Input Dataset" in the canvas, one at a time.
    (b) Name is entered in the "Tool Options," top of the right pane, in the text box labeled "Name:."
    (c) Name each so that proper inputs from the History can be matched up with Workflow inputs during run setup and execution.

        For example, if a particular input is a reference annotation Dataset in the GTF format, label the input "GTF reference annotation."
18. Hide intermediate analysis steps:
    (a) Click on the "Flag output" icon (asterisk, sometimes referred to as a snowflake) within the final output step(s).
    (b) This small icon is located in the bottom right corner of tool icons in the Workflow canvas. Hovering over the icon will pop-up display the text:

        "Flag this as Workflow output. All non-flagged outputs will be hidden."
19. Save Workflow using "Editor List" choice "Save."
20. Run Workflow using "Editor List" choice "Run."
21. Set "Input Datasets" in the Workflow displayed in the center panel to be the proper Datasets from the right History panel (the original History). This Workflow can be run on any similar data ongoing after a test run.
22. Click on the box next to "Send the results to new History," and when the option expands, type in a meaningful History name into the new text box.
23. Click on "Run Workflow."
24. Once launched, a message will appear in the middle pane indicating the location of the new History along with the tools used, analysis steps performed, and resulting Datasets that will be produced.
25. Follow the link to view the new results and compare to the original results, first setting hidden Datasets to viewable using the "Histories List" menu option "Include Hidden Datasets" (*see* **Note 9**).
26. Results will be complete, identical, and reproducible unless one or more of the inputs changed (tool version, data content) or a nondeterministic tool is used.

The second method describes how to create a Workflow de novo using the Workflow Editor.

27. In this example, a simple Workflow duplicating the analysis steps performed in Subheading 3.2 will be created. The tools used in this example, in actual use, or by following the above instructions will guide the contents and construction of the Workflow.
28. We will start with a copy of the History created in the previous section (Subheading 3.1), which contains a set of sequencing reads as Datasets (create a copy if needed; *see* **Note 3**). These Datasets will be the "Input Datasets" of our Workflow when executed.
29. Click on the masthead menu "Workflow" to reach the Workflow homepage.
30. Click on the button "Create new workflow" located near the upper right corner.
31. Within the Workflow Editor, start by selecting the "Input Dataset" icons from the left tool icon pane. These are located at the bottom. For each input Dataset, click to to create the input module and drag into the desired position. Name/label as in the prior Workflow method as desired.
32. Next, add in the other tools to represent the analysis steps for the Subheading 3.2 methods. Drag each tool over and arrange.
33. When a tool is clicked on, the right pane will display the tool options. These are exactly the same options as presented when using tools directly. Modify parameters and add names or annotation to suit requirements.
34. Connect tools together by using the flexible "noodles" that extend out from the right side of tool icons (representing outputs) and insert into the left side of tool icons (representing inputs).
35. Choose which output Datasets will be hidden or not using the same method explained in **step 18** above (or a variation).
36. Duplicate **steps 19–26** above substituting the input Datasets and History for the one prepared in **step 28** above.
37. Reuse, expand, and explore more ways to use Workflows!

Bonus method: How to make your Workflows appear as tools in your tool menu on any server (including the Main public Galaxy instance).

38. Click on the masthead menu "Workflow" to reach the Workflow homepage.
39. At the bottom of the page under "Other options," click on the button "Configure your Workflow menu."
40. On the next form, in the farthest right column labeled "Show in menu," check the box for the Workflows that you want to appear in your tool menu and click "Save."

41. To return to the analysis interface, click on the masthead menu "Analyze Data."
42. Scroll down to the bottom of the left Tool Pane, under the section "Workflows," the Workflows that you checked will be listed individually.
43. "All Workflows" is present by default and is a quick way to bring up Workflows in the center pane, without leaving the active History.

3.6 Sharing and Publishing with Galaxy

Central to Galaxy's mission of enabling accessible, reproducible, and transparent research are the Sharing and Publishing functions common to all core objects, namely, Workflows, Visualizations, Pages, and Histories (including the Datasets contained within). Galaxy objects can be Shared or Published using a single click within the Galaxy interface. Sharing a Galaxy object can be done directly with another Galaxy user by utilizing their account's e-mail address on the same Galaxy server or by generating and communicating a Share link (email, publication, or similar) that will allow anyone that knows the link to access the shared item. When an item is Published within Galaxy, it is made publicly available under a central "Shared Data" hub for the object type where all users can view, search, tag, vote, import, and use. A Galaxy account is not required to view the content included in a Share link. Shared and Published objects may also be embedded into Galaxy Pages, which in turn can be Shared or Published. Galaxy Pages are particularly significant in that they provide a customized and organized means to communicate exact data sources, methods, results, and discussion related to analysis. The creation of a Galaxy Page utilizes an interactive word-processing style editor directly within the web browser. These are commonly used as a platform for publication supplemental materials and tutorials. Workflows, Datasets, and Histories can be directly imported from a Page and into the user's own workspace to be modified or reused. To learn more about the Share or Publish features in Galaxy, please see http://wiki.galaxyproject.org/Learn/Share and http://vimeo.com/galaxyproject/sharepublish.

1. We will start with the History created in the prior section (Subheading 3.4), which now contains a populated analysis History (*see* **Note 3**).
2. In the History pane, click on the top right gear icon to expand the "Histories List" menu.
3. Select the option "Share or Publish" to bring up the associated form in the center pane of the Galaxy interface.
4. There are three Share or Publish options grouped into two sections. Option A or B, plus C may be made active for any single History at a time. Selections can be modified at any time while on the form or by returning to the form.

5. Group 1: "Make History Accessible via Link and Publish It." A single option may be chosen:
   (a) Option A, button: "Make History Accessible via Link." This generates a web link that you can share with other people so that they can view and import the History.
   (b) Option B, button: "Make History Accessible and Publish." This makes the History accessible via link (see above) and publishes the History to Galaxy's *Published Histories* section, where it is publicly listed and searchable.
6. Group 2: "Share History with Individual Users":
   (a) Option C, button: "Share with a user." This directly shares a History with another user having an account on the same Galaxy server.
7. Sharing and Unsharing the History with a single user directly:
   (a) Click on the button from Option C from **step 6a** above, to initiate sharing the History with a single user directly.
   (b) A new form will open with an empty text box labeled "Galaxy user emails with which to share Histories."
   (c) Type into the box either a known user's account email address or optionally enter the email address "outreach@galaxyproject.org." Click "Submit."
   (d) The user's email will now appear as a button under the section "Share History with Individual Users." Click on the button to "Unshare" when no longer needed.
8. Sharing and Unsharing the History with one or more users via a link:
   (a) Click on the button from Option A from **step 5a** above, to generate the Share link.
   (b) A new button "Disable Access to History Link" will also appear. This offers control to disable History's link so that it is not accessible when finished sharing.
   (c) The share link may be customized by clicking on the pencil icon at the far right end of the link and modifying the displayed text, if desired.
   (d) Copy the link and paste it into an email, text message, document, or any other means desired to communicate the location of the shared History.
   (e) Any recipient of the Share link may enter it into a web browser (as described in Subheadings 2 and 2.1) in order to access the History. An account is not required.
   (f) Unshare the History by clicking on the button "Disable Access to History Link."
9. Sharing and Publishing the History with all users, publicly into "Shared Data, Published Histories." This action also generates

the Share link from **step 8**. How to unshare and unpublish are included.

(a) Click on the button from Option B from **step 5b** above, to generate the Share link and Publish the History to "Published Histories."

(b) A new button "Disable Access to History via Link and Unpublish" disables this History's link so that it is not accessible and removes the History from Galaxy's Published Histories section so that it is not publicly listed or searchable.

(c) Actions for the Share links are the same as described in the above sections, **step 8c–e**.

(d) Click on "Shared Data" in the upper masthead to open the pull-down menu, and select "Published Histories."

(e) Locate your Published History by searching by keyword or publication data.

(f) Click on the button "Disable Access to History via Link and Unpublish" at any time to unshare and unpublish your History after returning to the "Share or Publish" form for the History.

10. All four of Galaxy's core objects (Histories, Workflows, Visualizations, Pages) allow Sharing and Publishing through these exact same methods, using identical "Share or Publish" interfaces.
11. Published data appears under the masthead menu "Shared Data."
    (a) All four of these same objects in **step 10** above, when in a "Published" state, will be displayed under the masthead menu "Shared Data," in the corresponding object section.
    (b) For example, "Pages" that have been published are found under the menu "Shared Data" option "Published Pages."
12. The introduction (above) for this subsection includes links to wiki and video demonstrations of Share and Publish operations in detail.

3.7 Installing a Local Galaxy Instance

You only need to download and install a local Galaxy if you plan to (1) develop it further, (2) add new tools, (3) plug in new data sources, or (4) run a local production server for your site because you have Sensitive data (e.g., clinical) or Large Datasets or processing requirements that are too big to be processed on a public server. To obtain the latest directions on running your own Galaxy instance, go to http://getgalaxy.org within your web browser. This page will also have additional information and trouble-shooting tips:

1. Open a command-line prompt (i.e., a terminal or shell).
2. Confirm that you have compatible version of Python installed (Subheadings 2.6 or 2.7) by typing "python --version" followed by the return key.

3. Confirm that you have Mercurial installed by typing "hg–version" followed by the return key. *see* **Note 10**, if you do not have Mercurial installed.
4. Download the Galaxy source code by typing "hg clone https://bitbucket.org/galaxy/galaxy-dist" followed by the return key.
5. Change your current working directory into the freshly created Galaxy root by typing "cd galaxy-dist," followed by the return key.
6. Update your Galaxy source code to the stable release branch by typing "hg update stable," followed by the return key. In the future, you can update your local Galaxy instance to the latest Galaxy version by entering "hg pull -u."
7. To start your Galaxy server, type "sh run.sh" followed by the return key. The first time that you start your Galaxy instance, it will download additional required dependencies (known as Python eggs) and automatically create local copies of several configuration files.
8. Once your Galaxy instance has started, load http://localhost:8080 within your web browser.
9. To stop the Galaxy server, use "Ctrl-C" from within the shell.
10. To add yourself as an administrator to your Galaxy instance, open the file "universe_wsgi.ini" and add your email address to the admin_users variable, for example, "admin_users = you@example.com."
11. Restart the Galaxy server to make the new Admin user active.
12. Stay current with release updates by following *Distribution News Briefs* (http://wiki.galaxyproject.org/DevNewsBriefs) and consider subscribing to the Galaxy-Dev mailing list for Galaxy community and team support (http://wiki.galaxyproject.org/MailingLists).

3.8 Running Galaxy in the Cloud

A third option for accessing Galaxy is to utilize cloud computing resources. Currently, to use Galaxy on commercial cloud resources, you will need to have an Amazon Web Services (AWS) account. A complete set of up-to-date instructions are also available at http://wiki.galaxyproject.org/CloudMan:

1. Register an account with AWS by going to http://aws.amazon.com.
2. Create a new IAM user via the AWS console.
3. Make note of the created Access Key ID and Secret Access Key.
4. Load http://usegalaxy.org/cloudlaunch in your web browser.
5. Enter your Access Key ID and your Secret Access Key into the boxes on this page.
6. Provide a name for your Galaxy cluster; this can be any value. The name should be unique for a particular AWS user, as it is

possible to have multiple CloudMan instances running at a single time and the name acts as an identifying key while running and resuming a cluster.

7. Set a password for your new Galaxy cluster. This password is the CloudMan console password and is only used to restrict access to the CloudMan administration interface.
8. Click Submit to launch your Galaxy CloudMan-managed cloud instance.
9. After a few minutes, the private Galaxy cloud instance will start and can be accessed from the provided link, with the form of "While it may take a few moments to boot, you will be able to access the cloud control panel at ec2-75-101-202-210.compute-1.amazonaws.com/cloud."
10. At the authentication prompt, enter the password that was specified in **step 7**. The username field can be left blank.
11. Since we are starting a new Galaxy cluster, you can accept the default settings and click "Choose platform type." It is also possible to specify a Dataset storage size different from the default (10 GB) depending upon the needs of the analysis; this size can later be changed via the CloudMan interface, but the Galaxy instance will be inaccessible while it is resized.
12. Within the CloudMan Console, the cloud cluster can be terminated, additional worker nodes can be added, and a link to Galaxy can be accessed. Additionally, the CloudMan Admin panel can be accessed by clicking on "Admin" at the top right of the masthead.
13. Add a new Galaxy Admin user by entering your email address into the "Set Galaxy admin users" textbox in the CloudMan Admin panel.
14. Click "Access Galaxy" to load the Galaxy Analysis interface.
15. You can now register a new Galaxy user, including the one that corresponds to the Admin user that was just created.
16. When you are finished, be sure to terminate the Galaxy cluster from the CloudMan interface, or else you will continue to be charged for usage. When you are completely finished with the cluster, be sure to check the box next to "Also delete this cluster"; this will delete the EBS volumes and S3 buckets that have been created for use in the cluster and allowed the cluster to be persisted without requiring compute nodes to be constantly running.
17. Consider subscribing to the Galaxy-Dev mailing list for Galaxy community and team support (http://wiki.galaxyproject.org/MailingLists).

3.9 Installing New Tools via the Galaxy ToolShed

The Galaxy ToolShed (http://usegalaxy.org/toolshed) enables sharing of Galaxy tools across the Galaxy community.

It is a software distribution hub for biomedical software that supports versioning, dependency, and datatype management, as well as Workflow and data integration. The ToolShed supports a wide array of tool types allowing nearly any software utility (written in any programming language), ranging from simple scripts written in interpreted languages such as Python to complex software packages distributed as source code that require compilation and installation as well as external dependencies, to be automatically installed into a Galaxy instance with a few clicks.

Here, we will install an example tool, the FreeBayes variant detector. Video examples of ToolShed tool installations into a CloudMan Galaxy can be viewed at http://vimeo.com/channels/galaxytoolshed:

1. Log in to a Galaxy instance where you are an administrator (*see* Subheadings 3.7 and 3.8).
2. Access the Administrator interface by clicking the "Admin" link in the masthead.
3. In the left-hand pane, click "Search and browse tool sheds." The ToolShed selection screen will appear in the main pane.
4. Click on the button labeled "Galaxy main tool shed." The primary public Galaxy ToolShed will load in the pane.
5. In the search box at the top left of the page, search for "freebayes." As the ToolShed is a community resource, several different versions of the FreeBayes tool may be available.
6. Choose the FreeBayes tool repository created by the Galaxy development team by clicking on the button "freebayes" that has the owner listed as "devteam." If a pop-up appears, click "Preview and Install."
7. In the top left-hand corner, click "Install to Galaxy."
8. Take note of the warning at the top of the Page, indicating that the ToolShed is a public resource where community members can add code and as such not all of this external code has been verified by the Galaxy development team nor the community-based team tasked with approving community contributions, known as the Intergalactic Utilities Commission (IUC). We are installing a tool added by the official Galaxy development team ("devteam"), so we can be rather sure of it being non-malicious.
9. Be sure that "Handle tool dependencies" is checked. This will download and install a local copy of the versioned FreeBayes binaries for Galaxy to run.

10. Select a tool section to install the new tool. In this case, we will create a new section, "NGS: Variant Detection," by entering that text into the box labeled "Add new tool panel section."
11. Click "Install." Galaxy will download and install the new tool. The installation progress can be monitored in real time as it advances through the various stages (e.g., new, downloading, installing dependencies, installed).
12. Switch back to the Galaxy Analysis interface and confirm that the tool has been installed into the new tool section.

4 Notes

1. The FASTQ format has become the de facto standard format for the representation of sequencing reads [12]. There are several different FASTQ variants in use, with the Sanger variant being the preferred form. In the fastqsanger format, the quality scores are Phred-scaled and encoded using ASCII characters, with one character used per base. The score value is equal to the ordinal value of the ASCII character subtracted by 33.
2. The Galaxy API provides a programmatic interface to communicate with a Galaxy instance when directly interacting with the web-based GUI is not practical or desired. For example, external programs that want to access features of Galaxy can make use of the RESTful API.
3. It is important to use meaningful names for Histories. History names serve as a straightforward method of keeping multiple Histories organized. It is a good idea to use a separate History for each analysis or logical portion of each analysis. Access all account Histories using the "History List" menu option "Saved Histories."
4. Galaxy's datatypes utilize a hierarchical system where more restricted datatypes are children of less defined types. This allows tools to function on specific data formats or on types of data. For example, a tool which utilizes the generic "fastq" datatype will accept "fastqsanger" and "fastqillumina" as input. However, a tool which accepts "fastq" for input will not accept generic "text" Datasets as input, despite "fastq" being a child of "text."
5. Tools can sometimes be difficult to locate. You can use the Tool Search to search for tools based upon name and key words. Enter a word or phrase into the "search tools" text box at the top of the Tool pane on the left-hand side.
6. When uploading a file, it is possible to manually declare a datatype instead of utilizing the auto-detect functionality. This can be helpful to select a specific sub-datatype during upload;

for example, all variants of the FASTQ format are detected as the base "fastq" class; however, selecting the specific FASTQ variant (e.g., "fastqsanger") during upload will allow a user to bypass the grooming step if they are sure of the correctness and encoding type of their uploaded FASTQ file.

7. It is possible to change the declared datatype of a Dataset after it exists in a History by clicking the pencil icon and selecting a new datatype from the drop-down list. This will not alter the actual Dataset content, just the way that Galaxy will interpret the Dataset. To convert Dataset content, select the "Convert Format" tab or search for a stand-alone tool in the Tool pane that will convert the Dataset to different formats.
8. There are several different variations to the FASTQ format, depending upon the value of and the encoding method used for base quality scores. The datatype "fastqsanger" is the preferred variant datatype to use for most tools. The FASTQ Grooming tool can be used to convert between the different variants. Color-space reads will use the "fastqcssanger" datatype.
9. Workflows can be configured to hide intermediate analysis Datasets; however, all Datasets are still present in the result History. The "History list" options that make hidden datasets accessible are (a) "Include Hidden Datasets" to temporarily unhide and rehide or individually permanently unhide and (b) "Unhide Hidden Datasets" to permanently unhide all hidden Datasets at once. Hidden Datasets are created by Workflows only and a Dataset that has permanently been marked as unhidden cannot return to hidden status.
10. Mercurial is a source-code revision control system. It can often be installed by the packaging system used by your operating system; for example, in Ubuntu, typing "sudo apt-get install mercurial" into a shell will perform the installation. If it is not possible to use Mercurial to obtain the Galaxy source code, a downloadable archive is also available from https://bitbucket.org/galaxy/galaxy-dist. Using Mercurial is the preferred method as it will greatly simplify updating Galaxy to the newest versions.

Acknowledgments

Efforts of the Galaxy team (E. Afgan, D. Baker, D.B., D. Bouvier, M. Cech, D. Clements, N. Coraor, C. Eberhard, D. Francheteau, J. Goecks, S. Guerler, J.J., G. Von Kuster, R. Lazarus, Anton Nekrutenko, and James Taylor) were instrumental in making this work possible. We extend a special thank you to the Galaxy community for their continuing contributions, both inspirational and technical.

References

1. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455. doi:10.1101/gr.4086505, gr.4086505 [pii]
2. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. Chapter 19: Unit 19 10 11-21. doi: 10.1002/0471142727.mb1910s89
3. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86. doi:10.1186/gb-2010-11-8-r86, gb-2010-11-8-r86 [pii]
4. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011:bar049. doi:10.1093/database/bar049 bar049 [pii]
5. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32(Database issue):D493–D496. doi:10.1093/nar/gkh103 32/suppl_1/D493 [pii]
6. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. doi:10.1186/gb-2013-14-4-r36, gb-2013-14-4-r36 [pii]
7. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515. doi:10.1038/nbt.1621 nbt.1621 [pii]
8. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. doi:10.1093/bioinformatics/btp324 btp324 [pii]
9. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. doi:10.1186/gb-2009-10-3-r25 gb-2009-10-3-r25 [pii]
10. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9(9):R137. doi:10.1186/gb-2008-9-9-r137 gb-2008-9-9-r137 [pii]
11. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006. doi:10.1101/gr.229102
12. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38(6):1767–1771. doi:10.1093/nar/gkp1137 gkp1137 [pii]

Chapter 3

edgeR for Differential RNA-seq and ChIP-seq Analysis: An Application to Stem Cell Biology

Olga Nikolayeva and Mark D. Robinson

Abstract

The edgeR package, an R-based tool within the Bioconductor project, offers a flexible statistical framework for detection of changes in abundance based on counts. In this chapter, we illustrate the use of edgeR on a human embryonic stem cell dataset, in particular for RNA-seq and ChIP-seq data. We focus on a step-by-step statistical analysis of differential expression, going from raw data to a list of putative differentially expressed genes and give examples of integrative analysis using the ChIP-seq data. We emphasize data quality spot checks and the use of positive controls throughout the process and give practical recommendations for reproducible research.

Key words Differential count analysis, edgeR, RNA-seq, ChIP-seq, Reproducible research, Integrative analysis, Human embryonic stem cells

1 Introduction

With the advent of large-scale yet low-cost sequencing of short DNA fragments, RNA sequencing (RNA-seq), or more precisely cDNA sequencing, has become an important and popular tool for studying transcriptomes. In a single experiment, researchers can explore gene expression levels, alternative isoforms, RNA editing, and allele-specific expression. Many variants of RNA-seq protocols exist, but the general strategy is to start with a pool of RNA molecules from cells of interest (e.g., embryonic stem cells). Generally, replicate pools are collected for each experimental condition of interest and conditions are compared (e.g., pluripotent versus differentiated cells). Typically, the protein-coding messenger RNA subpopulation is selected for (e.g., RNA with a poly-A tail) or ribosomal RNA is depleted. Specific protocols exist to capture other subclasses of RNA, such as microRNAs. RNA is then reversed transcribed into cDNA, fragmented into a suitable size range (e.g., between 200 and 800 base pairs), and sequenced in single- or paired-end mode.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_3, © Springer Science+Business Media New York 2014

Depending on the biological question, many bioinformatics tools are readily available for several tasks, such as mapping reads to a reference genome (and/or transcriptome), assembling transcripts, and quantifying (changes in) expression. Our focus in this chapter is on the statistical analysis of differential expression, including a full pipeline from raw data to a list of candidate differentially expressed genes. In particular, we perform a reanalysis of the identification of "signature" genes from a recent publicly available RNA-seq dataset describing differentiation of embryonic stem cells to polyhormonal and functional endocrine cells; our analysis goal is to find signature genes (i.e., significantly higher expressed in a cell type compared to other cell types). Indirectly, we are able to show how very general differential analyses can be conducted, on other types of data than RNA-seq. While the primary focus is on RNA-seq data and an R/Bioconductor tool called *edgeR* for differential gene expression ("empirical analysis of digital gene expression in R"), we also demonstrate some integrative analysis with chromatin immunoprecipitation sequencing (ChIP-seq) of histone modifications on the corresponding stem cell subpopulations.

The strategy adopted by *edgeR* (and similar packages) is, using a transcript catalog, to simply count reads that map to features of interest (i.e., genes), without regard to the multiple isoforms that may be expressed [1]. The starting point is a count table, representing the read counts for each gene across multiple samples. Gene counts are then modeled using sophisticated statistical methods to arrive at a set of differentially expressed features. This pipeline is a simplification with respect to the biology of alternative isoforms, but the framework is quite general and can allow arbitrary comparisons of experimental conditions, with a facility to easily account for batch effects, blocking factors, time course, pairing of samples, etc. This is a notable advantage of *edgeR*-style analyses, whereas many existing tools can only perform pairwise comparisons. In addition, differential splicing analyses can be conducted within the same count framework in an additional step (not shown here), by counting the reads with exons as features [2].

The statistical methods underpinning differential analyses in *edgeR* use the negative binomial (NB) model as the natural extension to Poisson sampling. The early RNA-seq papers have shown, in the absence of the lane and batch effects, that technical variability (i.e., re-sequencing the same cDNA population at different times) is well described by the binomial distribution, which is well approximated by the Poisson distribution [3]. The NB model can account for biological variability (i.e., sequencing cDNA from two or more biological replicates) by imposing a mean-variance relationship that is fit well by RNA-seq data [4]. As with the analysis of microarray data, one of the critical aspects of a differential expression analysis is to constrain the estimation of the variance

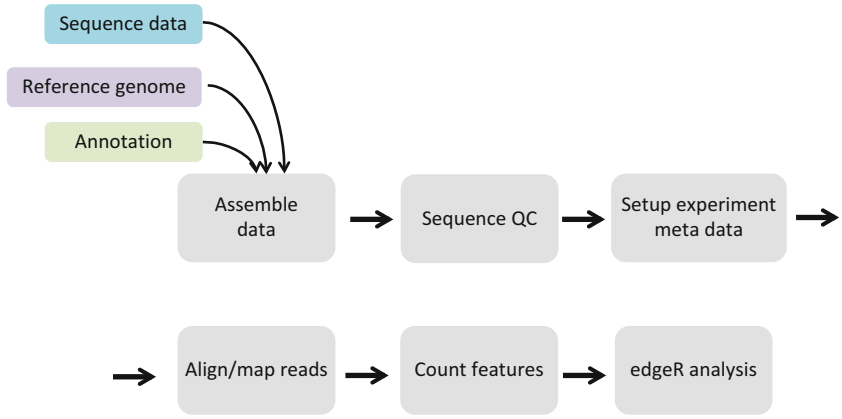


Fig. 1 Entire process overview: main steps in going from raw sequence and annotation data to *edgeR* analysis

parameter; in small samples, it is difficult to get accurate variability estimates, and sharing of information across all features measured is of vital importance. Similarly, in RNA-seq count data, it remains critical to get accurate estimates of the dispersion parameter. Many variations now exist in the literature (see Anders et al. for a recent collection of references [1]) for moderating dispersion; in particular, *edgeR* uses a weighted likelihood strategy to shrink toward a local dispersion-mean trend [5, 4].

To illustrate the pipeline, this chapter gives a start-to-finish edgeR analysis for finding cell-type-specific "signature" genes, with additional integrative analyses between the RNA-seq and ChIP-seq data on the same cell types to show the versatility of the R/Bioconductor environment [6]. We make a concerted effort to make such analyses thorough, including many spot checks of data quality and positive controls, where possible. In addition, we come from a standpoint of reproducible research, in terms of collecting R commands and output in an executable document (e.g., Sweave) and keeping track of the software versions used, so that researchers can reproduce the data analyses. As we proceed, we give some useful tips on how to organize such an analysis and how to semiautomate some tasks, such as batch data downloading and writing Unix commands. Note that this chapter is not a replacement of existing resources, such as the comprehensive user's guide (http://www.bioconductor.org/packages/release/bioc/html/edgeR.html).

Figure 1 shows the main steps of the Methods described below. One starts with collecting various types of relevant data, such as the raw sequence data (e.g, for RNA-seq or ChIP-seq experiments), as well as the reference genome and annotation information. After the raw and reference data have been located, a quality control step takes place to ensure that the raw data is suitable for downstream

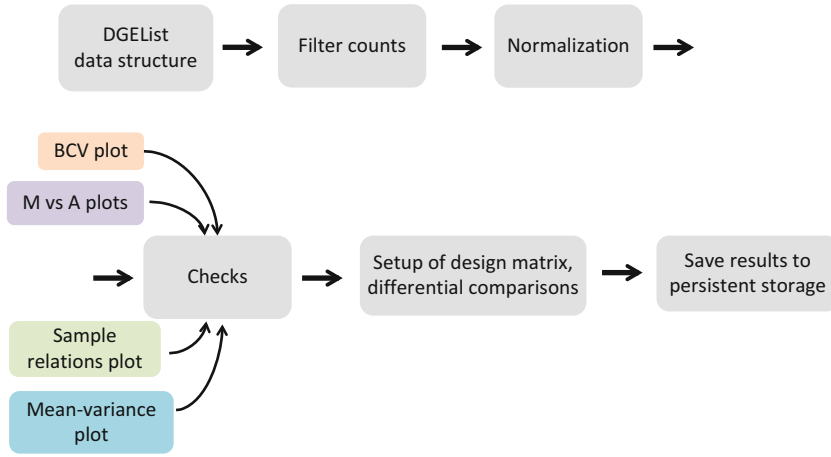


Fig. 2 *edgeR* process overview: main parts of a typical analysis

processing. In order to simplify the subsequent steps, a metadata table, containing relevant characteristics of the samples and the experimental design, is created semiautomatically. This step is followed by using an algorithm to align and map the sequencing reads to a reference genome. After the reads have been mapped, they are counted according to catalog of features of interest; the resulting counts serve as input to the edgeR analysis.

Figure 2 highlights the main steps of a typical *edgeR* analysis. First, a `DGEList` object is created and contains the feature counts as well as the information about which group the analyzed samples belong to. As the *edgeR* analysis progresses, additional elements are added to the `DGEList` object (e.g., dispersion estimates). The next step involves filtering counts. For example, it is advisable to remove low-abundance features using a reasonable counts-per-million threshold in at least some number of samples. During the normalization step, counts are normalized by the effective library sizes.

Prior to proceeding to formal statistical analyses, spot checks are performed to confirm that replicates bear resemblance to each other and the expression values are consistent with existing knowledge, if such information is available. In particular, an MDS plot is often informative to highlight consistency of replicates, potential batch effects or poor quality samples, and often offers a glimpse into the underlying biology. Various additional plots, such as the BCV (biological coefficient of variation), M (log-ratio) vs. A (expression strength), and mean–variance are used to explore other features of the analysis. The design matrix definition is a critical step that allows analysts to customize differential comparisons of

interest and highlights the flexibility of generalized linear models (GLMs) in such analyses. After the analyses are performed, the results (typically tables of statistics) are stored on disk for future retrieval.

2 Materials

1. **Operating System**

   We describe how to implement the methods below on a Unix-like operating system, such as Mac OS X or Linux, with a bash shell or a similar command-line environment. The commands described below should be either run in a terminal shell (shown in black fixed width font) or within *R* [7] (shown in fixed width font, with results shown in bold fixed width font), as appropriate. The original code is available from the authors.

2. **Software**

   For implementation, the following software is needed:

   - aligner—we illustrate the use of *tophat2* [8]. Other possibilities include *GSNAP* [9], *MapSplice* [10], and *Subread* [11]
   - alignment visualization tool—we use *Savant* [12, 13]. Another popular alternative is the Integrated Genome Viewer, *IGV* [14]
   - *R* statistical computing environment, downloadable from http://www.r-project.org/ [15]
   - *Bioconductor* packages, including `ShortRead` [16], `edgeR` [17, 4], `GenomicFeatures` [18], `rtracklayer` [19]. All of the packages used in the Methods are explicitly mentioned in the text
   - *samtools* [20]

   As most of these software packages are being regularly updated, make sure to use the most recent stable version as well as read the accompanying documentation prior to use, because some settings and recommendations may change over time. *See* Subheading 3.5 below for a description of the versions used to create this document.

3. **Software Installation**

   Download and install the following tools:

   - *samtools* from http://samtools.sourceforge.net
   - *tophat2* from http://tophat.cbcb.umd.edu and *bowtie2* from http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
   - the latest release version of *R* from http://cran.r-project.org

To download and install *Bioconductor* packages, open a terminal window and start *R* by typing `R`. After *R* starts, type the following commands:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("BiocUpgrade")
> biocLite(c("ShortRead","edgeR","rtracklayer","GenomicFeatures",
+            "QuasR","Rsamtools","limma","biomaRt","Repitools"))
```

biocLite is an automatic installation tool that ensures that the installed packages match versions with other packages, as well as the *R* version. All *Bioconductor* packages used in the following analysis are mentioned in the text. To install any additional packages, use the `biocLite("<package name>")` command.

4. **Sample Data**
We will use the data sets from a recent article in Cell [21] that describes differentiation of human embryonic stem cells (hESCs) into various pancreatic lineages and includes multiple RNA-seq and ChIP-seq datasets. In particular, these data correspond to various stages, including definitive endoderm (DE), primitive gut tube cells (GT), posterior foregut (FG), pancreatic endoderm (PE), as well as the in vitro differentiated polyhormonal cells (PH) and the in vivo differentiated functional endocrine (FE) cells.

   The RNA-seq data include, considering replicates, 27 total samples, each collected by sequencing 100bp of strand-specific complementary DNA (*see* **Note 1**). The ChIP-seq data comprise 24 samples.

3 Methods

3.1 Getting Started

1. Create metadata table
The dataset analyzed here includes both the RNA-seq and ChIP-seq experiments, the raw data of which come in a standardized (compressed) FASTQ format. The datasets are available on the ArrayExpress website, under the accession number E-MTAB-1086 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1086/). It is useful to start by compiling a metadata table, which summarizes the relevant characteristics (in terms of downstream statistical modeling) of the available samples (*see* **Note 2**). In general, this step needs to be customized for every analysis, according to the samples available and the goals of the experiment.

   Extract information about the samples (one row in the table for each) from the `E-MTAB-1086.sdrf.txt` file, by reading it into *R*:

```
> u <- url("http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1086/E-MTAB-1086.sdrf.txt")
> md <- read.table(u,sep="\t", header=TRUE, comment.char="", stringsAsFactors=FALSE)
```

Create a table called `samps` (for samples), by selecting important columns of original table from ArrayExpress, `md` (*see* **Note 3**), and assign shorter names:

```
> samps <- md[,c("Comment.LIBRARY_STRATEGY.","Comment.ENA_RUN.",
+                "Comment.FASTQ_URI.","FactorValue..histone.antibody.",
+                "FactorValue..CELL_TYPE.","FactorValue..CELL.PROPERTY.")]
> colnames(samps) <- c("lib.type","ENAid","fastq","antibody",
+                      "long.cell.type","cell.prop")
```

To be able to classify the samples by their cell type corresponding to their differentiation state, two-letter cell type abbreviations are assigned based on the `cell.prop` column. In this case, based on the format that the metadata arrives in, some amount of custom manual (though reproducible) processing can be applied to prepare for the downstream statistical analysis.

```
> un <- unique(samps$cell.prop)
> labels <- c("DE","n/a","GT","FG","PE","FE","PH","PE")
> names(labels) <- un
> labels
```

```
      embryonic stem cell directed to definitive endoderm
                                                      "DE"
                                                       n/a
                                                     "n/a"
      embryonic stem cell directed to primitive gut tube
                                                      "GT"
       embryonic stem cell directed to posterior foregut
                                                      "FG"
     embryonic stem cell directed to pancreatic endoderm
                                                      "PE"
                          in vivo-matured endocrine cells
                                                      "FE"
      CD200+ polyhormonal cells from embryonic stem cell
                                                      "PH"
CD142+ late pancreatic endoderm from embryonic stem cell
                                                      "PE"
```

Both embryonic stem cells (ES) and islet cells (IS) are denoted as `NA` in the `cell.prop` column of `samples`. Therefore, the `long.cell.type` column is used to determine whether a particular sample is of the ES or IS type.

```
> samps$cell.type <- labels[samps$cell.prop]
> samps$cell.type[samps$long.cell.type=="islet cell"]<- "IS"
> samps$cell.type[samps$cell.type=="n/a"] <- "ES"
> samps$cell.type <- factor(samps$cell.type,
+                           levels=c("ES","DE","GT","FG","PE","PH","FE","IS"))
```

The short names of samples are defined by concatenating the two-letter cell type and the last four digits of the ENA number, and the `short.name` column is then added to the `samps` table.

```
> samps$short.name <- paste0(samps$cell.type,".",gsub("ERR2","",samps$ENAid))
```

Use the `head` command to view a part of the `samps` matrix:

```
> head(samps,2)
```

```
  lib.type     ENAid
1 ChIP-Seq ERR208014
2 ChIP-Seq ERR208022
                                                                      fastq
1 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR208/ERR208014/ERR208014.fastq.gz
2 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR208/ERR208022/ERR208022.fastq.gz
                                        antibody              long.cell.type
1 H3K4me3 Millipore 04-745 Lot#: NG1643014 embryonic stem cell
2 H3K27me3 Millipore 07-449 Lot#: DAM1588246 embryonic stem cell
                                             cell.prop cell.type  short.name
1 embryonic stem cell directed to definitive endoderm          DE    DE.08014
2 embryonic stem cell directed to definitive endoderm          DE    DE.08022
```

Subset the `samps` table by using the `lib.type` column to separate the RNA-seq files from the ChIP-seq ones:

```
> sampsR <- samps[samps$lib.type=="RNA-Seq",]
> sampsC <- samps[samps$lib.type=="ChIP-Seq",]
```

In order to tabulate the samples available, use the `table` and `with` functions. For example, create a summary table of how many RNA-seq and ChIP-seq samples are available for each cell type as follows:

```
> with(samps, table(cell.type,lib.type))
```

```
         lib.type
cell.type ChIP-Seq RNA-Seq
       ES        3       3
       DE        3       3
       GT        3       3
       FG        3       3
       PE        7       6
       PH        2       3
       FE        3       4
       IS        0       2
```

Create a summary table of the ChIP-seq data by antibody type:

```
> sampsC$antibody <- sapply(sampsC$antibody,
+                           function(u) strsplit(u," ")[[1]][1],
+                           USE.NAMES=FALSE)
> with(sampsC, table(cell.type,antibody))
```

```
         antibody
cell.type H3K27me3 H3K4me3 input
       ES        1       1     1
       DE        1       1     1
       GT        1       1     1
       FG        1       1     1
       PE        2       2     3
       PH        1       1     0
       FE        1       1     1
       IS        0       0     0
```

In order to view other combinations of columns of the `samps` object, simply adjust the arguments of the `table` function accordingly.

2. Download reference genome and gene model annotation
   Download the reference genome and gene model annotation from a public resource, such as Ensembl (*see* **Note 4**), by running the following commands in Unix:

```
wget ftp://ftp.ensembl.org/pub/release-71/fasta/homo_sapiens/dna/\
Homo_sapiens.GRCh37.71.dna.toplevel.fa.gz

wget ftp://ftp.ensembl.org/pub/release-71/gtf/homo_sapiens/\
Homo_sapiens.GRCh37.71.gtf.gz
```

3. Build reference index
   Prior to aligning the reads, the reference genome needs to be converted into an aligner-specific index, in order to speed up the alignment process. An index file only needs to be generated once for a given reference genome and aligner.
   Create a *bowtie2*-specific index in Unix:

```
bowtie2-build -f ensembl_Homo_sapiens.GRCh37.71.fa ensembl71_Hs_GRCh37
```

   The output of this command will be a set of six BT2 files, with names starting with `ensembl71_Hs_GRCh37`, as defined above (*see* **Note 5**).

4. Organize files and directories
   Prior to downloading the sample data, we create separate directories (*see* **Note 6**) for the RNA-seq and ChIP-seq raw and mapped data files by using the `mkdir` command in Unix:

```
mkdir annotation
mkdir rna_seq_data
mkdir chip_seq_data
mkdir rna_seq_tophat
mkdir chip_seq_mapped
```

   The annotation directory will contain all the relevant annotation files for alignment and counting.

   We proceed below in the *root* directory (*see* **Note 7**), with the assumption that the subdirectories defined above are already created.

5. Download sample files
   To download the RNA-seq sample files, first set the working directory to `rna_seq_data`. Then use the following commands to extract the basenames of files and automatically download them.

```
> bn <- basename(sampsR$fastq)
> for(i in 1:length(sampsR$fastq))
+   download.file(sampsR$fastq[i], bn[i])
> all(file.exists(bn))
```

The `all` and `file.exists` commands are used to check that all files have been downloaded successfully.

To download the ChIP-seq data, first change the working directory to `chip_seq_data` and then execute analogous commands.

```
> bnchip <- basename(sampsC$fastq)
> for(i in 1:length(sampsC$fastq))
+   download.file(sampsC$fastq[i], bnchip[i])
> all(file.exists(bnchip))
```

3.2 Reads to Alignments: FASTQs to BAMs

3.2.1 Processing RNA-seq Data

1. Quality control with ShortRead

   The `ShortRead` package in *R* provides tools to evaluate the quality of the raw sequences (*see* **Note 8**). The `qa` and `report` functions perform quality assessment summaries and generate a report. The `dirPath` argument of the `qa` function specifies the path to the directory containing the FASTQ files to be evaluated.

   Generate an HTML quality report:

```
> library("ShortRead")
> fqQC <- qa(dirPath="rna_seq_data", pattern=".fastq", type="fastq")
> report(fqQC, type="html", dest="rnaseq_QAreport")
```

   Open the report in a web browser to inspect the resulting quality assessment.

   The per-cycle quality figure (Fig. 3) is one of the summaries that is automatically generated by the `qa` function. Here one sees a summary of quality across all of the RNA-seq samples—fortunately, each one of the 27 samples has high quality reads across most of the base-calling cycles. The slight score decrease is typical for later cycles (toward the end of reads).

2. Read alignment with *tophat2*

   This step involves converting the raw sequence data (in FASTQ format) to alignments onto the reference sequence (in BAM format) (*see* **Note 9**). The RNA-seq (cDNA) reads are expected to map with gaps, at least when a cDNA fragment overlaps an exon–exon junction, while ChIP-seq (genomic DNA) should map without gaps and are covered separately.

   First, the RNA-seq reads are aligned to the reference genome using *tophat2* tool. The commands necessary to run *tophat2* are generated in *R*. Assuming the directory structure is setup as mentioned above, all commands from *R* can be run with relative directory paths from the root directory:

```
> gz <- dir("rna_seq_data", "fastq.gz", full=TRUE)
> nm <- gsub(".fastq.gz","",basename(gz))
> unlink("tophat_commands_RNA")
> for(i in 1:length(gz)) {
+   cat("nice -n 19 tophat2 -G annotation/Homo_sapiens.GRCh37.71.gtf \\\n",
+       "       -p 10 -o rna_seq_tophat/",nm[i]," annotation/ensembl71_Hs_GRCh37 \\\n",
+       "       ", gz[i],"\n\n", sep="", file="tophat_commands_RNA", append=TRUE)
+ }
```

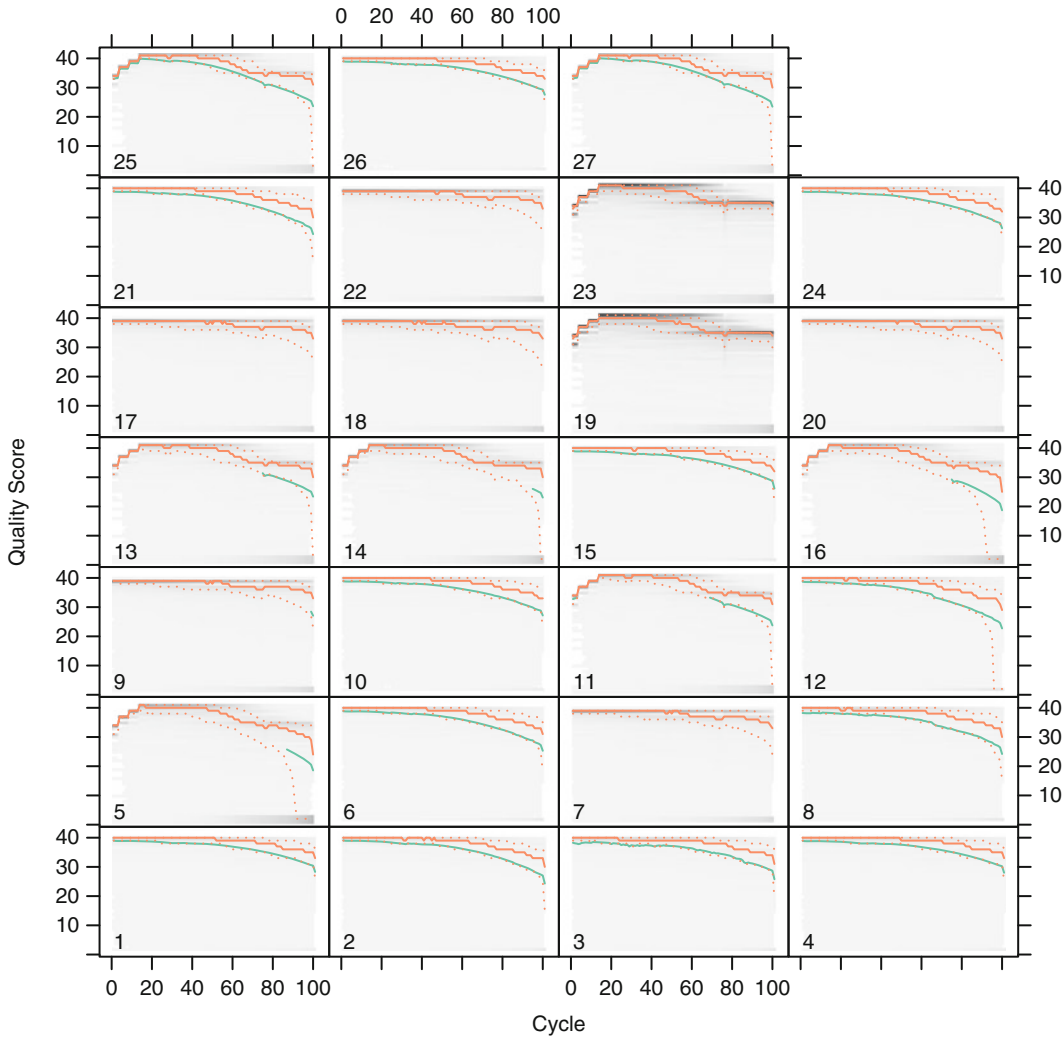


Fig. 3 perCycleQuality figure automatically generated by the ShortRead package for the RNA-seq samples. Each of the 27 panels corresponds to one sample and shows the quality scores across the read length

These will generate a text file, called `tophat_commands_RNA`, which contains the necessary *tophat2* commands. For example, one can look at the first few lines of this file from R:

```
> cat(system("head -n 7 tophat_commands_RNA",intern=TRUE),sep="\n")
nice -n 19 tophat2 -G annotation/Homo_sapiens.GRCh37.71.gtf \
           -p 10 -o rna_seq_tophat/ERR207980 annotation/ensembl71_Hs_GRCh37 \
           rna_seq_data/ERR207980.fastq.gz

nice -n 19 tophat2 -G annotation/Homo_sapiens.GRCh37.71.gtf \
           -p 10 -o rna_seq_tophat/ERR207981 annotation/ensembl71_Hs_GRCh37 \
           rna_seq_data/ERR207981.fastq.gz
```

To execute the *tophat2* commands (*see* **Note 10**), the `tophat_commands_RNA` can either be sourced in Unix by using the `source` command, or the commands contained in

the file can be copy-pasted into a Unix shell directly. The zipped FASTQ.GZ files do not need to be uncompressed prior to running *tophat2*.

The time required to run *tophat2* will depend on the number of cores, the command priority value, and the number of reads to be mapped (*see* **Note 11**). When *tophat2* successfully maps all reads, each output directory will contain several files, one of which is called `accepted_hits.bam`. These BAM files will be used for downstream processing.

3. Sort and index RNA-seq BAM files with *Rsamtools*
   Prior to further processing, the `accepted_hits.bam` files need to be sorted and indexed. These tasks are accomplished by using the `Rsamtools` library:

```
> library("Rsamtools")
> bm <- dir("rna_seq_tophat","accepted_hits.bam",
+           recursive=TRUE,full=TRUE)
> obm <- paste0(gsub("/accepted_hits.bam","",bm),"_s")
> for(i in 1:length(bm)) {
+       sortBam(file=bm[i],destination=obm[i])
+       indexBam(paste0(obm[i],".bam"))
+ }
```

Two files will be produced as a result: one with the _S.BAM extension, corresponding to sorted alignments, and the other with the _S.BAM.BAI extension, corresponding to the index.

4. Inspect RNA-seq alignments in *Savant*
   A genome browser, such as *Savant*, can be used to view the resulting alignments. To do this, open *Savant*, select the relevant genome, and load the alignments (BAM) as well as the annotation (GTF). If you are loading the annotation file for the first time, it will need to be converted into one of the *Savant* native formats (in this case from GTF into TABIX).

   Figure 4 is an example of alignment visualization—samples ER266342(ES), ER266348(FE), and ER266340(PH) have been plotted in the neighborhood of gene POU5F1 (ENSG00000204531) on chromosome 6, as well as the corresponding annotation. Here one notices that there are a lot more reads mapping to the ES sample than to the other two, which is consistent with Fig. 7.

3.2.2 Processing ChIP-seq Data

1. Quality control
   Generate a quality control assessment of the raw ChIP-seq data by using the `ShortRead` package, similarly to the RNA-seq procedure above, using:

```
> library("ShortRead")
> fqQC <- qa(dirPath="chip_seq_data", pattern=".fastq", type="fastq")
> report(fqQC, type="html", dest="fastQAreport_chip")
```

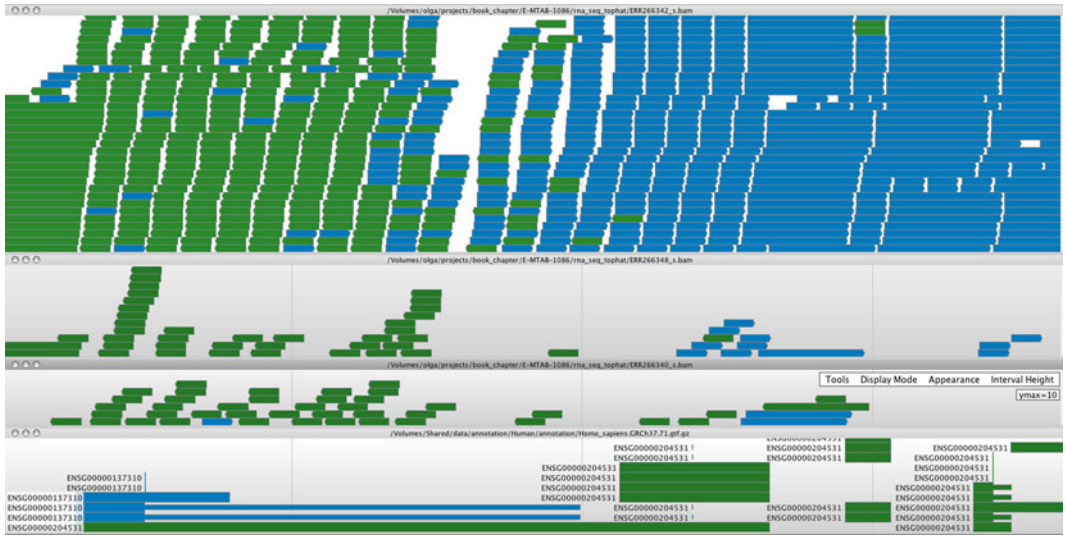


Fig. 4 Alignment visualization using the *Savant* browser. Reads mapping to samples ER266342(ES), ER266348(FE) and ER266340(PH) are shown in the *upper three panels*. The *bottom panel* shows annotation information in the region of POU5F1 on chromosome 6. *Blue* corresponds to positive strand, *green* to negative strand

Inspect the quality assessment report to make sure that the sample data can be used for further analysis.

Figure 5 shows that all of the 24 ChIP-seq samples have consistently high quality score across the base-calling cycles.

2. ChIP-seq read alignment with *bowtie2*
Since ChIP-seq reads originate from genomic DNA, we can map them to the reference genome without considering gaps. Thus, *bowtie2* can be used directly. First, create the commands in R, similarly to the *tophat2* commands for RNA-seq data above:

```
> gz <- dir("chip_seq_data", "fastq.gz", full=TRUE)
> nm <- gsub(".fastq.gz","",basename(gz))
> unlink("bowtie2_commands_chip")
> for(i in 1:length(gz)) {
+  cat("nice -n 19 bowtie2 -p 10 -x annotation/ensembl71_Hs_GRCh37 \\\n",
+      "          -U ",gz[i]," -S ", nm[i],"/",nm[i],".sam\n\n", sep="",
+      file="bowtie2_commands_chip",append=TRUE)
+ }
```

Use the `cat` and `system` commands to view the first two *bowtie2* commands (*see* **Note 12**):

```
> cat(system("head -n 5 bowtie2_commands_chip",intern=TRUE),sep="\n")
nice -n 19 bowtie2 -p 10 -x annotation/ensembl71_Hs_GRCh37 \
          -U chip_seq_data/ERR207979.fastq.gz -S ERR207979/ERR207979.sam

nice -n 19 bowtie2 -p 10 -x annotation/ensembl71_Hs_GRCh37 \
          -U chip_seq_data/ERR207984.fastq.gz -S ERR207984/ERR207984.sam
```

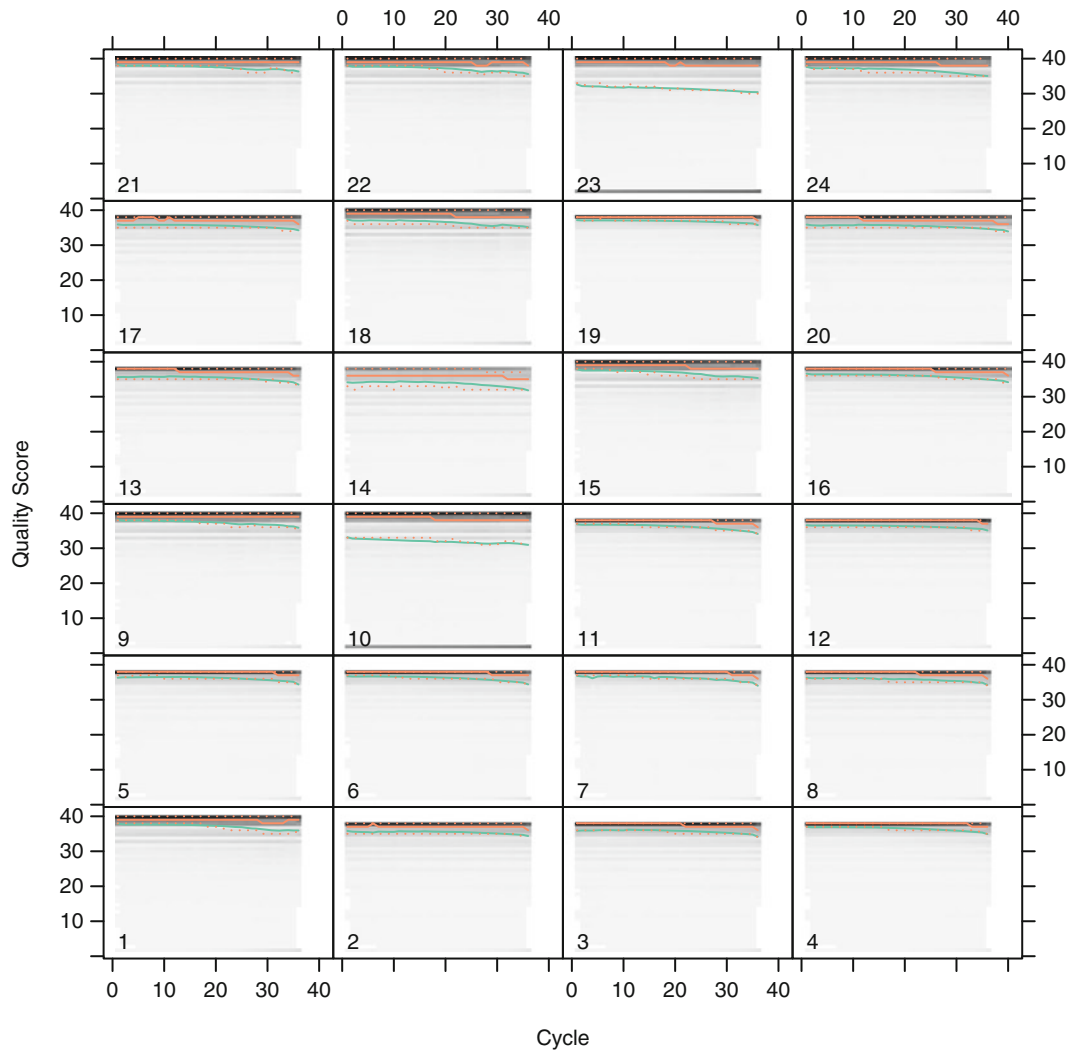


Fig. 5 perCycleQuality figure automatically generated by the *ShortRead* package for the ChIP-seq samples

3. Organize, sort, index SAM/ BAM files
   The output of *bowtie2* alignment is a SAM file, but `QuasR` requires BAM format as input. We use *samtools* to convert one file type into another and to sort and index the resulting BAM file.

   Set the working directory to `chip_seq_data` and generate the *samtools* commands in `R`:

```
> unlink("samtools_commands_chip")
> for(i in 1:length(gz)) {
+  cat("samtools view -bS chip_seq_data/",nm[i],"/",nm[i],".sam | \\\n",
+      "          samtools sort - chip_seq_data/",nm[i],"_s\n",
+      "samtools index chip_seq_data/",nm[i],"_s.bam;\n\n",sep="",
+     file="samtools_commands_chip",append=TRUE)
+ }
```

Look at the top few lines of the file right from *R* using:

```
> cat(system("head -n 7 samtools_commands_chip",intern=TRUE),sep="\n")

samtools view -bS chip_seq_data/ERR207979/ERR207979.sam | \
         samtools sort - chip_seq_data/ERR207979_s
samtools index chip_seq_data/ERR207979_s.bam;

samtools view -bS chip_seq_data/ERR207984/ERR207984.sam | \
         samtools sort - chip_seq_data/ERR207984_s
samtools index chip_seq_data/ERR207984_s.bam;
```

Here, the `view` command converts the SAM file (ASCII text) to BAM (binary), which contains the same information but results in a large reduction in the file size (*see* **Note 13**). The pipe character, `|`, directs the resulting BAM output into the `sort` command, avoiding the creation of intermediate files. Finally, the index commands produce the BAI index file, which many downstream tools need.

3.3 QuasR on the Loose: From BAMs to Counts

The `QuasR` (Quantify and Annotate Short Reads in *R*) package, [22], provides a convenient way to go from alignments contained in BAM files to read counts corresponding to a particular genomic feature (e.g., gene or transcription start site). `QuasR` (*see* **Note 14**) was chosen for this analysis as it allows to perform counting completely within *R* and it natively supports multithreading. Similar to above, this section is split into RNA-seq and ChIP-seq parts, since they differ in what features are counted.

3.3.1 Counting Features for RNA-seq Data

1. Create the `qFiles` text file, which contains the BAM files to be processed and their corresponding sample names:

```
> qFiles <- data.frame(FileName=paste0(sampsR$ENAid,"_s.bam"),
+                      SampleName=as.character(sampsR$short.name))
> sampF <- "rna_seq_tophat/qFiles.txt"
> write.table(qFiles, file=sampF, quote=FALSE, sep="\t", row.names=FALSE)
```

2. Create the `qProject` object by using the reference genome file and the samples text file as input to the `qAlign` function:

```
> library("QuasR")
> genomeF <- "annotation/ensembl_Homo_sapiens.GRCh37.71.fa"
> projR <- qAlign(sample=sampF, genome=genomeF, paired="no")
```

3. Use the `rtracklayer` library to import the GTF annotation file as a `GRanges` object:

```
> library("rtracklayer")
> annoF <- "annotation/Homo_sapiens.GRCh37.71.gtf"
> annoGR <- import(annoF, format="gtf", asRangedData=FALSE)
```

4. Filter the annotation object to include only lincRNA, protein coding, and antisense elements:

```
> chrs <- c(1:22,"X","Y")
> keep <- annoGR$gene_biotype %in% c("lincRNA","protein_coding","antisense") &
+         seqnames(annoGR) %in% chrs
> annoGR <- annoGR[keep]
```

The `keep` object contains indices of the `annoGR` object which fulfill two conditions: the `gene_biotype` property corresponds to one of the desired annotation element types and the `seqnames` of the `annoGR` object correspond to one of the chromosome names, as contained in the `chrs` object. The `keep` object is then used to subset the `annoGR` object, leaving only the annotation elements of interest.

5. Rename annotation by using the gene identifiers to label the annotation elements. Also, use the `chrs` (chromosomes) object and the `seqlevels` function to change the chromosome names of the `annoGR` object:

```
> names(annoGR) <- annoGR$gene_id
> seqlevels(annoGR) <- chrs # scrub off "extra" chromosomes
```

The `save` command is used to save an object as .RDATA for later retrieval (using the `load` command):

```
> save(annoGR,file="annoGR.Rdata")
```

6. Finally, run the `qCount` function in order to summarize the RNA-seq reads by genomic feature (*see* **Note 15**).

```
> library("parallel")
> cl <- makeCluster(10)
> countsR <- qCount(projR, annoGR, reportLevel="gene", orientation="same", clObj=cl)
> save(countsR,file="countsR.Rdata")
```

7. After the counting computation ends, use the `stopCluster` command to close the cluster, so that the CPU and memory resources allocated to this computation can be returned to the system (*see* **Note 16**).

```
> stopCluster(cl)
```

8. View the `countsR` object, which is a matrix of counts, with rows corresponding to genes and columns to sample names. The `width` column indicates the width of a particular feature, which in this case is a gene.

```
> head(countsR,2)
```

```
                  width ES.07985 DE.07981 GT.66339 FG.08004 PE.07980 FE.66350
ENSG00000254468     355        0        0        0        1        0        0
ENSG00000177951    3868       44       50       24       37       38       41
                ES.66342 DE.66333 GT.66341 FG.66346 PE.66344 FE.66331 PH.66332
ENSG00000254468        1        1        0        0        1        1        0
ENSG00000177951      162      164       78       84       77      129       43
                PH.66336 PE.66345 PE.66330 FE.66348 FE.66334 ES.66335 DE.66349
ENSG00000254468        0        0        0        0        0        0        0
ENSG00000177951       64       58       53       37       47       73       64
                GT.66337 FG.66351 PE.66338 PH.66340 PE.66329 IS.66347 IS.66343
ENSG00000254468        2        1        0        0        1       10       32
ENSG00000177951      104       59      111      392      297     1286     7411
```

3.3.2 Counting Features from ChIP-seq Data

The `QuasR` procedure for ChIP-seq samples is similar in structure to the RNA-seq procedure above, except that we want to count features a bit differently.

1. First, make a samples list file, `qFchip.txt`, using the metadata table, which includes the sample's short name and its corresponding sorted BAM file:

```
> qFchip <- data.frame(FileName=paste0(sampsC$ENAid,"_s.bam"),
+                      SampleName=paste0(sampsC$antibody,".",sampsC$short.name))
> sampFchip <- "chip_seq_mapped/qFchip.txt"
> write.table(qFchip, file=sampFchip, quote=FALSE, sep="\t", row.names=FALSE)
```

2. Create the `qProject` object:

```
> genomeF <- "annotation/ensembl_Homo_sapiens.GRCh37.71.fa"
> projC <- qAlign(sample=sampFchip, genome=genomeF, paired="no")
```

3. Create TSS (transcription start site) annotation using the `annoGR` object:

```
> exonGR <- annoGR[annoGR$type=="exon"]
> transGRL <- split(exonGR, exonGR$transcript_id)
> tssGR <- unlist(range(transGRL))
```

Here the exonic regions of the `annoGR` object are selected by using the `type=="exon"` statement. The `transGRL` object is a `GRangesList` of exons, organized by transcript. The `tssGR` object contains the start and end coordinates of each of the transcripts.

4. Define promoter regions (here, 2,000 base pairs up- and downstream):

```
> prGR <- promoters(tssGR, upstream=2000, downstream=2000)
```

The `promoters` function is used to define a region around each TSS. We can use these 161,448 TSSs to create a count for each TSS.

5. Run the `qCount` function in order to summarize the ChIP-seq reads according to the features in the `prGR` object:

```
> library(parallel)
> cl <- makeCluster(10)
> countsC <- qCount(projC, prGR, clObj=cl)
> stopCluster(cl)
> save(countsC,file="countsC.Rdata")
```

3.4 edgeR Analysis

Once the feature counts are obtained, one can start the count-based analysis using the *edgeR* package. The count-based framework is very flexible and allows one to analyze various types of data—as long as they can be represented as counts. Here we illustrate some of the *edgeR* features, including the standard use case of finding differentially expressed genes; analysts can analogously use *edgeR* to discover histone modifications that are differentially enriched (not shown here). The examples illustrated below are not meant to be exhaustive, but rather to illustrate some of the functionality provided by *edgeR*. In particular, we repeat the cell-type-specific signature analysis of [21] according to a statistical criterion.

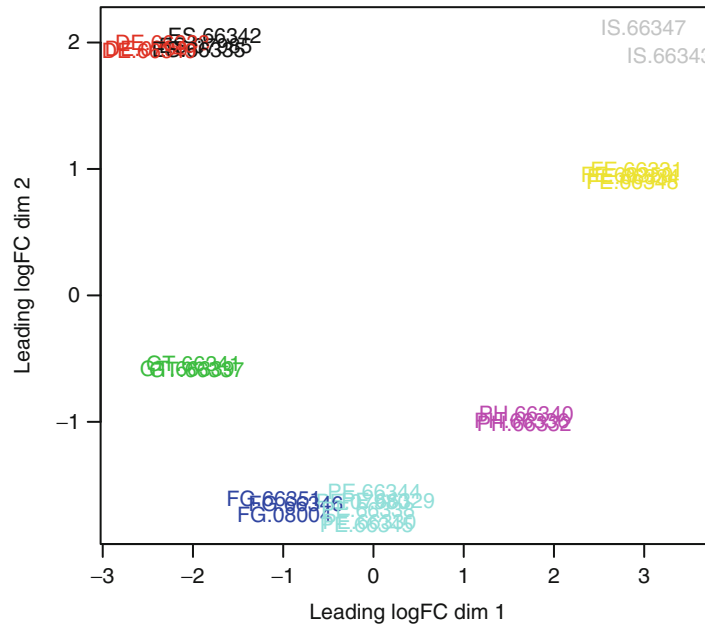


Fig. 6 MDS plot for RNA-seq samples: relationship between all samples according to two automatically determined dimensions (dim 1 and 2). "logFC" is log fold change

3.4.1 RNA-seq Data

1. Load the *edgeR* library and create a `DGEList` object, which is the container to store the raw counts and associated summaries (e.g., normalization factors) needed for the statistical methods.

```
> library("edgeR")
> d <- DGEList(counts=countsR[,-1], group=sampsR$cell.type)
> d <- calcNormFactors(d)
> d$genes <- data.frame(width=countsR[,"width"])
```

2. As a first look at the data, create a multidimensional scaling (MDS) plot as a spot check, by using the `plotMDS` function:

```
> plotMDS(d, col=as.numeric(d$samples$group))
```

This plot type, shown in Fig. 6, displays pairwise similarity of each sample in two automatically determined dimensions (*see* **Note 17**).

3. Compute count per million values by using the `cpm` function. We recommend filtering out genes that do not achieve at least a minimum abundance (e.g., 1 read per million) in at least the smallest group of replicates. Here the smallest number of replicates is two, corresponding to the islet cells (IS), so we use 2:

```
> cps <- cpm(d)
> k <- rowSums(cps>1)>=2
> d <- d[k,]
> cps <- cpm(d)
```

4. The RPKM values (reads per kilobase model) can be calculated using the rpkm function (*see* **Note 18**):

```
> rpks <- rpkm(d,d$genes$width,2)
```

5. If positive or negative controls are already known from existing work, it is always worthwhile to make additional checks before the formal statistical analysis. In doing this, we can confirm that the data being analyzed is consistent with the available body of knowledge.

   Here we look at the gene expression levels of several key genes, as mentioned in the manuscript [21]. To do this, semi-automatically query the *biomaRt* [23, 24] database to match gene symbols (`hgnc_symbol`) to identifiers used in the analysis (`ensembl_gene_id`) by using the `useMart` and `getBM` functions:

```
> pos.genes <- c("SOX17","POU5F1","SOX2","G6PC2","SOX9","PDX1")
> library("biomaRt")
> mart <- useMart("ensembl",dataset="hsapiens_gene_ensembl")
> bm <- getBM(attributes=c("ensembl_gene_id","hgnc_symbol"),
+             filters="hgnc_symbol", values=pos.genes, mart=mart)
> bm <- bm[bm$ensembl_gene_id %in% rownames(cps),]
> bm
```

```
   ensembl_gene_id hgnc_symbol
2  ENSG00000152254       G6PC2
3  ENSG00000139515        PDX1
4  ENSG00000204531       POU5F1
11 ENSG00000164736        SOX17
12 ENSG00000181449         SOX2
13 ENSG00000125398         SOX9
```

6. And use the `bm` lookup table to create a plot of the relative gene expression levels, here using the counts per million:

```
> par(mfrow=c(2,3))
> o <- order(sampsR$cell.type)
> cols <- as.numeric(factor(sampsR$cell.type[o]))
> for(i in 1:nrow(bm)) {
+   gid <- bm$ensembl_gene_id[i];
+   barplot(cps[gid,o], las=2, main=paste0(gid," // ",bm$hgnc_symbol[i]), col=cols)
+ }
```

   Figure 7 confirms many existing stage-specific genes: SOX17 is a known DE marker, OCT4 (POU5F1) and SOX2 are largely ES-specific, SOX9 and PDX1 are PE-specific, and G6PC2 is FE-specific.

7. Create a design matrix, here with a factor level for each cell type, that will be used later to construct the differences of interest:

```
> group <- sampsR$cell.type
> design <- model.matrix(~0+group, data=d$samples)
> colnames(design) <- levels(d$samples$group)
> design
```

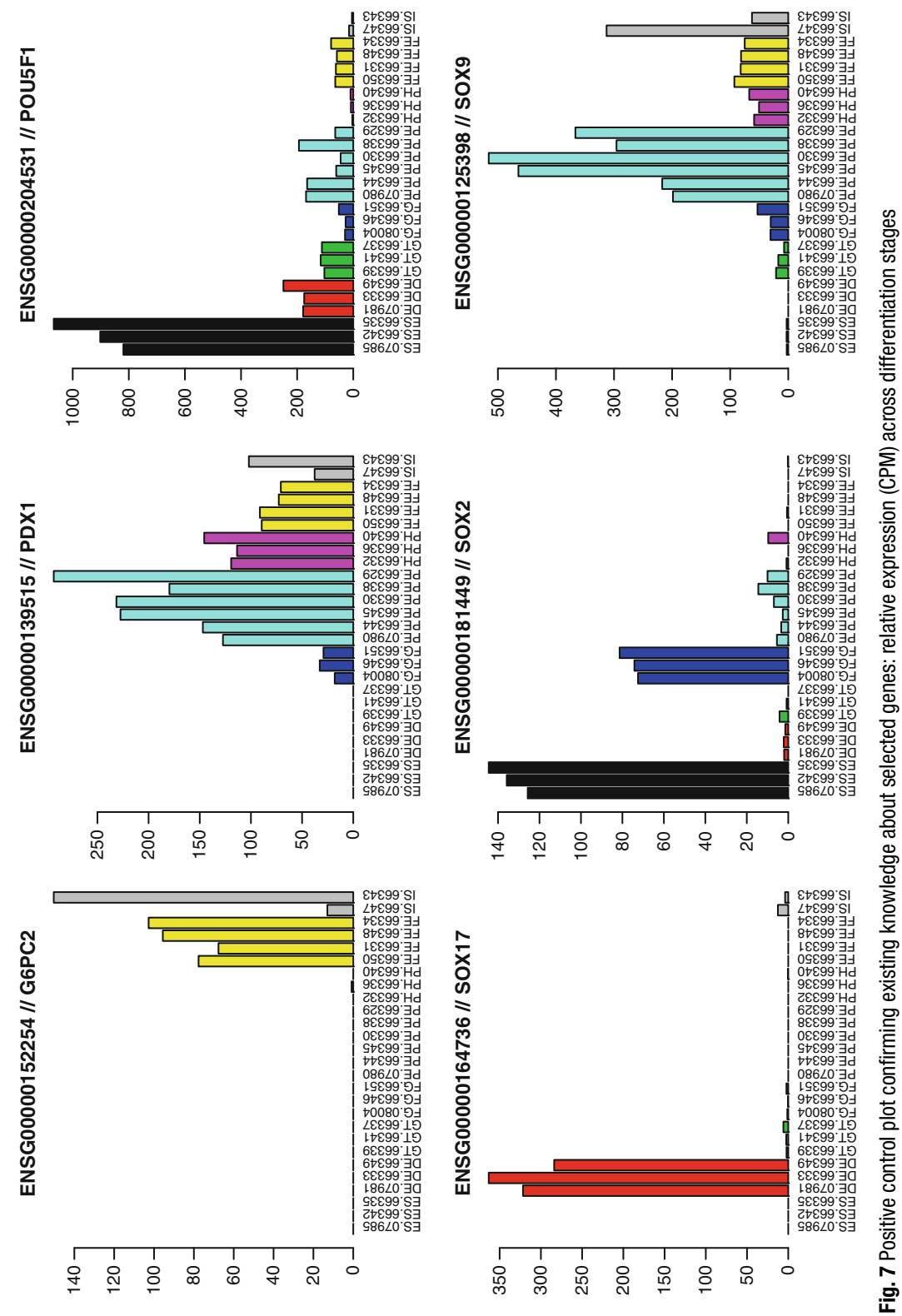


Fig. 7 Positive control plot confirming existing knowledge about selected genes: relative expression (CPM) across differentiation stages

```
          ES DE GT FG PE PH FE IS
ES.07985  1  0  0  0  0  0  0  0
DE.07981  0  1  0  0  0  0  0  0
GT.66339  0  0  1  0  0  0  0  0
FG.08004  0  0  0  1  0  0  0  0
PE.07980  0  0  0  0  1  0  0  0
FE.66350  0  0  0  0  0  0  1  0
ES.66342  1  0  0  0  0  0  0  0
DE.66333  0  1  0  0  0  0  0  0
GT.66341  0  0  1  0  0  0  0  0
FG.66346  0  0  0  1  0  0  0  0
PE.66344  0  0  0  0  1  0  0  0
FE.66331  0  0  0  0  0  0  1  0
PH.66332  0  0  0  0  0  1  0  0
PH.66336  0  0  0  0  0  1  0  0
PE.66345  0  0  0  0  1  0  0  0
PE.66330  0  0  0  0  1  0  0  0
FE.66348  0  0  0  0  0  0  1  0
FE.66334  0  0  0  0  0  0  1  0
ES.66335  1  0  0  0  0  0  0  0
DE.66349  0  1  0  0  0  0  0  0
GT.66337  0  0  1  0  0  0  0  0
FG.66351  0  0  0  1  0  0  0  0
PE.66338  0  0  0  0  1  0  0  0
PH.66340  0  0  0  0  0  1  0  0
PE.66329  0  0  0  0  1  0  0  0
IS.66347  0  0  0  0  0  0  0  1
IS.66343  0  0  0  0  0  0  0  1
attr(,"assign")
[1] 1 1 1 1 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"
```

8. Calculate the dispersion estimates, relative to the design matrix:

```
> d <- estimateGLMCommonDisp(d, design)
> d <- estimateGLMTrendedDisp(d, design)
> d <- estimateGLMTagwiseDisp(d, design)
```

9. Fit the GLM (i.e., estimate the relative gene-level abundance) according to the design matrix:

```
> f <- glmFit(d, design)
```

10. Use plotBCV function to plot the biological coefficient of variation (*see* **Note 19**).

```
> plotBCV(d)
```

Figure 8 shows an example of a typical RNA-seq-based BCV plot.

11. Create the `contrasts` matrix to determine signature genes of each of the differentiation stages. We compare the expression values of each stage to the average of all the other ones. Other possibilities exist to define stage-specific genes (*see* **Note 20**).

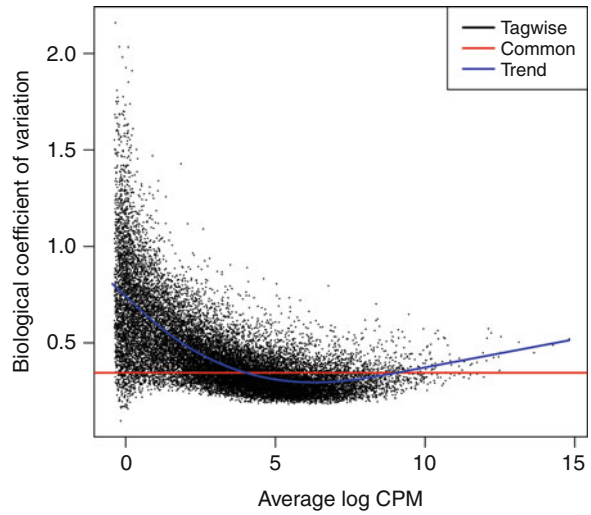


Fig. 8 BCV plot (biological coefficient of variation) plot for the RNA-seq samples. Each *black point* represents a gene and its corresponding dispersion estimate (Y-axis) and average log CPM (X-axis). Here, we can see the typical relationship, whereby estimated dispersion decreases with average mean

```
> contrasts <- makeContrasts(ESsig = ES-(DE+FE+FG+GT+PE)/5,
+                            DEsig = DE-(ES+FE+FG+GT+PE)/5,
+                            GTsig = GT-(ES+DE+FE+FG+PE)/5,
+                            FGsig = FG-(ES+DE+FE+GT+PE)/5,
+                            PEsig = PE-(ES+DE+FE+FG+GT)/5,
+                           FEsig = FE-(ES+DE+FG+GT+PE)/5, levels=design)
> contrasts
```

```
      Contrasts
Levels ESsig DEsig GTsig FGsig PEsig FEsig
    ES   1.0  -0.2  -0.2  -0.2  -0.2  -0.2
    DE  -0.2   1.0  -0.2  -0.2  -0.2  -0.2
    GT  -0.2  -0.2   1.0  -0.2  -0.2  -0.2
    FG  -0.2  -0.2  -0.2   1.0  -0.2  -0.2
    PE  -0.2  -0.2  -0.2  -0.2   1.0  -0.2
    PH   0.0   0.0   0.0   0.0   0.0   0.0
    FE  -0.2  -0.2  -0.2  -0.2  -0.2   1.0
    IS   0.0   0.0   0.0   0.0   0.0   0.0
```

12. Conduct a likelihood ratio test for each contrast and filter the top differentially expressed features. Since there are many differentially expressed genes for these contrasts, we consider only those with a low estimated false discovery rate and require the relative expression to be at least fourfold higher (log-fold change, lfc, of two):

```
> glmLRTs <- apply(contrasts, 2, function(u) glmLRT(f, contrast=u))
> sigList <- lapply(glmLRTs, function(u, fdr=.001, lfc=2) {
+   tt <- topTags(u, n=nrow(d))
+   tt$table[tt$table$FDR < fdr & tt$table$logFC > lfc,]
+ })
> sapply(sigList,nrow)
```

```
ESsig DEsig GTsig FGsig PEsig FEsig
  983   898   417   599   755  2488
```

Note that the `apply` and `lapply` simplify running the commands across the columns of the contrast matrix. An equivalent alternative is to use a `for` loop and store the results of each comparison in a list object.

13. Next, we perform another spot check that requires a customized capture of information from the original publication's Supplementary Material; specifically, we overlap our sets of signature genes called from the statistical analysis to those published in [21].

    First, we read the lists of signature genes from the Supplementary XLS file (http://www.sciencedirect.com/science/MiamiMultiMediaURL/1-s2.0-S1934590912007060/1-s2.0-S1934590912007060-mmc2.xls/274143/FULL/S1934590912007060/2af27e3306cf41d22af0daa12e61414d/mmc2.xls) into *R* (*see* **Note 21**):

```
> library(gdata)
> nm <- sapply(1:5,function(u) read.xls("mmc2.xls",sheet=u,nrow=1,sep=",",
+                                       header=FALSE,stringsAsFactors=FALSE$(V1)
> sigGenes <- lapply(1:5,function(u) read.xls("mmc2.xls",sheet=u,skip=2,sep=",",
+                                      header=TRUE,stringsAsFactors=FALSE)$RefSeq.ID)
> names(sigGenes) <- nm
> all.sig <- unique(unlist(sigGenes))
```

14. Next, we need to convert the identifiers from RefSeq to Ensembl, to match our processing of the data.

```
> library("biomaRt")
> mart <- useMart("ensembl",dataset="hsapiens_gene_ensembl")
> bm <- getBM(attributes=c("ensembl_gene_id", "refseq_mrna"),
+             filters="refseq_mrna", values=all.sig, mart=mart)
> head(bm)
```

```
  ensembl_gene_id refseq_mrna
1 ENSG00000175899   NM_000014
2 ENSG00000196839   NM_000022
3 ENSG00000135744   NM_000029
4 ENSG00000148218   NM_000031
5 ENSG00000101986   NM_000033
6 ENSG00000134982   NM_000038
```

15. For each signature, we need to map the RefSeq identifiers to Ensembl identifiers, which we can do easily by manipulating the *bm* lookup table created above:

```
> sigGenesENS <- lapply(sigGenes, function(u) unique(bm$ensembl_gene_id[bm$refseq_mrna %in% u]))
> sapply(sigGenesENS,length)
```

```
      Table S2 (A) Definitive endoderm (DE) signature genes
                                                        673
  Table S2 (B) Primitive gut tube (GT) signature genes
                                                        153
   Table S2 (C) Posterior foregut (FG) signature genes
                                                        560
```

```
 Table S2 (D) Pancreatic endoderm (PE) signature genes
                                                    232
Table S2 (E) Functional endocrine (FE) signature genes
                                                    710
```

16. We use the VennDiagram package to create a Venn diagram for each of the five signature genes' comparisons:

```
> library("VennDiagram")
> pdf("venn_comparison.pdf")
> for (u in 1:length(sigGenesENS)){
+ venn.plot <- venn.diagram( x=list("Xie et al."=sigGenesENS[[u]],"edgeR"=rownames(sigList[[u+1]])),
+  filename=NULL, scaled=TRUE, fill=c("gray", "blue"),
+  main=gsub("sig", " signature genes", names(sigList[u+1])),
+  main.cex=2, cex=2.5, cat.cex=1.5, cat.pos=0, ext.txt=TRUE, height=3000, width=3000)
+ grid.draw(venn.plot)
+ grid.newpage()
+ }
> dev.off()
```

To combine the resulting PDF figures onto one plot, run the following commands in Unix:

```
pdf2ps venn_comparison.pdf venn_comparison.ps
psnup -6 venn_comparison.ps venn_comparison6
ps2pdf12 venn_comparison6.ps venn_comparison6.pdf
pdftk venn_comparison6.pdf cat 1E output venn_comparison6_ls.pdf
```

Here the `pdf2ps` command performs the format type conversion and the `psnup` command places the figures onto one plot. Finally, the `ps2pdf12` command converts the PS file back to PDF and the `pdftk` command rotates the file landscape orientation.

Figure 9 shows the resulting Venn diagram, highlighting a strong but not perfect overlap with the original analysis.

17. Make a heatmap of signature genes, by cell type. First, the expression counts are selected by subsetting the `cpms` object with the signature genes contained in `gnames`. The `apply` function is used to scale all expression values for a given gene as a percentage of the maximum expression value. Afterward, the counts are averaged across all samples of the same cell type by using the `sapply` and `rowMeans` functions to provide a metric to organize the order of genes.

```
> gnames <- unique(unlist(lapply(sigList, rownames)))
> keep <- grep("IS|PH",colnames(cps),invert=TRUE)
> cpsH <- sqrt(cps[rownames(cps) %in% gnames,keep])
> cpsH <- t(apply(cpsH, 1, function(x) 100*x/max(x)))
> uct <- unique(samps$cell.type[keep])
> cpsA <- matrix(0, nrow=nrow(cpsH), ncol=length(uct),
+                dimnames=list(rownames(cpsH),uct))
> for(i in 1:ncol(cpsA)) {
+   rn <- rownames(sigList[[i]])
+   cl <- grep(colnames(cpsA)[i],colnames(cpsH))
+   cpsA[rn,i] <- rowMeans(cpsH[rn,cl,drop=FALSE])
+ }
```

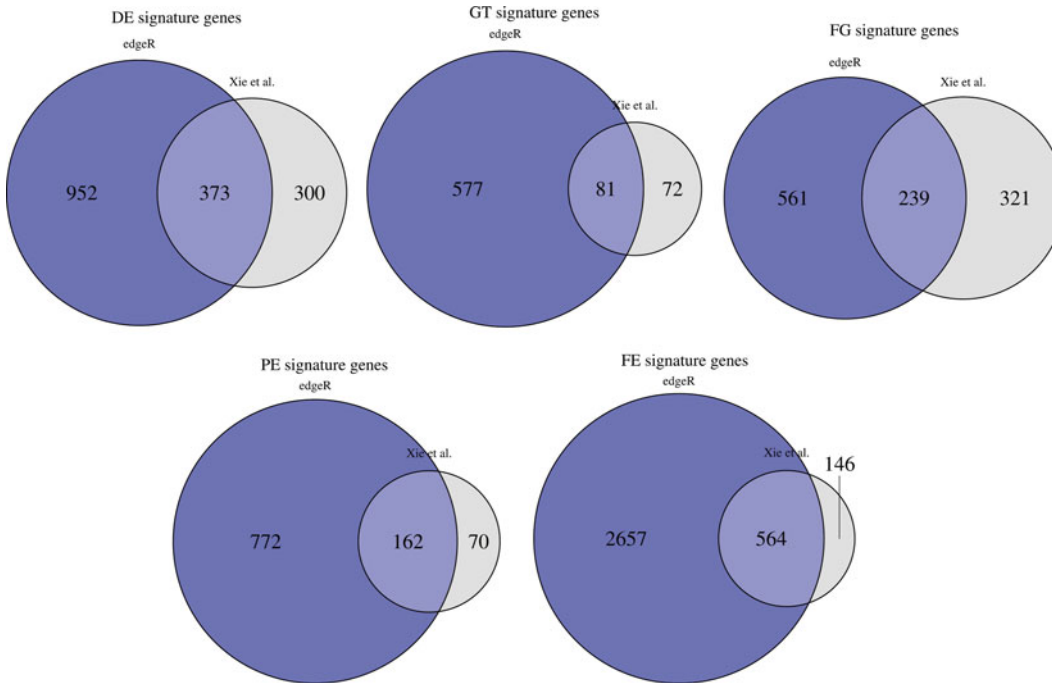


Fig. 9 Venn diagram of signature genes: comparison of number of genes identified by *edgeR* and in the original publication [21], by cell type

To plot the heatmap, we use the `heatmap.2` function. The `colorRampPalette` allows replacement of the default color scheme.

```
> library("gplots")
> cpsA <- as.data.frame(cpsA)
> ro <- order(cpsA$ES, cpsA$DE, cpsA$GT, cpsA$FG, cpsA$PE, cpsA$FE)
> co <- order(sampsR$cell.type[keep])
> hm <- heatmap.2(cpsH[ro,co], scale="none",
+    col=colorRampPalette(c("yellow","green","blue"),space="Lab")(128), dendrogram="none",
+    Colv=NULL, key=FALSE, trace="none", labRow="")
```

Figure 10 shows the collection of signature genes, organized according to their signature status.

18. Generate a smear plot by using the `plotSmear` command (*see* **Note 22**) and annotate the plot with the set of differentially expressed genes and some positive controls:

```
> par(mfrow=c(1,2))
> plotSmear(glmLRTs[["ESsig"]],de.tags=rownames(sigList[["ESsig"]]),
+           deCol = "blue", main="ES signature genes")
> w <- which( rownames(glmLRTs[["ESsig"]])=="ENSG00000204531" )
> x <- glmLRTs[["ESsig"]]$table$logCPM[w]
> y <- glmLRTs[["ESsig"]]$table$logFC[w]
```

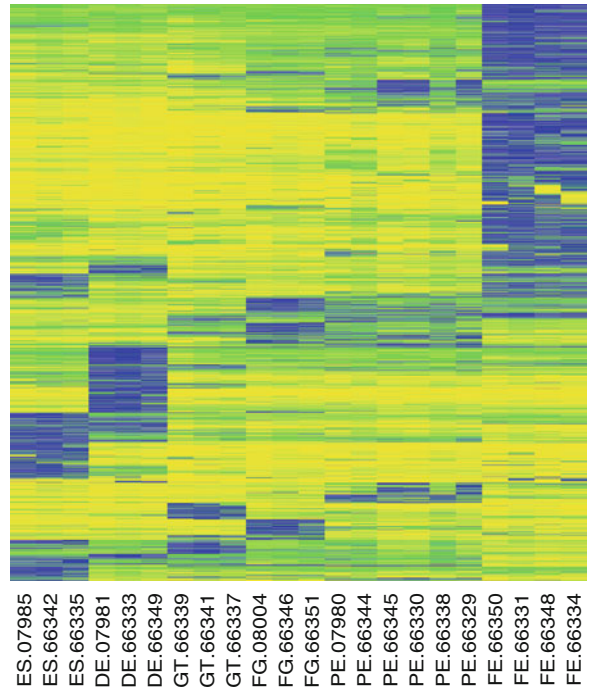


Fig. 10 Heatmap of differentiation stage signature genes identified from the RNA-seq samples. *Rows* correspond to individual genes. *Column* blocks correspond to samples, arranged by cell type according to the differentiation order. *Colors* correspond to expression signal of particular gene, scaled to percentage of the maximum expression across all samples. *Yellow* corresponds to low expression, *blue* to high

```
> points(x,y,col="orange",cex=2)
> text(x,y,"OCT4/POU5F1",adj=-.1)
> plotSmear(glmLRTs[["DEsig"]],de.tags=rownames(sigList[["DEsig"]]),
+           deCol = "blue", , main="DE signature genes")
> w <- which( rownames(glmLRTs[["DEsig"]])=="ENSG00000164736" )
> x <- glmLRTs[["DEsig"]]$table$logCPM[w]
> y <- glmLRTs[["DEsig"]]$table$logFC[w]
> points(x,y,col="orange",cex=2)
> text(x,y,"SOX17",adj=-.1)
```

Figure 11 shows the resulting smear plot.

3.4.2 ChIP-seq Data

1. Create a `DGEList` object for ChIP-seq count data:

```
> dc <- DGEList(counts=countsC[,-1], group=sampsC$cell.type)
```

2. Create an MDS plot as a spot check for the ChIP-seq data:

```
> plotMDS(dc, col=as.numeric(factor(sampsC$cell.type)))
```

Based on the MDS plot shown in Fig. 12, it seems that the type of chromatin mark has greater effect on the position on the MDS plot than the differentiation stage. In addition, one can see some organization of the changes in the epigenome.

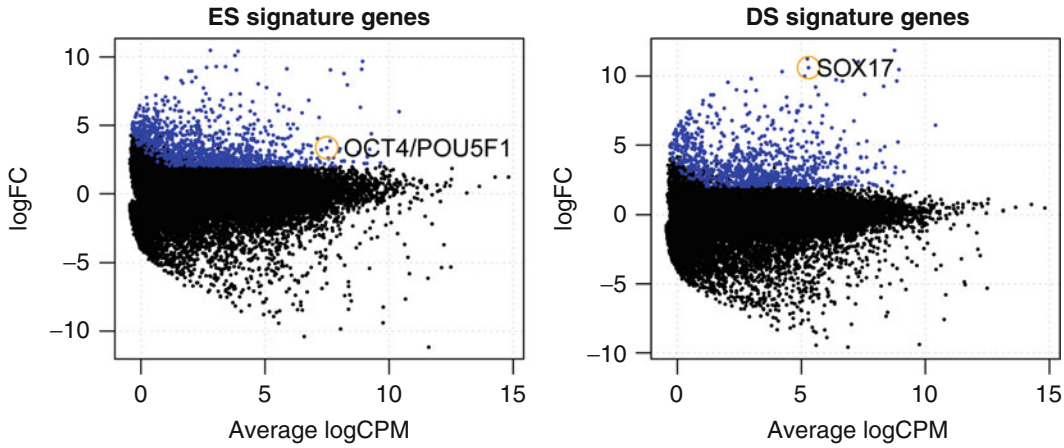


Fig. 11 Smear plot generated by *edgeR* during RNA-seq analysis. The x-axis (log-average of read counts) is the expression strength, while the y-axis (log-ratio of difference of interest) reflects the relative log-fold-change for the contrast of interest (here, ES expression versus the average of the other cell types)

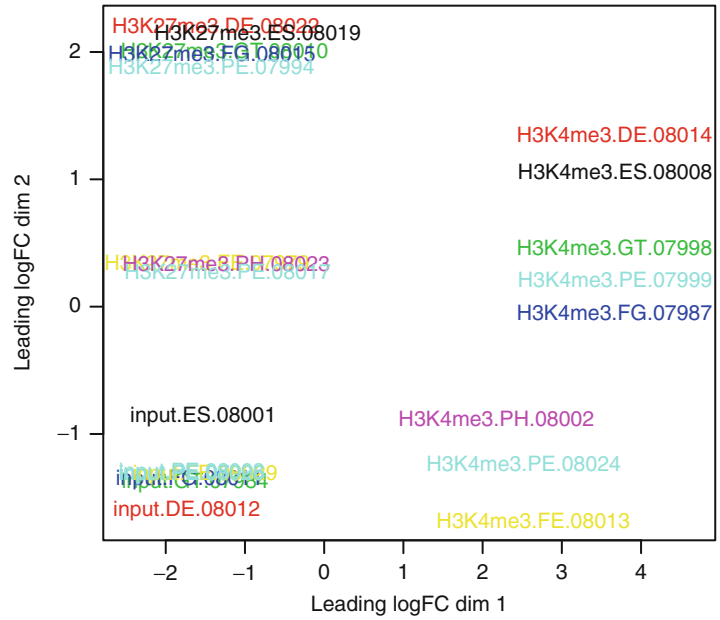


Fig. 12 MDS plot for ChIP-seq samples

Differential analyses of the TSS-level ChIP-seq counts can proceed with *edgeR* analogously to the RNA-seq analysis above (not done here).

3.4.3 Integrative Analyses

1. Define canonical transcription start site (ctss) annotation. First, create a `txlen` object that contains the exon widths, grouped by transcript identifiers, and then use the `sum` function to calculate the lengths of all transcripts and arrange them in descending order:

```
> txlen <- split(width(exonGR),exonGR$transcript_id)
> txlen <- sort(sapply(txlen, sum), decreasing=TRUE)
```

2. Match the names of the transcripts contained in `txlen` to the transcript identifiers of the `annoGR` object:

```
> m <- match(names(txlen), annoGR$transcript_id)
```

3. Group transcript lengths by `annoGR` gene identifiers, take the name of the largest (here, first) transcript to represent the entire gene and subset to only those used in the expression analysis (some were filtered based on low expression):

```
> txleng <- split(txlen, annoGR$gene_id[m])
> txleng <- sapply(txleng, function(u) names(u)[1])
> txleng <- txleng[ names(txleng) %in% rownames(cps) ]
```

4. Finally, define the `ctssGR` object, containing the canonical transcription start site information (by gene) and clean up:

```
> m <- match(txleng, names(tssGR))
> ctssGR <- tssGR[m]
> names(ctssGR) <- names(txleng)
> save(ctssGR, file="ctssGR.Rdata")
> rm(exonGR,transGRL,tssGR,prGR,m)
> gc()
```

```
            used     (Mb) gc trigger     (Mb)   max used     (Mb)
Ncells    4201496    224.4     7540570    402.8    7540570    402.8
Vcells  152331936   1162.3   304420522   2322.6  304420512   2322.6
```

As with intermediate files on disk, it may be advisable to remove intermediate objects created during an R session, since they consume memory. Users can use the `rm` command to remove objects not needed. This does not always free the memory immediately, but one can force a "garbage collection" process using the `gc` command.

5. Use the `featureScores` function and the `ctssGR` object to collect the observed ChIP-seq in the neighborhood around the TSSs:

```
> library(Repitools)
> k <- c(5,8)
> bf <- file.path("chip_seq_data",qFchip$FileName[k])
> names(bf) <- qFchip$SampleName[k]
> fs <- featureScores(bf, ctssGR, up=5000, down=5000, freq=50, s.width=c(500,1000))
> save(fs,file="fs.Rdata")
```

6. Use the `binPlots` function to create a bin plot for the H3K4me3 chromatin mark, as shown in Fig. 13:

```
> m <- match(names(ctssGR), rownames(rpks))
> expr <- rowMeans(rpks[m,grep("ES",colnames(rpks))])
> expr <- log(expr+.25)+rnorm(length(expr),sd=.001)
> binPlots(fs[1],ordering=expr,ord.lab="ES cell RPKM",
+          n.bins=50,plot.type="heatmap")
```

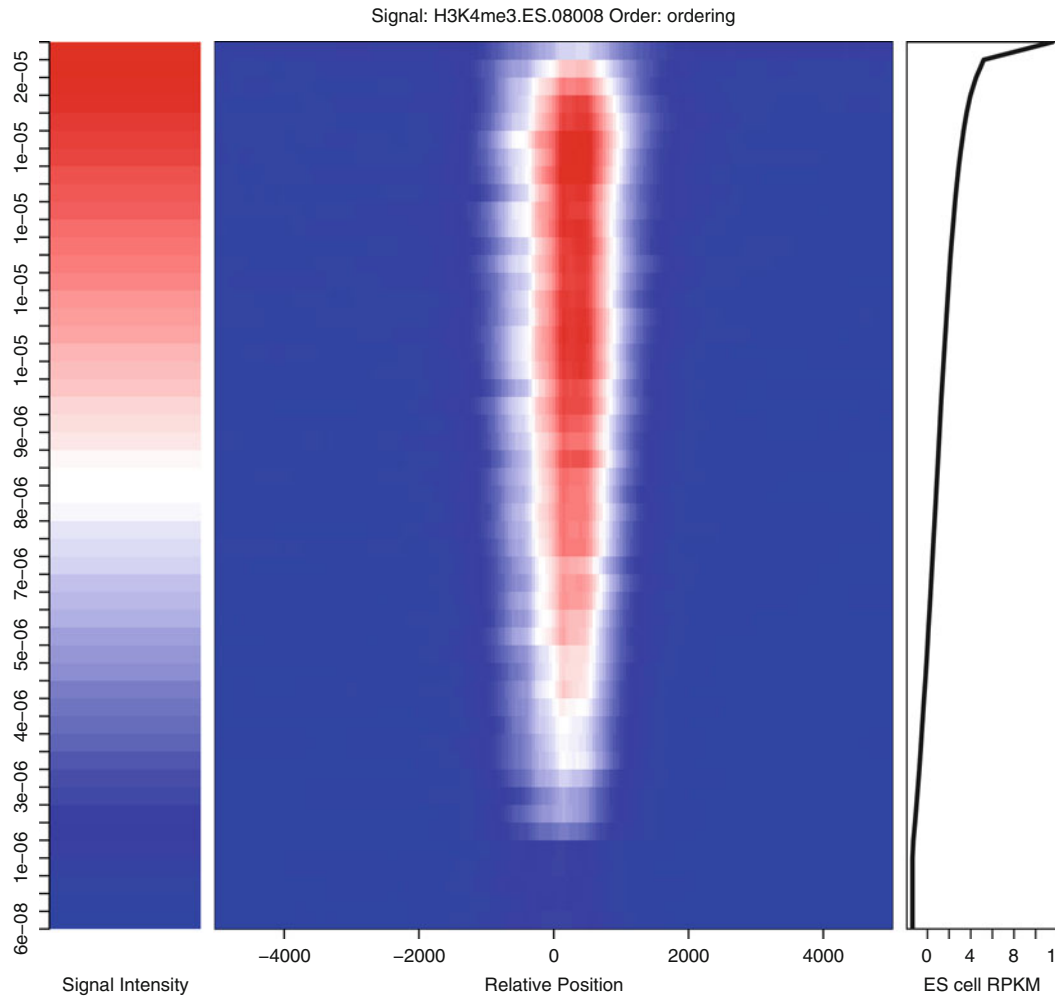


Fig. 13 Integrative analysis: H3K4me3 signal according to expression strength. Along the X-axis is the location of signal relative to a gene's TSS. The colors in the heatmap highlight the epigenome signal, averaged over all the genes in the bin (defined by expression-RPKM)

7. Similarly, this plot can be represented by lines, as shown for H3K27me3 signal in Fig. 14:

```
> binPlots(fs[2],ordering=expr,ord.lab="ES cell RPKM",
+          n.bins=10,plot.type="line")
```

3.5 Versions

```
> sessionInfo()
R version 3.0.1 (2013-05-16)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
[1] LC_CTYPE=en_CA.UTF-8       LC_NUMERIC=C
[3] LC_TIME=en_CA.UTF-8        LC_COLLATE=en_CA.UTF-8
[5] LC_MONETARY=en_CA.UTF-8    LC_MESSAGES=en_CA.UTF-8
[7] LC_PAPER=C                 LC_NAME=C
```

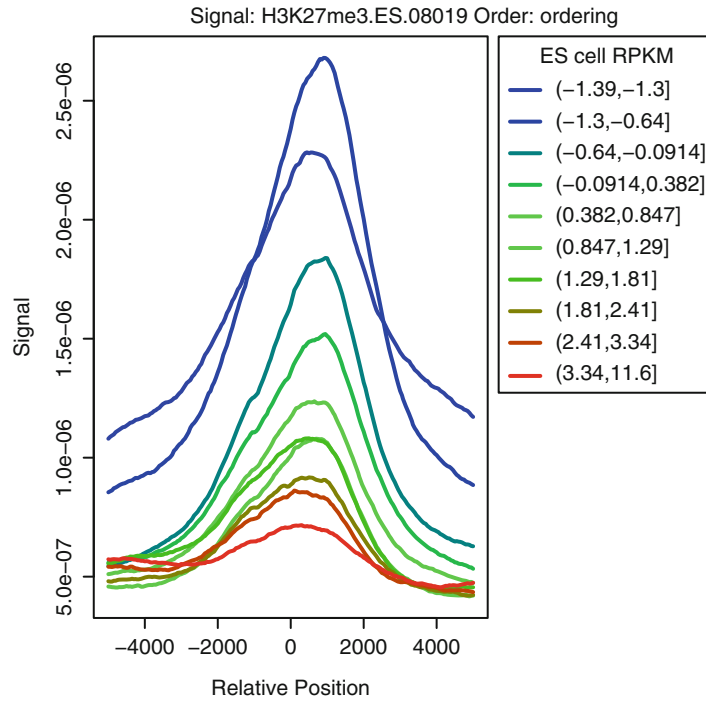


Fig. 14 H3K27me3 signal according to expression strength as a line plot. Line colors reflect the expression level, while line shapes correspond to averaged H3K27me3 signal across the promoters

```
 [9]   LC_ADDRESS=C       LC_TELEPHONE=C
[11]  LC_MEASUREMENT=en_CA.UTF-8 LC_IDENTIFICATION=C

attached base packages:
 [1]  parallel  grid     splines  stats  graphics  grDevices utils
 [8]  datasets  methods  base

other attached packages:
 [1]  QuasR_1.0.5           Rbowtie_1.0.3        Repitools_1.6.0
 [4]  GenomicRanges_1.12.4  IRanges_1.18.2       BiocGenerics_0.6.0
 [7]  gplots_2.11.3         MASS_7.3-28          KernSmooth_2.23-10
[10]  caTools_1.14          gtools_3.0.0         VennDiagram_1.6.4
[13]  gdata_2.13.2          biomaRt_2.16.0       edgeR_3.2.4
[16]  limma_3.16.7          cacheSweave_0.6-1    stashR_0.3-5
[19]  filehash_2.2-1

loaded via a namespace (and not attached):
 [1]  AnnotationDbi_1.22.6  Biobase_2.20.1       BiocInstaller_1.10.3
 [4]  Biostrings_2.28.0     bitops_1.0-5         BSgenome_1.28.0
 [7]  DBI_0.2-7             digest_0.6.3         GenomicFeatures_1.12.3
[10]  hwriter_1.3           lattice_0.20-15      RCurl_1.95-4.1
[13]  Rsamtools_1.12.3      Rsolnp_1.14          RSQLite_0.11.4
[16]  rtracklayer_1.20.4    ShortRead_1.18.0     snowfall_1.84-4
[19]  stats4_3.0.1          tools_3.0.1          truncnorm_1.0-6
[22]  XML_3.98-1.1          zlibbioc_1.6.0
```

4 Notes

1. Many variations exist on these protocols (e.g., strand-specificity, paired-end sequencing) and slight modifications may need to be considered in the processing steps (e.g., mapping). For more details regarding the library preparation in this case, such as the polyA capture process, refer to the original paper [21].
2. A nice feature of these public databases of sequencing (and microarray) data is that they require researchers to give a full description the experiments conducted. This information, the so-called metadata, is typically available in machine-readable format that can be automatically captured, organized, and used to automate some tasks. No data analysis is fully automated, but the data analyst can reproducibly capture the necessary information and reduce the amount of manual effort (e.g., adding sample labels in a spreadsheet program). In many cases, most of the metadata can be parsed directly from the information within files from the data providers. Similarly, individual research labs should make organized systems for keeping a log of the experiments done, with all the useful metadata, such as date, operator, batch, etc.
3. It is helpful to assign meaningful (but short) variable names. Here we typically use an abbreviation for the variable name, such as `md` for metadata. Also, we sometimes encode the variable's data type into its name. For example, variables with `GR` are of the `GRanges` data type and variables with `GRL` are of the `GRangesList` data type.
4. Reference genomes might be available from multiple sources (e.g., Ensembl, UCSC). The most important consideration is ultimately that the gene annotation and the genome sequence are on the same coordinate system, so users should be sure to choose reference genome and reference annotation with matching coordinates.
5. Since making a set of index files can be time-consuming, it may prove useful to check whether a pre-built index is already available online, either from *bowtie2* or from Illumina's "iGenomes' project (http://tophat.cbcb.umd.edu/igenomes.shtml). In addition, we note that the building an index needs to happen only once for a given genome build.
6. A large number of sample files are typical of bioinformatics analyses. For example, this experiment is a collection of both the RNA-seq and ChIP-seq samples, with 27 and 24 raw data files, respectively. There are many ways to organize files and directories for such analyses. One simple approach that many new users try is to collect all files for a project in a single directory. In practice, this strategy often becomes unmanageable, due to various intermediate files created during processing.

Therefore, we recommend additional organization, such as subdirectories for various sub-tasks.

7. If necessary, the `getwd` and `setwd` commands in *R* can be used to determine and adjust the current working directory.
8. It is important to know that quality of the sequencing reactions is high prior to any formal statistical analysis. Depending on the sequencing quality, it might be necessary to exclude some files from further processing or adjust mapping parameters (e.g., trimming). In some cases, soft trimming is implemented within the alignment algorithm.
9. BAM stands for binary alignment format and is the *de facto* standard for read alignments.
10. The `nice -n 19` parameter is optional and gives the resulting process a lower priority value (i.e., a high "niceness" value). This option is often helpful in a multiple-user server environment, where several members of a research group are sharing computing resources. In the absence of other jobs, *tophat2* would receive high priority. The `-G` option lists the annotation file, in GTF format; this is important to help the aligner map reads that span exon–exon junctions. The `-p` specifies the number of compute cores to use, which will vary according to the resources available. The `-o` option is used to indicate the output directory; here, we use the sample accession numbers. After all the options, the path of the *bowtie2* index is given, followed by the pathname of the FASTQ files.
11. When sufficient computing resources (CPUs and memory) are available, it might be helpful to split the *tophat2* commands into parts and run them in parallel. In addition, we find the Unix tool *screen* (http://en.wikipedia.org/wiki/GNU_Screen) indispensable for managing mapping jobs. In our case, we simultaneously ran two jobs (each using 10 cores), with other jobs running in addition to the *tophat2* tasks, and it took approximately 6 days (approximately 6 h per RNA sample).
12. The *bowtie2* commands also include several options, similar to the *tophat2* commands. The `-p` option indicates the number of cores to use. The `-x` parameter indicates the basename of the index for the reference genome (excluding the file extension). The `-U` parameter indicates the file, containing unpaired reads, to be aligned. Here the `-U` parameter is followed by the location of the compressed FASTQ file. Unlike *tophat2*, *bowtie2* outputs a SAM file, the name of which is indicated by the `-S` parameter.

    As in the case with *tophat2* commands, the *bowtie2* commands can be split into groups to run in several processes, in order to improve job management and decrease the total running time.

13. Since files typically used in a Bioinformatics context are quite large, it is important to be mindful of disk space and to remove intermediate files after a certain processing stage has been achieved. For example, after the BAM files are created, the corresponding SAM files can be deleted to free up substantial disk space. The same applies to the `accepted_hits.bam` files—they can be removed after their sorted versions are created (Note: these BAM files can be recreated, if necessary, by realigning).
14. `QuasR` can be used to do the alignments direct from the *R* environment (e.g., see `?qAlign`), but given the large size of the experiment here, we preferred to manage the alignments as separate processes outside of *R*.
15. Here the `parallel` library is used to benefit from multicore processors that are available on most desktop and server computers, in order to speed up the counting process. Ten cores are used here. Also note the `reportLevel="gene"` and the `orientation="same"` arguments passed to `qCount`. The former indicates that the reads should be summarized by gene, whereas the latter indicates that only reads corresponding to the same strand should be considered, because a strand-specific protocol was used to obtain the RNA-seq data.
16. If no multicore processors are available, the user does not need to supply the `clObj` argument. In that case, a single process will be used, leading to a longer computational time.
17. On the resulting MDS plot, samples separate by their differentiation stage and biological replicates corresponding to the same differentiation stage cluster together. This is an early indication of a dataset of high quality, where clear differences between conditions but not within replicates drive the placing of samples in the MDS plot.
18. We use RPKMs here since comparisons *between genes* are used, whereas in the differential expression analysis above, comparisons *across samples* of the same gene are desired, so CPMs were perfectly reasonable.
19. Biological coefficient of variation (BCV) is a unitless measure signifying the degree of variability; in general, technical replicates have low BCV, genetically identical replicates of the same experimental condition (e.g., laboratory mice) have a medium BCV and outbred populations (e.g., human cohorts) have high BCVs. In addition, many datasets exhibit a downward BCV versus mean trend, although the exact shape can vary by dataset.
20. In the contrast matrix shown, the stage-specific genes were defined as those where the average of the stage of interest is different from the average of the remaining 5 stages. There may be situations where the gene is specific to 2 stages, but because of

the averaging, there is still strong evidence to support the change for a particular stage. An alternative more conservative definition is to define a stage-specific gene as one that is different from the stage of interest to all other stages. This would require an alternative contrast matrix: all 15 pairs of stages. One could then test the block of pairwise comparisons that corresponded to a stage of interest simultaneously with a single likelihood ratio test or even more stringently, reduce to only the set of genes that are beyond a particular criterion for every pairwise comparison; we do not pursue these possibilities here.

21. We assume that the mmc2.xls supplementary information file has been downloaded into the current working directory.
22. The "smear" plot is an RNA-seq version of the ubiquitous M-versus-A plots that were used in the presentation of microarray results. In particular, the X-axis (A-value; log-average of read counts) is the expression strength, while the Y-axis (M-value; log-ratio of difference of interest) reflects the relative log-fold-change for the contrast of interest (here, ES expression versus the average of the other cell types).

References

1. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Prot (in press)
2. Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res Adv Ac (2008):1–19. ISSN 10889051
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18(9):1509–17. ISSN 10889051
4. McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 1–10. ISSN 1362-4962
5. Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23(21):2881–2887
6. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10):R80
7. R Development Core Team R (2011) R: A language and environment for statistical computing. ISSN 16000706
8. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. ISSN 14656906
9. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881
10. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J (2010) MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38(18):e178
11. Liao Y, Smyth GK, Shi W (2013) The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41(10):e108
12. Fiume M, Williams V, Brudno M (2010) Savant: Genome Browser for high throughput sequencing data. Bioinformatics 26(1):1–7
13. Fiume M, Smith EJM, Brook A, Strbenac D, Turner B, Mezlini AM, Robinson MD, Wodak SJ, Brudno M (2012) Savant Genome Browser 2: visualization and analysis for population-scale genomics. Nucleic Acids Res 40(W1):1–7. ISSN 13624962

14. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform Adv pu:bbs017. ISSN 14774054.
15. Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5(3):299–314
16. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 25(19):2607–2608
17. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140
18. Carlson M, Pages H, Aboyoun P, Falcon S, Morgan M, Sarkar D, Lawrence M. GenomicFeatures: Tools for making and manipulating transcript centric annotations
19. Lawrence M, Gentleman R, Carey V (2009) rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25: 1841–1842
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079
21. Xie R, Everett LJ, Lim H-W, Patel Na, Schug J, Kroon E, Kelly OG, Wang A, D'Amour Ka, Robins AJ, Won KJ, Kaestner KH, Sander M (2013) Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. Cell Stem Cell 12(2):224–37. ISSN 1875–9777
22. Lerch A, Gaiditzis D, Stadler MB (2012) QuasR: quantify and annotate short reads in R
23. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005) BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21(16):3439–40. ISSN 13674803
24. Durinck S, Spellman P, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 4(8):1184–1191

Chapter 4

Use Model-Based Analysis of ChIP-Seq (MACS) to Analyze Short Reads Generated by Sequencing Protein–DNA Interactions in Embryonic Stem Cells

Tao Liu

Abstract

Model-based Analysis of ChIP-Seq (MACS) is a computational algorithm for identifying genome-wide protein–DNA interaction from ChIP-Seq data. MACS combines multiple modules to process aligned ChIP-Seq reads for either transcription factor or histone modification by removing redundant reads, estimating fragment length, building signal profile, calculating peak enrichment, and refining and reporting peak calls. In this protocol, we provide a detailed demonstration of how to apply MACS to analyze ChIP-Seq datasets related to protein–DNA interactions in embryonic stem cells (ES cells). Instruction on how to interpret and visualize the results is also provided. MACS is an open-source and is available from http://github.com/taoliu/MACS.

Key words ChIP-Seq, Peak calling, Transcription factor, Histone modification

1 Introduction

Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-Seq), since it was introduced in 2007 [1–4], has been widely adopted to detect where a protein binds to DNA on genome scale. Proteins are firstly cross-linked to chromatin; after chromatin is sheared into smaller fragments, a specific antibody is used to precipitate the protein of interest together with associated DNA; at last, the ChIPed fragments are purified and their ends are read out by high-throughput sequencing. ChIP-Seq is now the most widely used technology for genome-wide assays of protein–DNA interaction in embryonic stem cells (ES cells), especially by several consortium projects [5, 6]. Its applications have been extended to study not only transcription factors and modified histones but also nucleosomes, chaperones, and other DNA-binding enzymes.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_4, © Springer Science+Business Media New York 2014

ChIP-Seq analysis begins with computationally aligning reads from sequencer, typically 50–100 bp long, to a reference genome. Depending on sequencing depth, nature of protein, and efficiency of ChIP, certain proportion of sequenced reads are enriched at the locations of protein–DNA interaction sites [7]. A pivotal step of analysis is to capture the characteristic reads distribution at the interaction sites and then to detect the regions with significant enrichment over background noise. This step is commonly called as "peak calling" since majority of transcription factors and some histone modifications are point-source factors, which means they bind to DNA at narrow regions and have a peaky enrichment comparing to surrounding regions. For these factors, reads mapped to Waston and Crick strand of reference genome would have a specific signal lag depending on the length of interaction sites. On the other hand, most epigenetic marks have broadly spreading and much weaker patterns over thousands of base pairs long regions. Their enrichment pattern should be better described as "domains" rather than "peaks."

Model-based Analysis of ChIP-Seq (MACS) was developed to identify read-enriched regions from ChIP-Seq data, firstly published in 2008 [8]. MACS has been cited by more than 800 studies according to Google Scholar. Over the years, MACS has continuously benefited from user feedback and contribution and has evolved from a point-source-specific peak caller to a more generalized ChIP-Seq analyzer. In this protocol, we will focus on the recent MACS version 2 (referred as MACS later in this protocol), which has been used in ENCODE and modENCODE data processing [9]. MACS is a completely redesigned algorithm from previous version, modularized with 11 useful modules that can be invoked separately (Table 1). MACS evaluates noise from ChIP-Seq control sample and then calculates enrichment scores quantitatively over entire genome, so as to ensure sensitivity and specificity. MACS has the capability to process huge datasets, with small amount of memory and fast speed. Here we demonstrate how to use the "callpeak" and "bdgcmp" modules in MACS (2.0.10) to find enriched regions from the NANOG (an important homeobox protein involved in ES cells) ChIP-Seq dataset and H3K36me3 (a broad histone mark covering active gene bodies) ChIP-Seq dataset in human H1 ES cell line [5], then generate genome-wide enrichment profile, and virtualize them in genome browser. Please refer to MACS manual for usage of other modules.

2 Materials

2.1 Computer Hardware

Need a computer workstation or server. A minimum of 2 GB of RAM is needed for processing 60 million human ChIP sample reads and 60 million control sample reads together. Estimation can be seen in Fig. 1. The RAM may need to be increased for more deeply sequenced ChIP-Seq data. The following examples were run on a computer server with a 2.8 GHz CPU.

Table 1
Eleven modules in MACS. Refer to MACS manual or command-line message for details

callpeak	Main MACS Function to call peaks from alignment results
bdgpeakcall	Call peaks from bedGraph file
bdgbroadcall	Call broad domains from bedGraph file
bdgcmp	Deduct noise by comparing two signal tracks in bedGraph format
bdgdiff	Differential peak detection based on paired four bedGraph files
diffpeak	Another differential peak detection tool with more statistics
filterdup	Remove duplicate reads at the same position, then convert acceptable format to simpler BED format
predictd	Predict binding fragment length from alignment results
pileup	Pileup aligned reads with a given extension size
randsample	Randomly sample certain number or percentage of total reads
refinepeak	(Experimental) Take raw reads alignment, refine peak summits, and give scores measuring balance of forward–backward tags

2.2 Computer Software

We recommend running MACS in a Unix-based operation system including Unix variations, Linux, and Mac OS, although MACS does work under Windows with Cygwin, a Linux simulator. MACS is a command-line program which means it should be invoked under command-line interface—a shell. To run MACS on a remote server, Internet connection and telnet or SSH software are necessary. Since MACS is programmed in Python language, which is famous on its elegance and simplicity, a proper Python environment is required. MACS works with Python version 2.7, and we recommend to run it in an isolated Python environment using VirtualEnv software (*see* Subheading 3.1 for detail). In order to compile and install MACS through source code, Python packages NumPy and SciPy with their header files and GCC compiler are also required.

2.3 Optional Software

In order to map ChIP-Seq reads to reference genome, BWA [10], Bowtie [11], or any other aligner is needed. SAMtools [12] provides useful operations on alignment results, such as merging files, down-sampling, and simple statistics. R environment is needed to generate a PDF image of the model for DNA-binding fragment length prediction. BEDtools [13] software can be used to manipulate MACS output in BED format. Tools to convert MACS signal profiles in bedGraph format into smaller binary bigWig files [14] (especially bedGraphToBigWig tool), or BED files into bigBed format, can be found in UCSC toolkits. Data files can be visualized through Integrative Genomics Viewer (IGV)

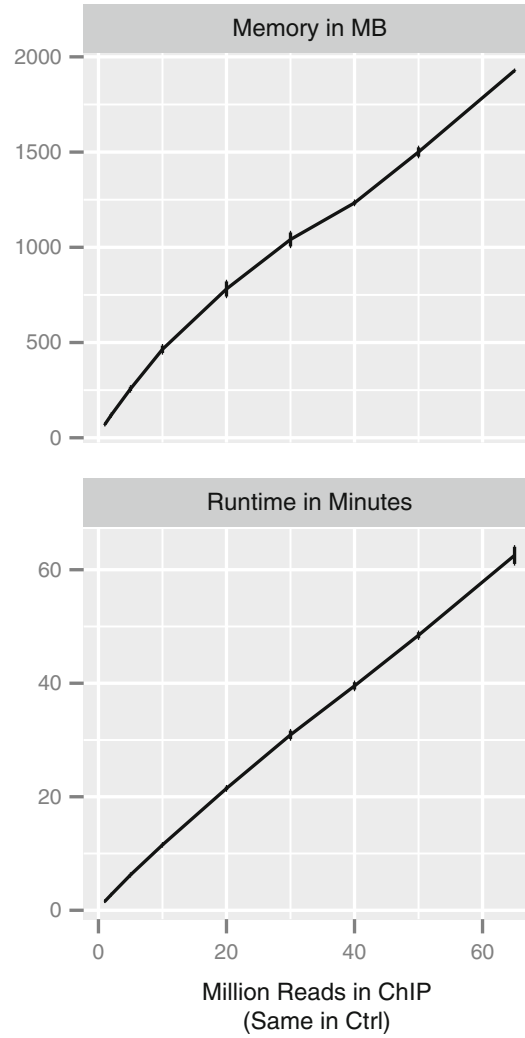


Fig. 1 CPU time and memory usage of MACS. Evaluation is done on a server with 2.8 GHz CPU and 512G mem, through subsampling a deeply sequenced human CTCF dataset from ENCODE (unpublished). The CPU time and memory usage is almost linearly proportional to sequencing depth

[15], a user-friendly interface on a local computer, or the UCSC Genome Browser website through custom tracks or track hub.

2.4 Data

We use NANOG and H3K36me3 ChIP-Seq in human H1 ES cell line, released by ENCODE consortium, as examples in this protocol. The raw sequences in FASTQ format and alignment results (mapped to human hg19/GRCh37 genome) in BAM format can be downloaded through UCSC website at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/ and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/. Each factor has two replicates of ChIP data and two replicates of control.

3 Methods

3.1 Install MACS

1. Set up the necessary operating system and computing environment as listed under Materials. It's recommended to set up a virtual Python environment through VirtualEnv tool (https://pypi.python.org/pypi/virtualenv). After VirtualEnv is downloaded, run these commands (*see* **Note 1**):

```
$ python2.7 virtualenv.py LocalMACS2
$ source LocalMACS2/bin/activate
(LocalMACS2)$
```

2. The easiest way to install MACS is through PyPI (https://pypi.python.org). After **step 1**, "pip" command will be available, while LocalMACS2 environment is entered (a "(LocalMACS2)" before dollar sign). The "pip" command provides a convenient way to install Python packages through Internet. Run these commands to install NumPy, SciPy, and MACS in order (*see* **Note 2**):

```
(LocalMACS2)$ pip install numpy
(LocalMACS2)$ pip install scipy
(LocalMACS2) $ pip install -1
https://pypi.python.org/pypi/MACS2/2.0.10.20130731
```

3. This step and the following **step 4** are provided as alternative, while **step 2** is failed. Make sure NumPy and SciPy are installed in the system. MACS source code can be found on PyPI website (https://pypi.python.org/pypi/MACS2/2.0.10.20130731) (*see* **Note 3**). Click the download button to get the source code package. Right click the download button to copy the URL and then download the package directly to the computer server. Use Unix command "curl" or "wget":

```
(LocalMACS2)$ curl -O
https://pypi.python.org/packages/source/M/MACS2/
MACS2-2.0.10.20130731.tar.gz
or
(LocalMACS2)$ wget
https://pypi.python.org/packages/source/M/MACS2/
MACS2-2.0.10.20130731.tar.gz
```

4. Unpack source code, change the working directory to "MACS2-2.0.10.20130731," and use the standard installation command for Python packages as follows (*see* **Note 4**):

```
(LocalMACS2)$ tar -zxvf MACS2-2.0.10.20130731.tar.gz
(LocalMACS2)$ cd MACS2-2.0.10.20130731
(LocalMACS2)$ python setup.py install
```

5. Now MACS has been installed into an isolated virtual environment named "LocalMACS2." Run "macs2" in command line to see the following information:

```
(LocalMACS2)$ macs2
usage: macs2 [ -h] [ --version]
```

```
{callpeak,diffpeak,bdgpeakcall,bdgbroadcall
,bdgcmp,bdgdiff,filterdup,predictd,pileup,ra
ndsample,refinepeak}
...
macs2: error: too few arguments
```

This means installation is successful (*see* **Note 5**).

3.2 Install Optional Software

6. Download BWA from "http://sourceforge.net/projects/bio-bwa/files/," SAMtools from "http://samtools.sourceforge.net/," R from "http://cran.r-project.org/," BEDTools from "http://code.google.com/p/bedtools/downloads/list," UCSC toolkits (we use bedGraphToBigWig and bedClip in this protocol) from "http://hgdownload.soe.ucsc.edu/admin/exe/," and IGV from "http://www.broadinstitute.org/igv/." Install each software package according to the corresponding instructions (*see* **Note 6**).

3.3 Run MACS to Call Peaks for Point-Source Factor

1. Data we used are two replicates for NANOG ChIP-Seq in human H1 ES cell with corresponding two replicates of control data (reverse cross-link library). Data files in FASTQ format can be downloaded from UCSC ENCODE portal or through the following commands:

```
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/wgEncodeHaibTfbsH1hescNanogsc33759V0416102RawDataRep1.fastq.gz
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/wgEncodeHaibTfbsH1hescNanogsc33759V0416102RawDataRep2.fastq.gz(LocalMACS2)$ wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/wgEncodeHaibTfbsH1hescRxlchV0422111RawDataRep1.fastq.gz
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/wgEncodeHaibTfbsH1hescRxlchV0422111RawDataRep2.fastq.gz
```

The first two files are for ChIP sample, and the later two are controls. Rename these four files as "NANOG_H1hesc_Rep1.fastq.gz," "NANOG_H1hesc_Rep2.fastq.gz," "RXL_H1hesc_Rep1.fastq.gz," and "RXL_H1hesc_Rep2.fastq.gz" accordingly (*see* **Note 7**).

2. Download pre-compiled BWA index file for human reference genome hg19/GRCh37 and unpack it:

```
(LocalMACS2)$ wget http://biomisc.org/data/hg19_bwa.tar.gz
(LocalMACS2)$ tar -zxvf hg19_bwa.tar.gz
```

Several files with prefix "hg19.fa_" can be found in the current working directory:

```
(LocalMACS2)$ ls hg19.fa_*
hg19.fa.amb  hg19.fa.ann  hg19.fa.bwt  hg19.
fa.pac hg19.fa.sa
```

See **Note 8** for detail about making BWA index files for other genome.

3. Run BWA on each FASTQ file to align reads to human hg19 genome. It's not necessary to uncompress the "fastq.gz" files. Here we use one of two ChIP sample file as example. Please replace the file name for other FASTQ files.

```
(LocalMACS2)$ bwa aln -f NANOG_H1hesc_Rep1.
sai  hg19.fa  wgEncodeHaibTfbsH1hescNanog-
sc33759V0416102RawDataRep1.fastq.gz
(LocalMACS2)$ bwa samse -f NANOG_H1hesc_Rep1.
sam hg19.fa NANOG_H1hesc_Rep1.sai
```

Then convert SAM file to BAM file using SAMtools:

```
(LocalMACS2)$ samtools view -bS -o NANOG_
H1hesc_Rep1.bam NANOG_H1hesc_Rep1.sam
```

At aligning all four FASTQ files, four BAM files should exist in the working directory.

```
(LocalMACS2)$ ls *.bam
NANOG_H1hesc_Rep1.bam  NANOG_H1hesc_Rep2.bam
RXL_H1hesc_Rep1.bam RXL_H1hesc_Rep2.bam
```

See **Note 9**.

4. Call NANOG peaks by using MACS callpeak module. Run the following command:

```
(LocalMACS2)$ macs2 callpeak --call-summits -B -g hs -t
NANOG_H1hesc_Rep1.bam NANOG_H1hesc_Rep2.bam -c
RXL_H1hesc_Rep1.bam  RXL_H1hesc_Rep2.bam  -n
NANOG_H1hesc_macs
```

See Table 2 for meanings of the options. *See* **Note 10** for suggestions on parameter tweaking.

5. Now two BED style output files, a XLS spreadsheet file, and two bedGraph files with prefix "NANOG_H1hesc_macs" can be found in working directory:

```
(LocalMACS2)$ ls NANOG_H1hesc_macs*
NANOG_H1hesc_macs_peaks.narrowPeak
NANOG_H1hesc_macs_summits.bed
NANOG_H1hesc_macs_peaks.xls
NANOG_H1hesc_macs_treat_pileup.bdg
NANOG_H1hesc_macs_control_lambda.bdg
```

See **Note 11** for brief description of each file.

Table 2
Common options in MACS

-t	Specify the file name for the ChIP-seq sample read alignment. MACScan automatically detect file formats and can use gzip compressed files directly. Various file formats are supported, such as SAM, BAM, BED, ELAND,and Bowtie. While multiple file names are given to -t option, such as theabove example, they will be pooled together
-c	Specify the file name for the control sample read alignment. Same as -t,multiple file names can be given together
-g	Specify the effective genome size, which is the approximate mappable genome size. Parameter can be shortcut such as "hg" for human, "mm" for mouse,or number such as 2.7e9 for human. While -g is not specified, defaultparameter is "hg," which means "-g hs" in the above example can be omitted
-n	Specify the name of this MACS run. The parameter will be applied as prefixto output file names and as prefix to peak name identifier
-B	Turn on generating signal files in bedGraph format containing the extendedread pileup and chromatin local bias at every base pair. This option is requiredin order to compute noise-deducted signal tracks. MACS will run much fasterwhile this option is off
--call-summits	Turn on a post-processing module in MACS which will smooth the signalswithin each predicted peak region and then find all possible subpeak summits.This option is extremely useful to deconvolve nearby transcription factorbinding sites
--broad	Turn on broad region calling for MACS. MACS will call weaker andstronger regions separately and then group them
-p or -q	Set the cutoff for peak calling. -p sets *p*-value cutoff based on local Poisson test,whereas -q sets *q*-value (FDR) cutoff through additional Benjamini–Hochbergprocess on *p*-values. -p and -q are mutually exclusive
--broad-cutoff	Cutoff for broad region calling. It will be used for identifying weakerand broad domains, and -p or -q will be used to identify strongerand narrow domains
--nomodel	Turn off prediction on DNA-binding fragment length
--extsize	With --nomodel on, an arbitrary value will be used to extend eachread toward 3′ end
--SPMR	Save signal per million reads for extended read pileup and localbias while -B is turned on

3.4 Run MACS to Call Regions for Broad Mark

1. Data we used are two replicates for histone mark H3K36me3 ChIP-Seq in human H1 ES cell with corresponding two replicates of control data. Data files in FASTQ format can be downloaded from UCSC ENCODE portal or through the following commands (*see* **Note 7**):

```
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescControlStdRawDataRep1.fastq.gz
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescControlStdRawDataRep2.fastq.gz
```

```
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescH3k36me3StdRawDataRep1.fastq.gz
(LocalMACS2)$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescH3k36me3StdRawDataRep2.fastq.gz
```

The first two files are for ChIP sample, and the later two are controls. Rename these four files as "H3K36me3_H1hesc_Rep1.fastq.gz," "H3K36me3_H1hesc_Rep2.fastq.gz," "Control_H1hesc_Rep1.fastq.gz," and "Control_H1hesc_Rep2.fastq.gz" accordingly.

2. Follow the **steps 2** and **3** as in Subheading 3.3 to map FASTQ files, and get the following BAM files.

```
(LocalMACS2)$ ls *.bam
H3K36me3_H1hesc_Rep1.bam    H3K36me3_H1hesc_
Rep2.bam  Control_H1hesc_Rep1.bam  Control_
H1hesc_Rep2.bam
```

3. Call H3K36me3 region by using MACS callpeak module with --broad option. Run the following command:

```
(LocalMACS2)$ macs2 callpeak --broad -B -g hs -t
H3K36me3_H1hesc_Rep1.bam    H3K36me3_H1hesc_
Rep2.bam -c
Control_H1hesc_Rep1.bam Control_H1hesc_Rep2.bam -n
NANOG_H1hesc_macs
```

The only difference between this step and **step 5** of Subheading 3.3 is that "--broad" replaces "--call-summits." *See* Table 2 for meanings of the options. *See* **Note 12** for parameter tweaking.

4. Now two BED style output file, a XLS spreadsheet, and two bedGraph files with prefix "H3K36me3_H1hesc_macs" can be found in working directory:

```
(LocalMACS2)$ ls H3K36me3_H1hesc_macs*
H3K36me3_H1hesc_macs_peaks.broadPeak
H3K36me3_H1hesc_macs_peaks.gappedPeak
H3K36me3_H1hesc_macs_peaks.xls
H3K36me3_H1hesc_macs_treat_pileup.bdg
H3K36me3_H1hesc_macs_control_lambda.bdg
```

See **Note 13** for brief description of each file.

3.5 Generate Noise-Deducted Signal Tracks

1. For each pair of "treat_pileup.bdg" and "control_lambda.bdg" files, MACS bdgcmp module can be used to compute noise-deducted signal track file, which is better for visualization or comparison across different libraries. Locate the following files in pairs that are generated in Subheadings 3.3 and 3.4:

```
NANOG_H1hesc_macs_treat_pileup.bdg
NANOG_H1hesc_macs_control_lambda.bdg
```

```
H3K36me3_H1hesc_macs_treat_pileup.bdg
H3K36me3_H1hesc_macs_control_lambda.bdg
```

2. Run the following command to compute fold-enrichment and log10 likelihood ratio tracks:

```
(LocalMACS2)$ macs2 bdgcmp -t
NANOG_H1hesc_macs_treat_pileup.bdg -c
NANOG_H1hesc_macs_control_lambda.bdg -o NANOG_H1hesc_macs
-m FE logLR -p 0.001
(LocalMACS2)$ macs2 bdgcmp -t
H3K36me3_H1hesc_macs_treat_pileup.bdg -c
H3K36me3_H1hesc_macs_control_lambda.bdg -o
H3K36me3_H1hesc_macs -m FE logLR -p 0.001
```

These commands will generate "NANOG_H1hesc_macs_FE.bdg" and "H3K36me3_H1hesc_macs_FE.bdg" which contain fold-enrichment (ChIP pileup signal divided by local bias) values for every base pairs along entire genome and "NANOG_H1hesc_macs_logLR.bdg" and "H3K36me3_H1hesc_macs_logLR.bdg" which contain log10 likelihood ratios between two Poisson models for ChIP and control (*see* **Notes 14** and **15**).

3.6 Visualization

1. Firstly, we need to convert MACS output files in genome browser-friendly formats. The outputs in narrowPeak, broadPeak, gappedPeak, and general BED format are of small file size and so can be easily opened and visualized in UCSC Genome Browser or IGV. On the contrary, the outputs in bedGraph are usually huge and hard to handle. We suggest to convert bedGraph to bigWig format, which is a binary file format and was specifically optimized for visualization and data transfer. Download a file containing human chromosome length to working directory:

```
(LocalMACS2)$ wget http://biomisc.org/data/hg19.len
```

Run the following commands to convert a bedGraph file (using NANOG_H1hesc_macs_FE.bdg as example):

```
(LocalMACS2)$ bedGraphToBigWig NANOG_H1hesc_
macs_FE.bdg hg19.len NANOG_H1hesc_macs_FE.bw
```

See **Note 16** on how to fix incorrect coordinates in bedGraph files.

2. BigWig files can be directly loaded by IGV, following IGV manuals.
3. An alternative to **step 2** is to use UCSC Genome Browser for visualization. Upload BED style files and bigWig files to a website so that each file can be accessed through a URL. We use the example "http://biomisc.org/data/NANOG_H1hesc_macs_FE.bw" as URL for the bigWig file of NANOG

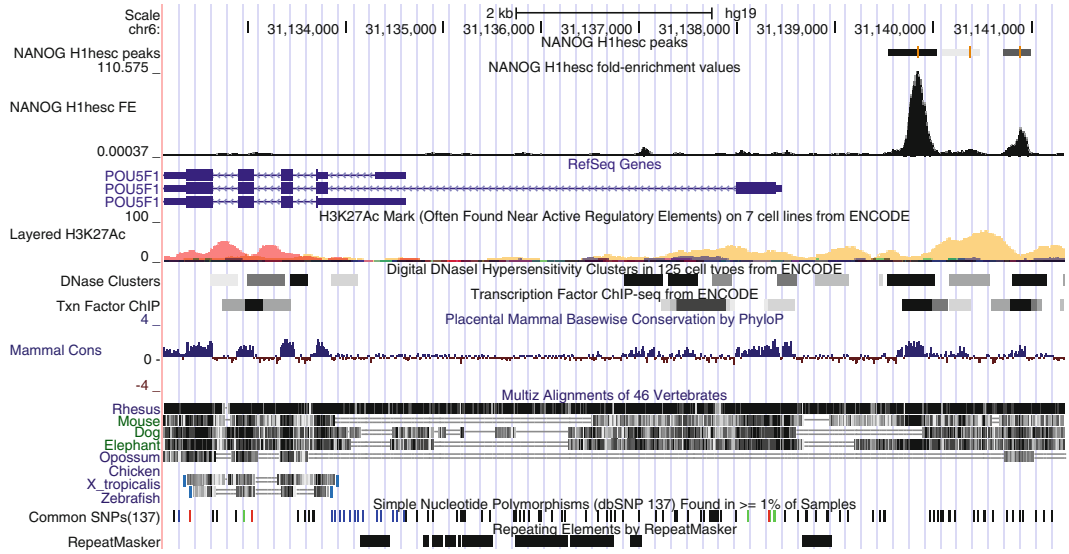


Fig. 2 Visualization through UCSC Genome Browser. The *top first track* shows the NANOG binding sites detected by MACS, where the *orange vertical lines* mark the peak summits. The *second track* shows the NANOG ChIP-seq fold-enrichment values

fold-enrichment values. While adding custom tracks to UCSC Genome Browser, enter the following information in the form on webpage and then click "submit" button (Fig. 2):

```
track type = bigWig name = "NANOG H1hesc FE"
description = "NANOG H1hesc fold-enrichment
values" bigDataUrl=http://biomisc.org/data/NANOG_H1hesc_
macs_FE.bw
```

4 Notes

1. After the virtual environment is successfully created and activated, a prompt string starting with "LocalMACS2" can be seen before dollar sign. This isolated clean Python environment named "LocalMACS2" can be fully controlled by users without help from system administrators. All library files and executables should be installed into the directory "LocalMACS2/" in future. If proper NumPy and SciPy have been installed system-wisely in the machine, add "--system-site-packages" option while creating virtual environment. In this way, the virtual environment can inherit system-wise Python libraries, to save the effort to install NumPy and SciPy separately.

2. If NumPy or SciPy fails to be installed through "pip," check their corresponding website on how to install them through source code packages and then try to use "pip" command to install MACS again.
3. Always check MACS page on PyPI (https://pypi.python.org/pypi/MACS2) for download link of the most recent release.
4. GCC compiler is necessary for installing MACS. If there are errors during installation, take the following actions:
   (a) Check whether Python, NumPy, and SciPy header files have been installed. If they are installed through "pip" or source code, their header files should be installed correctly. If system-wise Python, NumPy, and SciPy libraries have been installed through certain software package management system (examples are dpkg for Debian Linux or rpm for Redhat Linux), make sure packages "python-dev," "numpy-dev" and "scipy-dev" are installed as well. Then try again.
   (b) Install "Cython" (http://cython.org), remove all C files with suffix ".c" in source code folder, and then use "setup_w_cython.py" instead of "setup.py":

```
(LocalMACS2)$ find . -name '*.c' -exec rm {} +
(LocalMACS2)$ python setup_w_cython.py install
```

5. Never install MACS into source code directory. If MACS source code is unpacked into a folder named "MACS2-2.0.10.20130731," to install MACS through "python setup.py install --prefix MACS2-2.0.10.20130731" will cause unexpected errors. Using virtual environment to avoid any confusion is the most recommended.
6. Make sure the optional software can be executed in the shell of virtual environment. Specifically in our case, executable "bwa," "samtools," "bedtools," "bedGraphToBigWig," and "bedClip" should be placed in the directory "LocalMACS2/bin/."
7. UCSC ENCODE portal also provides BAM format alignment results for corresponding FASTQ files. Therefore, alternatively, replace the name "RawDataRep1.fastq.gz" or "RawDataRep2.fastq.gz" with "AlnRep1.bam" and "AlnRep2.bam" in the commands of **step 1** of Subheadings 3.3 and 3.4, download BAM files, and then skip alignment steps.
8. To make BWA index for a certain genome, download the genome sequence in FASTA format first. Go to UCSC Genome Browser download page for genome sequence files of common species. Then use the following command:

```
(LocalMACS2)$ bwa index genome.fa
```

   This will generate the following files that can be used in reads alignment steps of Subheadings 3.3 and 3.4: genome.

fa.amb, genome.fa.ann, genome.fa.bwt, genome.fa.pac, and genome.fa.sa.

9. By default, BWA will randomly pick an alignment if there are multiple best alignments for the same read (i.e., multiple mapping read). If necessary, please refer to BWA and SAMTools manuals on how to filter out multiple mapping reads using tags in SAM/BAM files.
10. Increase or decrease the "-p" or "-q" parameter to call stronger or weaker peaks. Pay attention to MACS runtime message on the binding fragment length prediction (a message as "#2 predicted fragment length is … bps"). A typical value should be around 200 bp. If the value is abnormally small or large, set "--nomodel --extsize …" to bypass the prediction. While comparing different libraries on the same ChIPed factor, to set a uniform fragment length through "--nomodel --extsize" is highly recommended.
11. The "peaks.narrowPeak" file is in UCSC BED6 + 4 format. It contains information about peak boundaries, summits, fold enrichments, *p*-values, and *q*-values. The "summits.bed" file contains the 1 bp location of each peak summit, so it's suitable for searching TF-specific DNA-binding motifs. The "peaks.xls " file is a tab-separated values file, ready to be loaded into spreadsheet software. The "treat_pileup.bdg" and "control_lambda.bdg" files are in bedGraph format and can be used for visualization or post-processing, such as generating noise-deducted signal tracks or predicting differential binding sites.
12. MACS will call two levels of regions—stronger but narrower regions controlled by "-p" or "-q" option and weaker but broader regions controlled by "-broad-cutoff" option (whether this is a *p*-value cutoff or *q*-value cutoff depends on whether "-p" or "-q" is in use). Although we recommend using the DNA-binding fragment length prediction in MACS, it's safe to use "--nomodel --extsize 147" for most histone mark ChIP-Seq data, where 147 corresponds to the length of DNA wrapped on a nucleosome.
13. The "peaks.broadPeak" file is in UCSC BED6 + 3 format. It contains information about peak boundaries, fold enrichments, *p*-values, and *q*-values. The "peaks.gappedPeak" file is in BED12 + 3 format which contains two levels of region calls linked in a structure like exons within genes: a weaker call to mark broadly spreading enrichment pattern and a stringent call to mark local sharp enrichment pattern. The "peaks.xls," "treat_pileup.bdg," and "control_lambda.bdg" files are similar to those described in **Note 10**.

14. MACS "bdgcmp" module can generate multiple types of noise-deducted signal tracks simultaneously. Simply append names of scoring method after "-m" option. Acceptable methods are "ppois" for Poisson *p*-values, "qpois" for Poisson *q*-values, "subtract" for subtraction between ChIP and control, "FE" for fold enrichment, "logFE" for log10 fold-enrichment, and "logLR" for log10 likelihood ratio. The "-p" option for "bdgcmp" is used to specify a pseudo-count for scoring methods. It should be used for logarithm-type methods to avoid domain errors.
15. Fold-enrichment tracks are easier to interpret; however, they are unstable, while both ChIP and control signals are low. Log10 likelihood ratios have more power to reduce background noise statistically; however, it is difficult to link them with biology.
16. To fix incorrect coordinates in bedGraph or BED file, use slop function in BEDTools and bedClip tool from UCSC. Run:

```
(LocalMACS2)$ bedtools slop -i NANOG_H1hesc_macs_FE.bdg -g hg19.len -b 0 | bedClip stdin hg19.len NANOG_H1hesc_macs_FE.bdg.fixed
```

This command will fix any illegal coordinates in bedGraph file that extend over chromosome boundaries. Rename ".bdg.fixed" as ".bdg" when finished.

Acknowledgment

This work is supported by Startup funds from University at Buffalo.

References

1. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein–DNA interactions. Science 316:1497–1502
2. Barski A et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
3. Robertson G et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657
4. Mikkelsen TS et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560
5. ENCODE Project Consortium et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
6. Bernstein BE et al (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28:1045–1048
7. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-Seq and RNA-seq studies. Nat Methods 6:S22–S32
8. Zhang Y et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
9. Landt SG et al (2012) ChIP-seq guidelines and practices used by the ENCODE and modENCODE consortia. Genome Res 22:1813–1831
10. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
11. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

12. Li H et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079
13. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
14. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26:2204–2207
15. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics 14: 178–192

Chapter 5

Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells

Shiliyang Xu, Sean Grullon, Kai Ge, and Weiqun Peng

Abstract

Chromatin states are the key to embryonic stem cell pluripotency and differentiation. Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-Seq) is increasingly used to map chromatin states and to functionally annotate the genome. Many ChIP-Seq profiles, especially those of histone methylations, are noisy and diffuse. Here we describe SICER (Zang et al., Bioinformatics 25(15):1952–1958, 2009), an algorithm specifically designed to identify disperse ChIP-enriched regions with high sensitivity and specificity. This algorithm has found a lot of applications in epigenomic studies. In this Chapter, we will demonstrate in detail how to run SICER to delineate ChIP-enriched regions and assess their statistical significance, and to identify regions of differential enrichment when two chromatin states are compared.

Key words ChIP-Seq, Histone modifications, Epigenetic modifications, Epigenome, SICER

1 Introduction

Chromatin structure plays a critical role in embryonic stem cell pluripotency and differentiation. Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-Seq) is now widely used to quantify chromatin states across genomes. Large amount of ChIP-Seq datasets of genome-wide profiling of epigenetic modifications and chromatin-binding proteins have been generated. The distribution of ChIP-Seq signals has been found to vary widely, ranging from a few nucleosomes to large chromatin domains encompassing multiple genes. For example, H3K4me2 and H3K4me3, which are usually associated with enhancers and promoters, tend to exhibit relatively localized sharp peaks [1, 2]. On the other hand, H3K36me3, a hallmark of elongation, or repressive mark H3K27me3 may span tens or even hundreds of kilo bases (for an example, please *see* Fig. 1).

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_5, © Springer Science+Business Media New York 2014

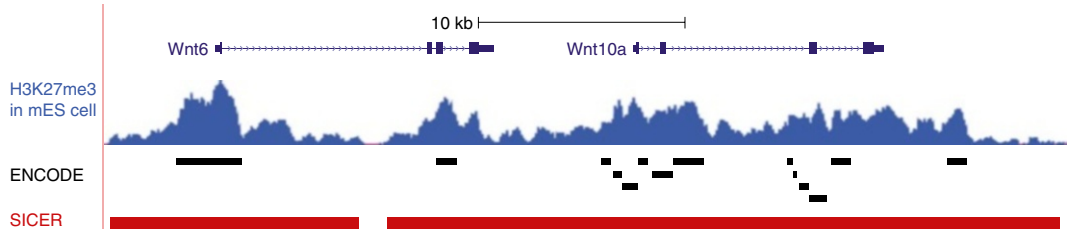


Fig. 1 SICER with default parameters identifies extended domain with H3K27me3 enrichment in mouse embryonic stem cell across Wnt6 and Wnt10a gene locus. *Top track*: H3K27me3 ChIP-Seq data from ENCODE [14]. *Middle track*: H3K27me3 enrichment peaks identified in ENCODE. *Bottom track*: H3K27me3 enrichment domains identified by SICER [4] with default parameters

Diffuse signals can be observed in many libraries. In addition to chromatin modifications, some histone-modifying enzymes [3], chromatin remodeling complexes, and RNA Pol II also exhibit extended domains of enrichment. Because the detection of diffuse signals often suffers from high noise level and lack of saturation in sequencing coverage, it is a challenging task to identify statistically significant ChIP-enriched domains. These generally weak and diffuse signals render approaches seeking strong local enrichment, such as those peak-finding algorithms designed for finding transcription factor (TF) binding sites, inadequate. Toward this end, we developed SICER (Statistical model for Identification of ChIP-Enriched Regions) [4], which achieves high sensitivity and specificity by identifying spatial clusters of ChIP-enriched signals that are unlikely to appear in a background model. As demonstrated in Fig. 1, SICER is able to identify extended domains of ChIP enrichment. Although SICER was designed for analyzing ChIP-Seq data with extended enrichment profile, upon a proper choice of parameters, it could also be applied to ChIP-Seq data with sharp peaks like those for transcription factors.

2 Overview of SICER

2.1 Motivation

Classic examples of the mechanism of domain formation of histone modifications include H3K9me3 in yeast. H3K9me3 recruits HP1, which in turn recruits H3K9 methyltransferase Suv39h. Suv39h modifies H3K9 on other nucleosomes in the vicinity, thereby self-propagating the heterochromatin state [5–7]. Another example is H3K27me3. This mark is deposited by the polycomb complex, PRC2, and is believed to recruit the PRC1 complex [8]. In Drosophila, it has been suggested that the looping action of PRC1 and PRC2 that both anchor at the polycomb response elements results in the spreading of H3K27me3 [8]. Inspired by the

mechanisms of domain formation illustrated by these examples, SICER [4] regards significant spatial clusters of ChIP enrichment as true signal. The main feature of the SICER algorithm is to pool together signals from nearby nucleosomes that are in the same modification state for identification of statistically significant enrichments. For ChIP-Seq libraries with diffuse profile, this feature alleviates the problem of lack of saturation and markedly improves the signal-to-noise ratio, where at any short scale of one or several nucleosomes, the ChIP enrichment does not appear to be significant enough. This approach also enables a systematic evaluation of statistical significance of identified ChIP-enriched regions against control library when available.

2.2 Algorithm

The key concept that SICER uses to capture spatial clustering of reads is island. To delineate the islands and assess the statistical significance of ChIP enrichment on them, SICER [4] carry out the following steps: (1) It partitions the genome into nonoverlapping windows of size w. (2) It identifies windows with enrichment (i.e., "eligible" windows). A window is deemed "eligible" ("ineligible") if the number of ChIP-Seq reads in this window is equal to or above (below) a read count threshold l_0. The threshold l_0 is determined based on a Poisson distribution $\sum_{l_0}^{\infty} P(l,\lambda) \le p_0$. $\lambda = wN/L$ is the average number of reads in a window. N is the number of reads in the library and L the effective genome length (further discussion on L can be found in Subheading 2.3.3). The threshold p_0 is defaulted to be 0.2 so that all windows with reasonable ChIP enrichment are "eligible." (3) It identifies islands as clusters of "eligible" windows separated by gaps of size no larger than a predetermined value g. A gap is a contiguous stretch of "ineligible" windows between two neighboring "eligible" windows. When $g = 0$, islands are uninterrupted clusters of "eligible" windows. *See* Fig. 2 for an illustration of the definition of islands. (4) It identifies "candidate" islands that exhibit significant clustering of "eligible" windows that are unlikely to appear by chance. SICER assigns a score $s(l)$ for each "eligible" window of read count l as $s(l) = -\ln P(l,\lambda)$. The score S for each island is defined as the aggregated score of all "eligible" windows in the island. Only islands with score $S > S_T$ are regarded as "candidate" islands, where S_T is an island-score threshold controlling statistical significance of ChIP enrichment on an island against random background. More specifically, S_T is determined by requiring that the expected number of islands with scores above S_T if reads are randomly distributed be less than an E-value threshold e. (5) If a control library is available, SICER will further filter the "candidate" islands using the control library, retaining only those that exhibit significant enrichment of ChIP signal compared to control on the islands. The statistical significance versus control library is

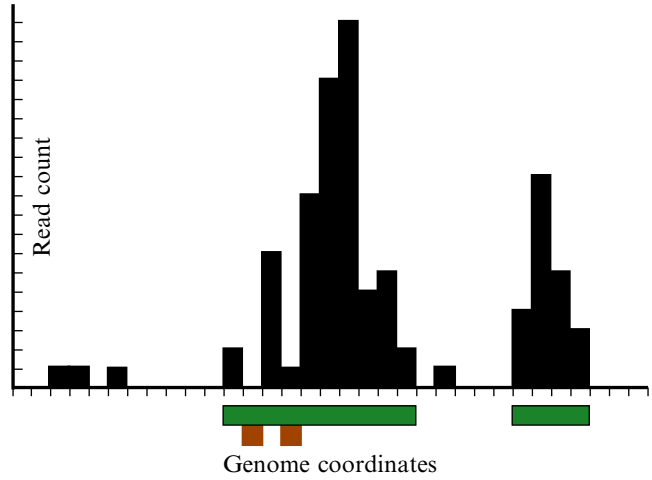


Fig. 2 Schematic illustration of definition of islands. Shown is a segment of a genomic landscape of ChIP-Seq reads. The *x*-axis denotes the genome coordinates, where each interval represents a window. The *y*-axis denotes the read count. Each black vertical bar represents the read count in the respective window. The regions underlined by the green horizontal bars are the two identified islands under $g = 1$ and $I_0 = 2$. The two windows underlined by brown boxes are gaps in the first island

characterized by a *p*-value based on Poisson distribution. A false discovery rate (FDR) is also reported using *p*-value adjusted for multiple testing [9]. Because of the presence of systematic biases in a typical ChIP-Seq library, it is highly desirable to have a matching control library.

The flow chart of SICER is shown in Fig. 3. SICER is essentially a filtering tool. The delineated ChIP-enriched regions can be used to associate with other genomic landmarks. Reads on those ChIP-enriched regions can be identified and used for profiling and other quantitative analysis. Further details of the SICER algorithm can be found in [4].

2.3 Considerations for Key Parameters

Choices of key parameters are important for satisfactory results. Of particular importance are window size w, gap size g, effective genome length L, as well as a parameter controlling statistical significance (*E*-value for random background and *p*-value or FDR for using a control library as background). Before a discussion of considerations on choices of these parameters, we would like to emphasize that visual examination on the genome browser is indispensable, and positive and negative control cases from known biology would be of tremendous help in making the appropriate choice of the parameters.

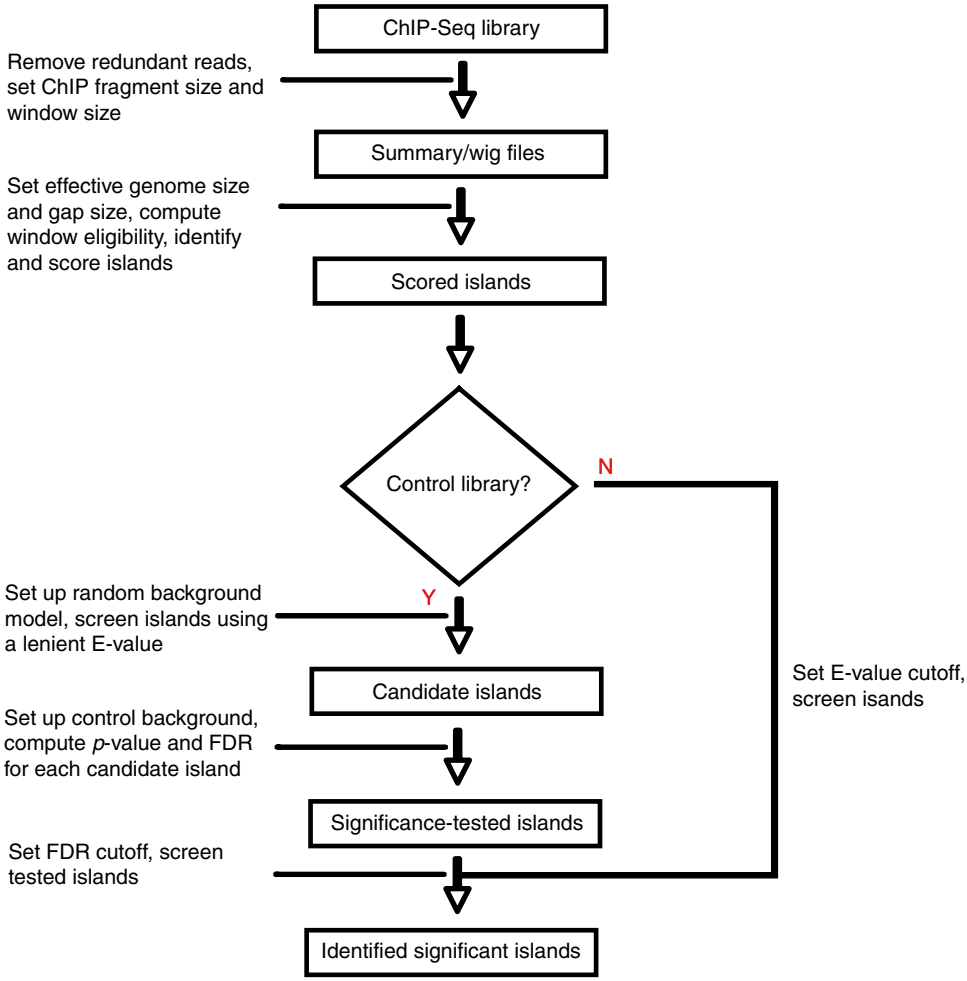


Fig. 3 Complete SICER flowchart. For ChIP-Seq library with or without control library, SICER always uses a random background model to identify candidate islands subject to a preset *E*-value. If a control library is available, SICER further screens islands using a false discovery rate (FDR) or a *p*-value cutoff against control library

2.3.1 *Window Size*

The choice of the window size, which directly affects the delineation of islands, is an important one. A window too narrow will exaggerate local fluctuation in each window, while a window too large will cause over-smoothing of data and lose resolution. Our experience has been that for transcription factors, a suitable window size choice is around 50–100 bps. On the other hand, for histone modifications and histone variants, a typical choice for window size w is 200 bps, a number approximately the length of a single nucleosome and a linker. As an example, we tested various window sizes (50, 100, 200, 500, and 1,000 bps) with a fixed gap size (3 windows) on the H3K27me3 dataset, and the resulting islands identified were shown in Fig. 4. It is clear that a larger window size results in more extended islands. For this particular dataset on this

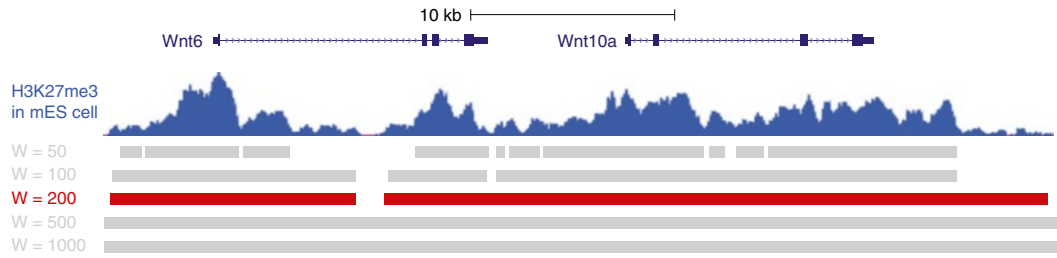


Fig. 4 The choice of the window size directly affects the delineation of islands. *Top track*: H3K27me3 ChIP-Seq data from ENCODE. *Bottom tracks*: islands with significant H3K27me3 enrichment identified by SICER with different choices of window size: (from *top* to *bottom*, in bps) 50, 100, 200, 500, 1,000. The gap size is always set to be 3 windows

locus, a window size of 200 bps appears to have a good balance of specificity (that didn't include too many regions of weak enrichment) and sensitivity (that didn't produce too many gaps within an extended island) and thus an appropriate choice. In general, we can estimate the window size using the approach developed by Shimazaki and Shinomoto [10, 11], which employed a cost function defined by the mean integrated squared error to find an optimized window size for a histogram. This approach cannot be used blindly. Although the automatically calculated window size results in improved island delineation in many cases, in some other cases it fails to output a reasonable value.

2.3.2 Gap Size

The adoption of gap reflects the unique strength of SICER in identifying broad ChIP enrichment from poor coverage and/or high background noises. Gap size g by definition must be a multiple of window size chosen. In general the wider the domains are, the larger the gap size should be. For instance, for localized histone modifications like H3K4me3, the gap size can be set to be equal to the window size, $g = w$, while for a histone modification with an extended profile (e.g., H3K27me3), $g = 3w$ likely works better. For more careful consideration, users can plot the aggregate score of all significant islands as a function of g. If the aggregate score reaches a maximum inside of the range of g explored, the gap size corresponding to the highest aggregate score would be a good choice. On the other hand, the aggregate score may increase monotonically with gap size (*see* Fig. 5 for an example). If the curve gradually increases toward saturation, we suggest choosing the gap size so that the corresponding aggregate score is sufficiently close to saturation. No sign of saturation suggests poor sequencing coverage (*see* Fig. 5), and the ultimate solution would be to increase sequencing depth. Another option in the absence of enough sequencing depth is to increase window size, the discussion of which can be found in the previous subsection. In general,

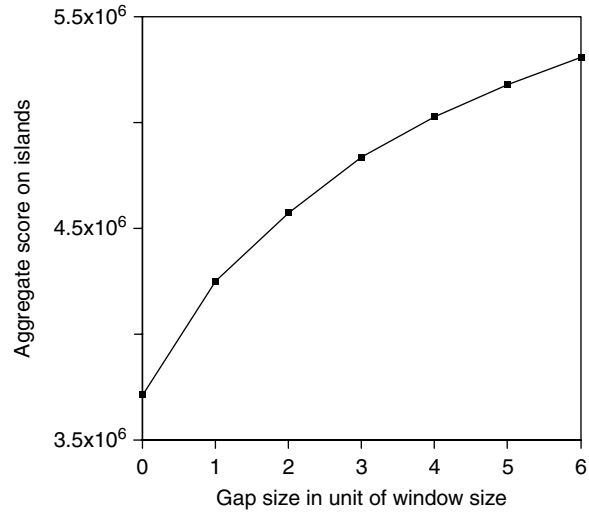


Fig. 5 Aggregate score on islands as a function of gap size. The window size is fixed at 200 bps. The H3K27me3 ChIP-Seq data exhibits monotonically increasing aggregate score with increasing gap size

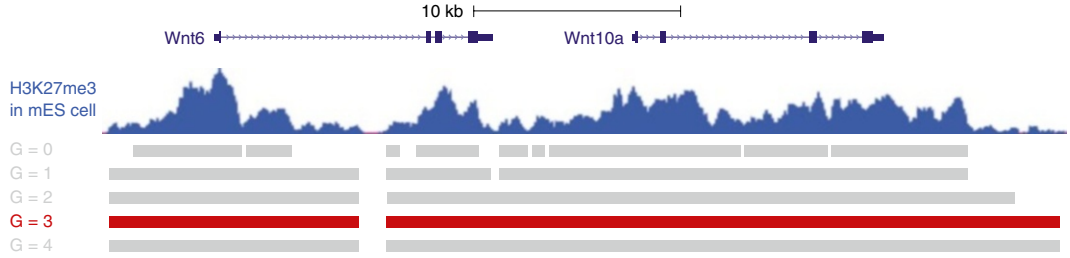


Fig. 6 The choice of the gap size directly affects the delineation of islands. *Top track*: H3K27me3 ChIP-Seq data from ENCODE. *Bottom tracks*: islands with significant H3K27me3 enrichment identified by SICER with different choices of gap size: (from top to bottom, in windows) 0, 1, 2, 3, 4. Window size is fixed at 200 bps

we recommend against a gap size beyond 4 windows, for fear of too much spurious clustering. Figure 6 shows the effect of gap size on the H3K27me3 island delineation at the Wnt6 and Wnt10 locus, in which case gap sizes of 2, 3, and 4 windows provide adequate results.

2.3.3 Effective Genome Size and Effective Genome Fraction

When short reads are mapped onto the reference genome, normally only those that map to unique genomic loci are retained for further analysis. As a result, genomic regions with degenerate sequences or sequences composed of character "N" are non-mappable. The effective genome length L is defined as the total length of mappable regions in the genome. The effective genome fraction is defined as L divided by the actual genome length. L depends on the species and sequencing protocol (e.g., read length, paired end, or single end).

Generally speaking, longer read length and paired-ended sequencing will lead to higher fraction of effective genome. *L* can be found or computed from Uniqueome [12].

2.3.4 Statistical Significance

In case of random background, an *E*-value cutoff is used to identify significant islands. *E*-value is the expected number of islands emerged merely due to local fluctuation from randomly distributed reads along the genome. A smaller *E*-value means higher stringency. In the case of random background, one can give a rough estimate of error rate empirically by dividing the *E*-value by the total number of significant islands identified. For example, if *E*-value is 500, the number of significant islands is 10,000, the empirical error rate is 5 %. If a control library is available, SICER uses a default permissive *E*-value of 1,000 in identifying candidate islands prior to incorporating control library information. SICER then computes *p*-value for each candidate island based on a Poisson distribution against read count in the control library and considers multiple-testing correction and uses FDR for statistical significance assessment. An FDR threshold of 0.01 or 0.001 is in general adequate, while an FDR of 10^{-8} or less can be used to find the high-confidence ChIP-enriched regions.

3 Material and Method

In this section, we demonstrate how to run SICER and understand SICER output by going through a concrete example, where we apply SICER in the most typical situation: a ChIP-Seq library along with a control library.

3.1 Material

Commands listed in the current manual are executed and time-benchmarked in the following system environment: Mac OS X 10.6.8, dual 2.93 GHz 6-core Intel Xeon processors, 64GB Memory, 64-bit Python 2.7.3 with NumPy, SciPy and PyLab packages, and BEDTools [13] installed (*see* **Note 1**).

The dataset analyzed is the ChIP-Seq data of H3K27me3 in mouse embryonic stem cell (ES Bruce4) from the mouse ENCODE project [14], downloaded from the UCSC genome browser [15]. There are 2 replicates of H3K27me3 ChIP-Seq libraries, together with 2 replicates of input control library. For simplicity the replicates are pooled together, yielding the ChIP library ES_H3K27m3.bed with 53 million reads and the control library ES_input.bed with 22 million reads.

The most recent release of SICER (Version 1.1) was downloaded from http://home.gwu.edu/~wpeng/Software.htm.

3.2 Method

3.2.1 Installation of SICER

After downloading SICER software package SICER_V1.1.tgz, the user shall launch Terminal, go to the directory where the downloaded file is, and then execute

$ tar -xvf SICER_V1.1.tgz -C /mydir

Here /mydir represents the directory where the user desires to have SICER installed.

3.2.2 Setting Paths Inside Master Scripts

The master scripts need to be customized to reflect the directory where SICER is located. Use a plain text editor (not rich format editors like Microsoft Word) to open the script SICER.sh, and change the first line right below the shebang (line starts with "#!") and comment box (lines start with "#").

PATHTO=/home/data/SICER1.1

to

PATHTO=/mydir/SICER_V1.1

Repeat this for the other main script, SICER-rb.sh. For SICER-df.sh and SICER-df-rb.sh, replace the first line below the comment box to

SICER=/mydir/SICER_V1.1/SICER

3.2.3 Preparation of Data Files

In ENCODE project database, ChIP-Seq data is available in various formats. For simplicity we use the pre-aligned BAM data. After downloading BAM files and accompanying index BAM.BAI files, change working directory to where the files are stored and execute

$ bamToBed -i wgEncodeLicrHistoneEsb4H3k27me3ME0C57bl6StdAlnRep1.bam>ES_H3K27me3_rep1.bed

$ bamToBed -i wgEncodeLicrHistoneEsb4H3k27me3ME0C57bl6StdAlnRep2.bam>ES_H3K27me3_rep2.bed

$ bamToBed -i wgEncodeLicrHistoneEsb4InputME0C57bl6StdAlnRep1.bam>ES_input_rep1.bed

$ bamToBed -i wgEncodeLicrHistoneEsb4InputME0C57bl6StdAlnRep2.bam>ES_input_rep2.bed

Then combine the 2 replicates in each ChIP-Seq sample via

$ cat ES_H3K27me3_rep1.bed ES_H3K27me3_rep2.bed> ES_H3K27me3.bed

$ cat ES_input_rep1.bed ES_input_rep2.bed>ES_input.bed

3.2.4 Execution of SICER

Once we have BED files of the ChIP-Seq library and control library, we can start the SICER analysis. Within the SICER directory, there are four SICER scripts, namely, SICER.sh, SICER-rb.sh, SICER-df.sh, and SICER-df-rb.sh. Among those SICER.sh is the master SICER script that processes ChIP library against a control library. If no control library is available, SICER-rb.sh can be executed in place of SICER.sh. SICER-df.sh and SICER-df-rb.sh are used to compare two epigenomes and will be discussed later (*see* **Notes 2–4**).

In Terminal, from where the ChIP and control libraries (ES_H3K27me3.bed and ES_input.bed in this example) are stored, launch SICER with (*see* **Note 2**)

```
$ sh /mydir/SICER_V1.1/SICER/SICER.sh.ES_H3K27me3.bed ES_input.bed.mm9 1 200 150 0.8 600 1e-3
```

SICER.sh takes 11 ordered command line parameters. The general command structure is

```
$ sh /mydir/SICER_V1.1/SICER/SICER.sh [Input directory] [ChIP file] [Control file] [Output directory] [Species] [Redundancy threshold] [Window size] [Fragment size] [Effective genome fraction] [Gap size] [FDR]
```

The detailed description of the command line arguments are listed below:

1. Input directory: where the ChIP and control libraries data files are stored. In this example, "." denotes current directory.
2. ChIP file: the file name of the ChIP library. In this example it is ES_H3K27me3.bed.
3. Control file: the file name of the control library. In this example it is ES_input.bed.
4. Output directory: where the output files should be stored. In this example, "." denotes current directory.
5. Species: the name of reference genome (*see* **Note 3**). In this example it is "mm9."
6. Redundancy threshold: number of redundant reads kept for analysis. In this example it is 1. Redundant reads refer to reads with exactly the same genomic location and orientation. For typical ChIP-Seq datasets, this is likely due to PCR amplification artifact. To remove this potential bias, we generally recommend removing the redundancy and retaining only 1 read for each set of redundant reads.
7. Window size: the width (in bps) of window in comparing ChIP with control library. In this example it is 200 as the default value recommended for histone modification marks.
8. Fragment size: the average size (in bps) of ChIP fragment. In this example it is 150 as the default recommended value. This parameter is used to assign a ChIP read to the center of the DNA fragment. Typical sonication outputs ChIP fragment of 150–300 bps long.
9. Effective genome fraction: In this example it is 0.8, which is recommended for single-end ChIP-Seq with 50 bp read length.
10. Gap size: the gap size (in bps) allowed in SICER filtering. In this example it is 600 bps (or 3 windows), as recommended for extended histone modification marks like H3K27me3. This parameter must be a multiple of window size.

11. FDR: the desired false discovery rate cutoff in identifying statistically significant islands. In this example FDR = 0.1%.

3.2.5 Analysis of SICER Output

It takes 35 min to finish the SICER.sh run on this data file. Please be advised that the process time increases with increasing ChIP-Seq and control library size.

SICER.sh should generate the following files after the complete workflow. Explanation of each file is as follows:

1. ES_H3K27me3-1-removed.bed. This file is in BED format and contains all reads in the ChIP library after removal of redundant reads.
2. ES_H3K27me3-W200-normalized.wig. This file is in WIG format and could be uploaded to the UCSC genome browser for visualization of the ChIP library with redundancy-removed but before island-filtering (ES_H3K27me3-1-removed.bed) with desired window size.
3. ES_H3K27me3-W200.graph. This file is in bedGraph format and is a summary of the redundancy-removed data file.
4. ES_H3K27me3-W200.scoreisland. This file stores all identified islands with respective scores and could be used to evaluate the choice of gap size.
5. ES_H3K27me3-W200-G600-islands-summary. This 8-column file is a summary of all islands identified in the ChIP library. The format is chromosome, start, end, read count in ChIP library, read count in control library, p-value, fold change, FDR).
6. ES_H3K27me3-W200-G600-islands-summary-FDR1e-3. This is a subset of identified islands whose FDR are less than the given threshold. It is the same 8-column format as the summary.
7. ES_H3K27me3-W200-G600-FDR1e-3-island.bed. This file has 4 columns and contains information about all identified ChIP-enriched region information under filtering parameters (window size 200, gap size 600, FDR cutoff 1e-3). The 4 columns are chromosome, start, end, and read count in ChIP library.
8. ES_H3K27me3-W200-G600-FDR1e-3-islandfiltered.bed. This file is in BED format and contains all reads that are within significant islands.
9. ES_H3K27me3-W200-G600-FDR1e-3-islandfiltered-normalized.wig. This file is in WIG format and could be uploaded directly to UCSC genome browser for visualization of the island-filtered ChIP library.
10. ES_input-1-removed.bed.

    This file is in BED format and contains all reads in the control library after removal of redundant reads to the threshold 1.

Of all these files, the ES_H3K27me3-W200-G600-islands-summary-FDR1e-3 and ES_H3K27me3-W200-G600-FDR1e-3-island.bed are the most important for further analysis. The first one contains the details about each significant island. The second one contains the redundancy-removed raw reads filtered by islands. In addition, the wig files can be used for visual examination of the raw and processed data on the genome browser.

Figure 1 shows the results of SICER on Wnt6 and Wnt10a gene locus. The top data track, showing raw data, suggests that H3K27me3 is enriched in broad domains across the entire Wnt6 and Wnt10a gene locus. SICER under default parameters does a good job in capturing the extended domains of H3K27me3 enrichment.

4 Notes

1. SICER prerequisites.

   SICER takes advantage of Shell scripting language and therefore runs in a Unix/Linux platform (e.g., Mac OS X and Ubuntu). SICER cannot be run directly under Windows. However, it potentially could work under a Unix/Linux simulator (e.g., Cygwin).

   We recommend running SICER with 64-bit Python 2.6 or Python 2.7. The current version of SICER is not compatible with Python 3.X. To check if the Python running is 32-bit or 64-bit, the user can run the following command in a terminal:

   ```
   $ python -c 'import struct; print struct.calcsize("P") * 8'
   ```

   In addition, SICER requires Python libraries including NumPy, SciPy, and Pylab. Instructions on installation of these packages can be found on the web sites of respective packages. To check their installation, one can launch the Python environment and run the following commands inside:

   ```
   >>>import numpy
   >>>import scipy
   >>>import pylab
   ```

   If all packages are installed and functioning correctly, there should be no message displayed.

2. Avoid running multiple SICER instances under the same directory.

   During the execution, SICER generates temporary files with hard-coded names. Therefore, multiple SICER instances running in the same directory will interfere with each other, leading to unpredictable results. This is particularly important if users are using a centralized cluster-computing management system (e.g., Condor).

3. Adding additional genome.

   SICER by default supports reference genomes mm8, mm9, rn4, hg18, hg19, sacCer1, dm2, dm3, pombe, and tair8. If the desired reference genome is not in the list, user can easily add customized reference genome information in the file GenomeData.py under /lib. For example, if a new reference genome "NewGenome" contains two chromosomes "chr1" and "chrX," with length 100 and 200, respectively, user will need to:

   (a) Add this entry to dictionary "species_chroms": "NewGenome": NewGenome_chroms

   (b) Add this entry to dictionary "species_chrom_lengths": "NewGenome": NewGenome_chrom_length

   (c) Add this list to GenomeData.py: NewGenome_chroms=["chr1," "chrX"]

   (d) Add this dictionary to GenomeData.py: NewGenome_chrom_lengths={"chr1": 100, "chrX": 200}

4. Compare epigenomes and identify differentially enriched regions.

   A frequently encountered case in epigenomic analysis is to identify significant differences between two conditions: wild-type cells versus treated cells, normal cells versus pathological cells, or undifferentiated cells versus differentiated cells. SICER-df.sh and SICER-df-rb.sh are designed to identify domains of differential enrichment. SICER-df.sh shall be used when matching control libraries for the two conditions are available, whereas SICER-df-rb.sh shall be used when random background has to be used. Here we focus on SICER-df.sh. For clarity, we will call the two conditions wild-type (WT) and knockout (KO). SICER-df.sh works as follows: (1) it first identifies significant islands in both WT and KO. This is done by using SICER.sh to process both WT and KO ChIP-Seq libraries against their respective control libraries. (2) It merges the identified ChIP-enriched regions from WT and KO libraries. The merged islands are regarded as the "candidate" islands and constitute the units of comparison. Therefore, the "candidate" islands are required to be significantly ChIP-enriched (compared with the respective control library) in at least one of the two conditions. (3) On each "candidate" island, signal level in KO is compared with that in WT to determine the significance of changes. Regions with increased or decreased enrichment fulfilling a desired FDR requirement are reported. Fold-change values are also reported in the output file.

   Figure 7 illustrates the result of SICER-df in identifying regions with differential enrichment of H3K9me2 during adipogenesis [16]. SICER-df finds several domains of decreased

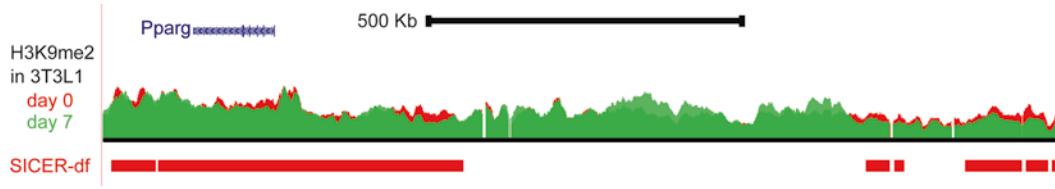


Fig. 7 SICER-df identifies regions with reduced H3K9me2 enrichment during adipogenesis. *Top track*: SICER-filtered H3K9me2 data before differentiation (day 0, *red*) and after differentiation (day 7, *green*) overlaid with the same vertical scale. Raw data behaves similarly. *Bottom track*: Regions with decreased H3K9me2 enrichment as identified by SICER-df. Here window size $w = 500$ bps (as calculated using Shimazaki and Shinomoto [10]), gap size $g = 3w$, FDR cutoff is 0.1 %. The profiles were smoothed for illustration purpose (with UCSC genome browser smoothing window set to 10 pixels). There are multiple genes in this locus but only PPARγ is shown

H3K9me2 enrichment in a 1.5 Mb region. In particular, an extended region of differential enrichment (~500 kbs) covers the PPARγ gene (*see* Fig. 7), a master regulator of the adipogenetic process. Interestingly, the specific removal of H3K9me2 correlates well with the induction of PPARγ during adipogenesis and supports the regulatory role of histone methyltransferase G9a-mediated repressive epigenetic mark H3K9me2 [16] on PPARγ expression.

Acknowledgement

This work was supported in part by the Intramural Research Program of the NIDDK, NIH to KG.

References

1. Barski A, Cuddapah S, Cui KR, Roh TY, Schones DE, Wang ZB, Wei G, Chepelev I, Zhao KJ (2007) High-resolution profiling of histone methylations in the human genome. Cell 129(4):823–837
2. Wang ZB, Zang CZ, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui KR, Roh TY, Peng WQ, Zhang MQ, Zhao KJ (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40(7):897–903
3. Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W, Zhao K (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. Cell 138(5):1019–1031
4. 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(15):1952–1958
5. Aagaard L, Laible G, Selenko P, Schmid M, Dorn R, Schotta G, Kuhfittig S, Wolf A, Lebersorger A, Singh PB, Reuter G, Jenuwein T (1999) Functional mammalian homologues of the Drosophila PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31. EMBO J 18(7):1923–1938
6. Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410(6824):120–124

7. Lachner M, O'Carroll N, Rea S, Mechtler K, Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410(6824):116–120
8. Schwartz YB, Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet 8(1):9–22
9. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57:289–300
10. Shimazaki H, Shinomoto S (2007) A method for selecting the bin size of a time histogram. Neural Comput 19:1503–1527
11. Song Q, Smith AD (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics 27(6):870–871. doi:10.1093/bioinformatics/btr030
12. Koehler R, Issac H, Cloonan N, Grimmond SM (2010) The uniqueome: a mappability resource for short-tag sequencing. Bioinformatics. doi:10.1093/bioinformatics/btq640
13. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. doi:10.1093/bioinformatics/btq033
14. The EPC (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9(4):e1001046. doi:10.1371/journal.pbio.1001046
15. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ (2012) ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. doi:10.1093/nar/gks1172
16. Wang L, Xu S, Lee J-E, Baldridge A, Grullon S, Peng W, Ge K (2013) Histone H3K9 methyltransferase G9a represses PPAR[gamma] expression and adipogenesis. EMBO J 32(1):45–59. http://www.nature.com/emboj/journal/v32/n1/suppinfo/emboj2012306a_S1.html

Part II

Visual Analysis and Interpretation of Large-Scale Interaction Networks

Chapter 6

Identifying Stem Cell Gene Expression Patterns and Phenotypic Networks with AutoSOME

Aaron M. Newman and James B. Cooper

Abstract

Stem cells have the unique property of differentiation and self-renewal and play critical roles in normal development, tissue repair, and disease. To promote systems-wide analysis of cells and tissues, we developed AutoSOME, a machine-learning method for identifying coordinated gene expression patterns and correlated cellular phenotypes in whole-transcriptome data, without prior knowledge of cluster number or structure. Here, we present a facile primer demonstrating the use of AutoSOME for identification and characterization of stem cell gene expression signatures and for visualization of transcriptome networks using Cytoscape. This protocol should serve as a general foundation for gene expression cluster analysis of stem cells, with applications for studying pluripotency, multi-lineage potential, and neoplastic disease.

Key words Stem cells, Pluripotency, Cluster analysis, Gene expression patterns, Transcriptome networks, Fuzzy clustering, AutoSOME

1 Introduction

From recent advances in somatic cell nuclear reprogramming [1, 2] and cloning [3] to the isolation of self-renewing cells that drive tumorigenesis [4–6], stem cell science is rapidly progressing. In parallel, microarrays and next-generation sequencing platforms have become invaluable for delineating critical stem cell genes and pathways [7–9]. To facilitate the identification and analysis of gene expression signatures that underlie distinct phenotypic states and to usefully visualize similarity among whole-transcriptome profiles, we developed a novel cluster analysis method called AutoSOME [10]. Unlike previous clustering approaches, such as hierarchical or K-means clustering [11], AutoSOME can process millions of data points on a desktop computer in practical time and requires no assumptions about cluster number or structure. We previously applied AutoSOME to transcriptome analysis of human embryonic and induced pluripotent stem cells (ESCs and iPSCs, respectively) and identified a large protein–protein interaction network significantly

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_6, © Springer Science+Business Media New York 2014

enriched in the pluripotency phenotype [10], consisting of ~four-fold more genes than previously estimated [7]. Separately, in a meta-analysis across stem cell labs, we demonstrated that previously reported gene expression differences between human ESCs and iPSCs [12] could be better explained by laboratory-specific factors than cell-type-specific gene expression [13].

Here, we provide a road map for the stem cell community to perform whole-transcriptome cluster analysis with AutoSOME. Using publicly available microarray gene expression data consisting of diverse human cell types, we show how to prepare and normalize input for AutoSOME analysis and anecdotally demonstrate how to use AutoSOME to (1) derive modular gene expression signatures that reflect key genetic programs and (2) create fuzzy cluster networks that capture similarity and heterogeneity among stem cell and somatic cell phenotypes. We illustrate how to render the latter using Cytoscape [14], a network analysis and visualization tool, and clusterMaker [15], a Cytoscape plugin implementing AutoSOME. Gene expression clusters will be identified easily and automatically, with the cluster number determined by the underlying statistics of the data and by a user-adjustable *p*-value threshold. Importantly, the approach described here is not limited to microarrays and can be applied to whole-transcriptome FPKM expression data from RNA-Seq studies.

2 Materials

A 64-bit desktop computer or laptop equipped with at least 2GB RAM and a multi-core processor is recommended, though a single CPU and/or 32-bit system may be used (*see* **Note 1**). AutoSOME is written in Java and has been extensively tested on Windows, Mac, and Linux platforms.

2.1 Software

1. Java 1.6+: AutoSOME requires Java version 1.6 or later to run properly. To check which version of Java is installed, if any, open a Unix shell (Mac/Linux) or DOS console (Windows) and execute the command, "java -version." If needed, download the latest release of the Java Runtime Environment (JRE) from Oracle (http://www.oracle.com/technetwork/java/javase/downloads/index.html; *see* **Note 1**).
2. AutoSOME: Using an academic email address, register online to download the latest AutoSOME Graphical User Interface (GUI) and command line tool (http://autosome.stanford.edu/download.jsp; *see* **Note 2**). Unpack the AutoSOME ZIP file to your working directory.
3. Cytoscape: Download and install Cytoscape version 2.8.3 (*see* **Note 3**) (http://www.cytoscape.org/download.html).

4. clusterMaker: Download clusterMaker version 1.11 at (a) http://autosome.stanford.edu/clustermaker.html or (b) http://chianti.ucsd.edu/cyto_web/plugins/displaypluginfo.php?name=clusterMaker, and save the clusterMaker.jar executable in your Cytoscape plugins folder (e.g., /Cytoscape_v2.8.3/plugins).

2.2 Microarray Data

1. Download the GSE22651 series matrix file (under "Download family") from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22651). Extract the series matrix GZIP file and move the resulting TXT file to your working directory (if double-clicking does not extract the file and a Unix environment is available, use gunzip; otherwise, instructions and free GZIP tools can be found online).
2. Repeat **step 1** for GSE23034, (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23034).

3 Methods

Subheading 3.1 summarizes AutoSOME input formats and describes how to launch AutoSOME, load input data, and filter and normalize gene expression data. Subheading 3.2 illustrates how to identify and visualize clusters of coordinated gene expression. Subheading 3.3 demonstrates how to cluster biological samples based on whole-transcriptome patterns and visualize the results as a fuzzy cluster network in Cytoscape. Subheading 3.4 describes an alternative approach to Subheading 3.3, in which both clustering and visualization are performed in Cytoscape using clusterMaker. This protocol is outlined in Fig. 1.

3.1 Data Formats, Input, and Normalization

1. AutoSOME accepts three input formats. The most common is a simple tab- or comma-delimited table of gene expression data, where the first column contains gene labels (e.g., HUGO or Entrez gene symbols, microarray probe sets, etc.) and the first row contains sample names (e.g., ESC1, ESC2, iPSC1, iPSC2, etc.). The table cell occupying the first row and first column should be labeled "Gene symbol," or comparable. The second type of input is a PCL file, used by Cluster 3.0 software [16]. The third format is a series matrix file, a standardized storage format for normalized gene expression data from a previous microarray study, freely available from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/).
2. To open AutoSOME, go to http://autosome.stanford.edu and press the large "Launch" button. This will deploy the AutoSOME GUI on your local computer using Java Web

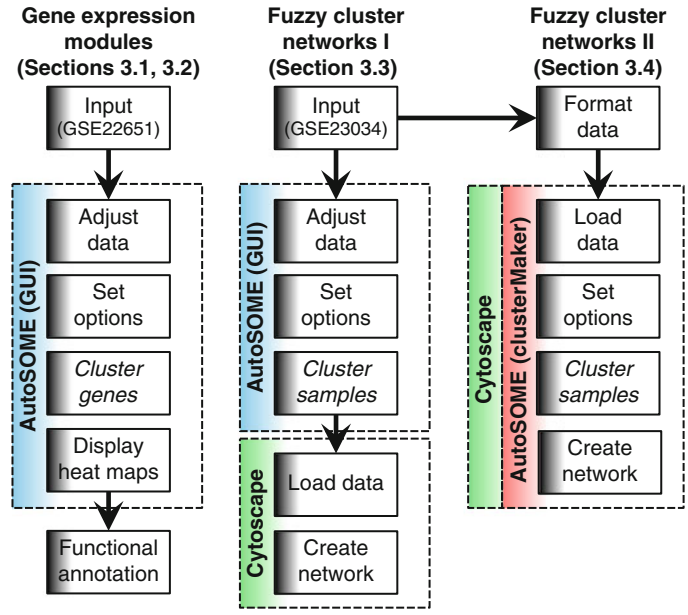


Fig. 1 Protocol schema

Start. It may be necessary to navigate through intermediate pop-up windows by pressing "OK" and/or agreeing to various terms before AutoSOME launches. Alternatively, deploy the AutoSOME GUI from a Unix shell or console window by running the following command from your working directory:

```
java -Xmx2g -Xms2g -jar autosome_vXXXXXX.jar
```

The first two arguments, –Xmx and –Xms, specify the amount of memory to allocate to AutoSOME (in this case, 2GB), and XXXXXX denotes the release date of the AutoSOME version in MMDDYY format (*see* **Note 4**).

3. Open a file browser by pressing the "Input" button (red text), located in the upper left corner of the GUI (Fig. 2). By default, AutoSOME expects a simple tabular input, as described in Subheading 1. However, since we are analyzing a series matrix file, press the checkbox next to "Gene Expression Omnibus Series Matrix File" in the blue "Data Format" box and open the previously downloaded series matrix file, "GSE22651_series_matrix.txt."
4. AutoSOME will display basic statistics about the input data (i.e., number of rows/columns, maximum and minimum expression values), followed by a prompt to filter data prior to cluster analysis. Press "Yes."
5. The removal of genes with low expression levels and/or low inter-sample variance can improve the identification of cluster boundaries (*see* **Note 5**). Since expression values in GSE22651

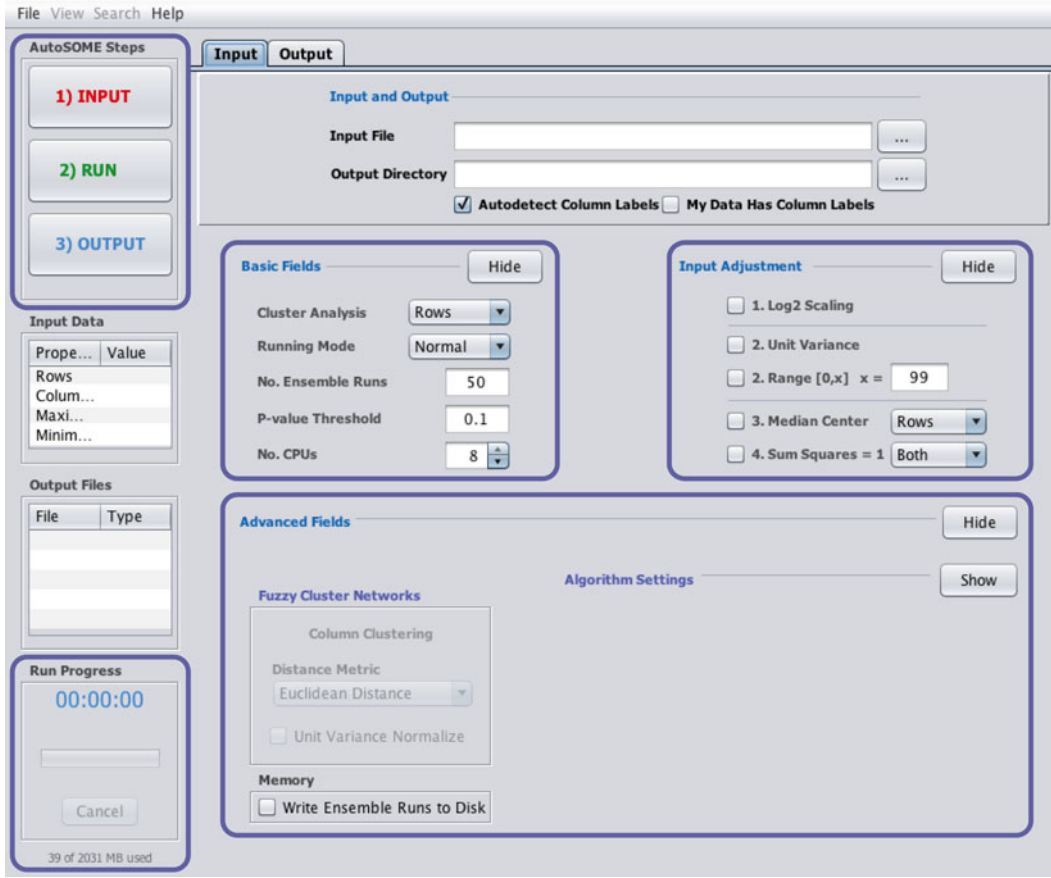


Fig. 2 AutoSOME GUI. Locations of major parameters and features are indicated by rounded rectangles

range from ~25 to ~27,500, they are not in log space and we can leave the corresponding checkbox unselected. Check "Remove rows with fold change less than," and input "4." Check "Remove all rows with mean value below," and input "100." Press "Apply." This leaves 11,364 of 48,785 rows (i.e., gene probes). Press "Accept."

6. Data normalization is critical for successful gene expression clustering, and familiarity with methods implemented by AutoSOME is advised (*see* **Note 6**). To access data scaling and normalization options, press "Show," located next to "Input Adjustment" (Fig. 2). For this exercise, select "Log2 Scaling," "Unit Variance," "Median Center Rows," and "Sum Squares = 1 Both" (*see* **Note 6**).

3.2 Clustering Gene Expression Data

1. AutoSOME basic fields (Fig. 2) dictate several critical aspects of cluster output, including stability and significance (*see* **Note 7**). Press "Show" next to "Basic Fields" to expand AutoSOME

Table 1
Summary of AutoSOME output files

AutoSOME running mode	File name	Description
Any	AutoSOME_**Input**_E**X**_Pval**Y**_**Z**_summary.html	Summary of cluster parameters and results
Any[a]	AutoSOME_**Input**_E**X**_Pval**Y**_**Z**.txt	Clustered data (rows or columns)
Both[a]	AutoSOME_**Input**_E**X**_Pval**Y**_rows_columns.txt	Clustered data (rows and columns)
Columns[b]	AutoSOME_**Input**_E**X**_Pval**Y**_Edges.txt	Fuzzy cluster edges and weights
Columns[b]	AutoSOME_**Input**_E**X**_Pval**Y**_Nodes.txt	Fuzzy cluster node labels
Columns	AutoSOME_**Input**_E**X**_Pval**Y**_Matrix.txt	Edge weights in matrix form

Bolded variables are as follows: *Input*, name of input file; *X*, number of ensemble runs; *Y*, *p*-value threshold; *Z*, rows or columns (depending on running mode)
[a]To revisit previous results in the GUI, open this file (File>"Open AutoSOME results")
[b]Open these files in Cytoscape [14] to create a fuzzy cluster network

parameters (Fig. 2). As indicated by the "Cluster Analysis" field (the top parameter), AutoSOME provides the option of clustering rows (i.e., genes), columns (i.e., biological samples), or both (genes, then columns). In this example, we will cluster rows. Set "No. Ensemble Runs" to 100 and the *p*-value threshold to 0.05, and if desired, decrease "No. CPUs" (by default, all available cores are used). Leave "Running Mode" set to "Normal."

2. Press the large "Run" button (green text) in the upper left corner of the GUI (Fig. 2). A progress bar will be displayed in the lower left, along with elapsed time and the current clustering stage (*see* **Note 8**). Depending on system hardware, running time for this example may range from seconds to minutes.
3. When complete, AutoSOME will redirect the GUI to the "Output" tab and display a list of identified clusters ("Cluster Output"). Output files will also be written to disk (*see* Table 1). Select cluster 1 to display clustered data points (Fig. 3a). The first column contains gene identifiers (here, probe set IDs from the Illumina HumanHT-12V3.0 beadchip microarray platform), while the second column displays associated cluster confidence scores (*see* **Note 9**).

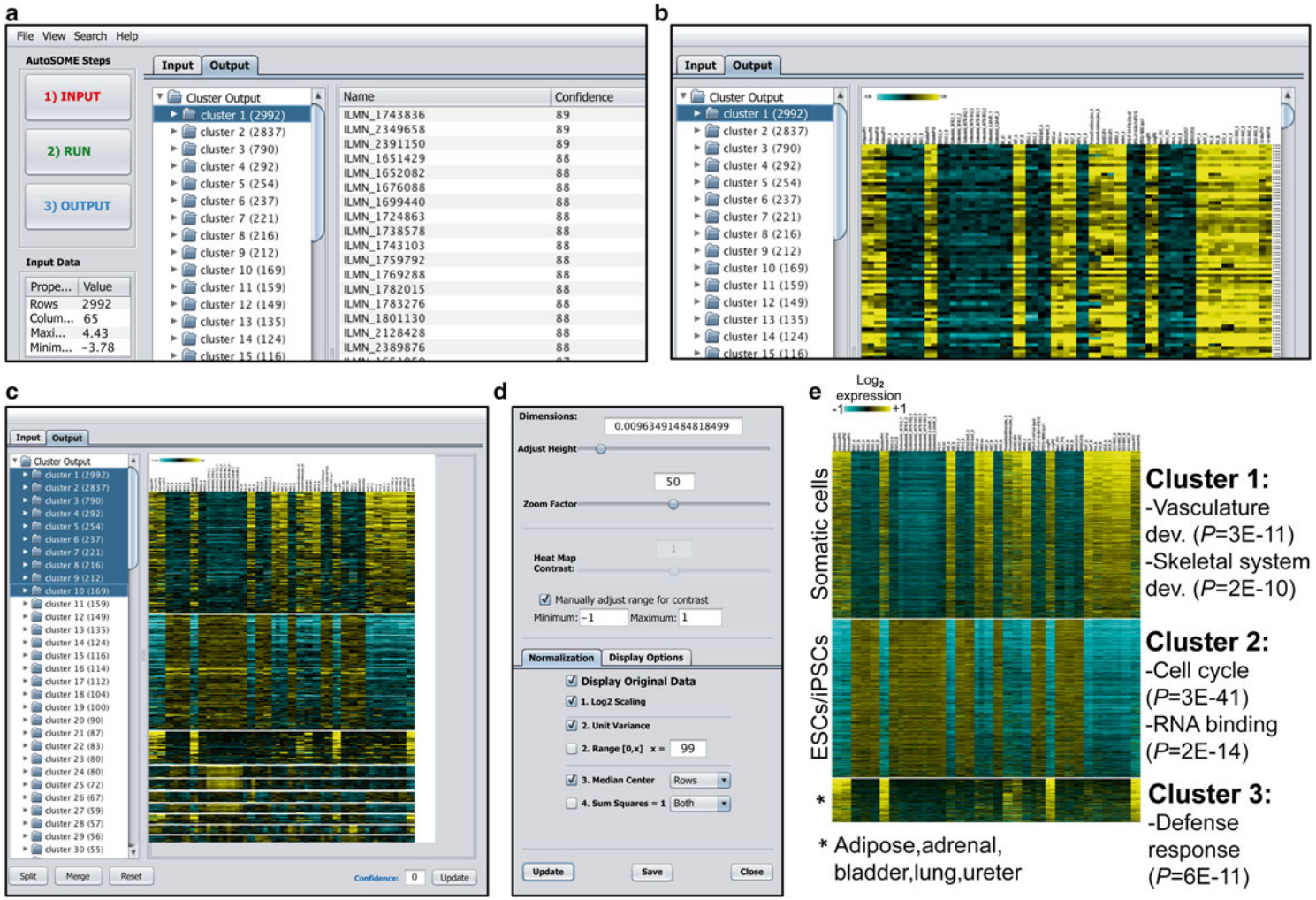


Fig. 3 Gene expression modules: display and annotation. (**a**) Raw contents of cluster 1. (**b**) Cluster 1 rendered as a heat map. (**c**) Clusters 1–10 displayed as a heat map. (**d**) "Image Settings" window. (**e**) Gene expression clusters distinguishing ESCs/iPSCs and diverse somatic cell populations (from GSE22651). Functional annotations were determined using DAVID [18] and *p*-values are Bonferroni corrected

4. Heat maps can be used to visualize AutoSOME cluster results, and several coloring schemes are provided, including red-green, yellow-cyan, and rainbow scales. For example, to display a yellow-cyan heat map, select cluster 1 and select View > heat map > yellow-cyan (Fig. 3b). Heat map visualization parameters, such as size, dimensions, and contrast, can be altered using the "Image Settings" tool, located in View > settings > "image settings." To resize a heat map to fit the viewing screen, click the right mouse button (or equivalent) when hovering over the heat map or select View > "fit to screen." For users that prefer alternate visualization software, such as Java TreeView [17] or R, AutoSOME results are readily transferrable (*see* **Note 10**).
5. To select and visualize multiple clusters simultaneously, use the SHIFT or CTRL keys. For example, to select clusters 1–10, select cluster 1 and while pressing SHIFT, select cluster 10. Display as a heat map, and fit to the screen, as described in **step 4** (Fig. 3c). To export in high resolution, open "Image Settings," increase "Zoom Factor" to at least 50 (Fig. 3d), and press "Save" to launch a file browser (alternatively, File > Export > "save image"). Though only partially visible, the entire image will be saved as Portable Network Graphics (PNG) file (*see* **Note 11**).
6. The AutoSOME GUI facilitates functional annotation of selected clusters by allowing users to easily copy/export gene lists into online gene set enrichment tools. In this exercise, we clustered probe set identifiers, not gene symbols. However, the Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) [18] accepts a variety of input identifiers, including Illumina beadchip, and can convert a number of gene formats to standard HUGO gene symbols (http://david.abcc.ncifcrf.gov/conversion.jsp). Select View > raw data to return to the original output display and select cluster 1. To copy cluster contents, highlight the contents of column 1 and press CTRL-C or copy contents from the appropriate output file (Table 1). While the use of DAVID is beyond the scope of this section, Fig. 3e illustrates possible functional annotation of clusters 1–3.

3.3 Fuzzy Cluster Networks I (Using AutoSOME GUI and Cytoscape)

1. Fuzzy cluster networks illuminate differences between individual biological samples and differences between clusters using an intuitive network schematic [10]. Complete Subheading 3.1, **steps 2–4** using input data set, GSE23034_series_matrix.txt, and do not prefilter.
2. Expand "Input Adjustment" to show scaling and normalization options. Check "Unit Variance" and leave the remaining settings unchecked (these data range from 1.74 to 13.92 and thus are already $\log_2$ adjusted; *see* **Note 6**).

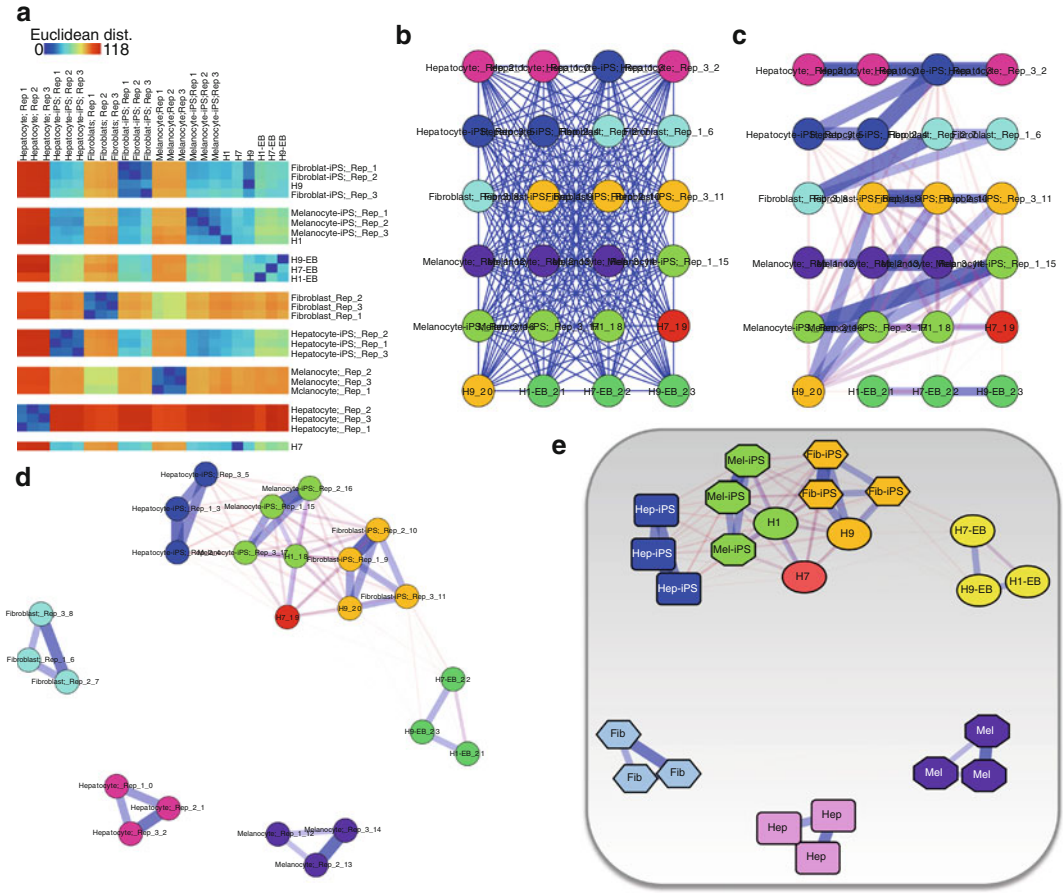


Fig. 4 Biological sample clusters: heat maps and fuzzy cluster networks. (**a**) Heat map depicting whole-transcriptome clusters (separated by horizontal white bars) of stem cell and somatic cell lines (from GSE23034). (**b**) Cytoscape network related to results in panel **a**, with nodes colored according to cluster assignments. (**c**) Network after application of edge appearance settings. (**d**) Fuzzy cluster network using a force-directed layout. (**e**) Final network after additional enhancements (e.g., node shape reflects cell of origin; some node colors from panels **b**–**d** were changed). Sample names are abbreviated as follows: *Fib,* fibroblast; *Hep,* hepatocyte; *Mel,* melanocyte; *EB,* embryoid body. H1, H7, and H9 are ESC lines; the notation *X*–*Y* indicates *Y* derived from *X*

3. Expand "Basic Fields" to access AutoSOME parameters (*see* **Note 7**). Set the "Cluster Analysis" field to columns (i.e., to cluster biological samples; *see* **Note 12**), "No. Ensemble Runs" to 500, and "*p*-value threshold" to 0.01.
4. Run clustering (as described in Subheading 3.2, **step 2**).
5. Inspect raw results and then visualize as a heat map, as described in Subheading 3.2, **step 4**. The similarity between each sample and all other samples is shown rather than gene expression data (Fig. 4a; *see* **Note 12**).
6. Locate and launch Cytoscape (*see* **Notes 3** and **12**).

7. AutoSOME generates two output files needed to create a fuzzy cluster network—a file containing node names (i.e., biological samples) and corresponding cluster IDs and an edge file, consisting of connections between nodes that reflect their similarity, ranging from −0.5 (never co-cluster over ensemble procedure) to +0.5 (always co-cluster) (Table 1). To render these data as a network in Cytoscape 2.8.3, import edges by selecting File > Import>"Network from Table (Text/MS Excel)…" and browse for the input file, AutoSOME_GSE23034_series_matrix_E500_Pval0.05_Edges.txt. Press "Open." Set "Source Interaction" to Column 1 and "Target Interaction" to Column 2 (leaving "Interaction Type" as default), activate "Column 3" (shown in the preview table) by selecting "X" (will become a check mark), and press "Import." To import network nodes, select File > Import>"Attribute from Table (Text/MS Excel)…" and browse for the input file, AutoSOME_GSE23034_series_matrix_E500_Pval0.05_Nodes.txt. Press "Import."
8. Enlarge the network window to fill the Cytoscape viewport. If details disappear, select View>"Show Graphics Details."
9. Cytoscape has numerous graphical layout parameters for enhancing cosmetic appeal. We prefer the following settings, accessible through the Cytoscape VizMapper™ (View > Open VizMapper™). Color nodes according to cluster assignments by first double-clicking "Node Color" in "Unused Properties" and choosing column 2 (i.e., cluster ID). Then select "Discrete Mapping" for mapping type (right column), right-click over the list of cluster IDs (left column), and select "Generate Discrete Values" > Rainbow 1 (Fig. 4b). To modify edge appearance, update the following fields with respect to column 3 (i.e., edge weights): (a) Edge Color {Continuous Mapping}: −0.5 = red, ≥0 = blue (double-click on the arrows to change values); (b) Edge Line Width {Continuous Mapping}: −0.5 = 0.5, +0.5 = 20; and (c) Edge Opacity {Continuous Mapping}: −0.5 = 0.5, +0.5 = 200. Finally, click on the network image underneath "Defaults," and in the pop-up window, change "NODE_BORDER_COLOR" to black (node tab) and "Background Color" to white (global tab). Press "Apply" (Fig. 4c).
10. Open the layout algorithm panel by selecting Layout> "Settings…" and then select the Force-Directed Layout algorithm. Set "The edge attribute that contains the weights" to Column 3, and set minimum and maximum edge weights to −0.5 and 0.5, respectively. Set "Default Spring Coefficient" to 5E-6, "Default Spring Length" to 50, "Default Node Mass" to 3, and "Force deterministic layouts" to true.

Press "Execute Layout" (Fig. 4d; *see* **Note 14**). To manipulate the network (e.g., scale and rotate), select View>"Show Tool Panel."

11. After rotating the network, we used the VizMapper™ to further enhance network appearance and emphasize the cell of origin for reprogrammed cells (i.e., iPSCs) and differentiated cells (i.e., embryoid bodies) (Fig. 4e). Notably, while pluripotent stem cells are topologically adjacent in the network, iPSCs cluster by somatic cell of origin (Fig. 4e), consistent with findings from the original study [8].

3.4 Fuzzy Cluster Networks II (Using ClusterMaker and Cytoscape)

1. As an alternative to Subheading 3.3, the clusterMaker plugin [15] allows AutoSOME to be run directly in Cytoscape, enabling one-stop clustering and visualization of biological samples. To import raw expression data into Cytoscape, the GSE23034 series matrix file will need to be reformatted. Open GSE23034_series_matrix.txt in Excel and if not done automatically, convert tab-delimited text to columns. Cut the row with "!Sample_title" in column A (here, row 41) and insert it directly under the row with "ID_REF" in column A (here, row 78). Delete all metadata rows (here, rows 1–77; highlight rows with mouse, right-click, and select delete). Lastly, delete "!series_matrix_table_end," located in column A in the last row (here, row 28,314). Save as a tab-delimited TXT file (GSE23034.txt), press "Continue" if prompted, and close Excel.
2. If Cytoscape is open, start a new session (File > New > Session). Otherwise, launch Cytoscape. Import edges by selecting File > Import>"Network from Table (Text/MS Excel)…" and browse for the input file, GSE23034.txt. Under Advanced, check "Show Text File Import Options" and select "Transfer first line as attribute names." Set "Source Interaction" to Column 1 and press "Import." Import nodes by selecting File > Import > "Attribute from Table (Text/MS Excel)…" and browse for the input file, GSE23034.txt. Under Advanced, check "Show Text File Import Options," select "Transfer first line as attribute names," and press "Import." Due to the size of the network (i.e., the entire transcriptome), it may not be automatically displayed.
3. Launch AutoSOME in clusterMaker by selecting Plugins > Cluster > "AutoSOME Clustering." Next to "Array Sources (Node Attributes)," select all samples from GSE23034 starting with "node" (i.e., select the top sample, hold SHIFT and select the bottom sample).
4. Under "AutoSOME Basic Tuning," change "Number of Ensemble Runs" to 500 and "*p*-value threshold" to 0.01. Press "Show Data Normalization" and select "Unit Variance," leaving the remaining fields at default values. Press "Show Fuzzy

Cluster Network Settings" and select "Perform Fuzzy Clustering." Set Distance Metric to "Euclidean" (*see* **Note 12**).

5. Press "Create Clusters" in the AutoSOME Cluster Settings window.
6. When finished, under "Data Output," set "Choose Visualization" to Network and press "Display" (Fig. 4b; *see* **Note 14**).
7. If network details are missing, select View>"Show Graphics Details."
8. Unlike the AutoSOME GUI, clusterMaker allows the number of edges in fuzzy cluster networks to be altered. For example, by reducing edges, modular cluster architectures can be more readily visualized. This can be achieved by changing "Maximum number of edges to display in fuzzy network," located under "Fuzzy Cluster Network Settings." To recreate the network, repeat **step 6** (Fig. 4c).

4 Notes

1. The Java Virtual Machine (JVM) allows Java software to be run on multiple computing platforms, but only allocates ~64 MB RAM by default. Additional memory must be allocated to the JVM before starting AutoSOME, and 32- and 64-bit JVMs have different memory ceilings, with the former maxing out at ~1.5 to 3GB and the latter limited only by available RAM. We recommend using a 64-bit version of Java (which requires a 64-bit operating system). Separately, AutoSOME running time will decrease linearly with increasing dedicated CPU cores, and a computer equipped with a multicore processor is advised.
2. AutoSOME is also available at http://jimcooperlab.mcdb.ucsb.edu/autosome.
3. Cytoscape has a large and active developer community that regularly releases new versions. During preparation of this chapter, Cytoscape version 3.0.1 was released, and major changes to the user interface were introduced. Because clusterMaker is not yet available for Cytoscape version 3.x, we used version 2.8.3 in Subheadings 3.3 and 3.4. For readers interested in Cytoscape version 3, *see* **Note 13**.
4. For common usage, 1.5 to 3GB of allocated memory should be sufficient. However, extremely large data sets may require upwards of 10GB (*see* **Note 1**). If insufficient memory is available, Java will throw an "OutOfMemory" error (printed in the terminal/console), and AutoSOME may cease to function normally. Users can monitor both memory usage and maximum

available memory in the lower left corner of the GUI. If the user's computer is constrained by physical memory, AutoSOME can lower its memory footprint by writing intermediate ensemble iterations to disk. This option is available under "Advanced Settings" in the GUI. If used, "No. CPUs" in "Basic Fields" should be set to 1.

5. AutoSOME scales favorably with large data sets, and thus data filtration is not strictly required. However, some users may wish to ignore genes with low expression or low variance, and in general, gene filtration makes cluster boundaries more apparent. Moreover, unlike RNA-Seq, microarrays have considerable noise at the low end, and distinguishing artifact from true signal among low expressing genes can be difficult without independent validation [19, 20].
6. To facilitate the identification of cluster structure in gene expression data, normalization is critical. Reasons to normalize include standardizing the representation of data across biological samples, reducing the impact of technical noise, and improving statistical properties [21]. AutoSOME includes a variety of common normalization methods useful for intensity-based expression data: (a) *Log_2 Scaling* transforms the data into a more intuitive space (typically approximating a skewed normal distribution) and reduces the influence of outliers; (b) *Unit Variance* standardizes every column to a mean of zero and standard deviation of 1, allowing different arrays to be fairly compared; (c) *Median Centering* subtracts the median from each row (i.e., gene) and/or column (i.e., biological sample) to allow *relative* patterns of gene expression to be compared; and (d) *Sum of Squares = 1* normalization dampens outliers in noisy data by setting the sum of squares of all data points in a row or column equal to 1. Strategies (a), (c), and (d) are also available in Cluster 3.0 [16]. In general, we recommend applying (a)–(d) for gene expression clustering and (a)–(b) for biological sample clustering. Note that (a) should not be performed again if the input data are already $\log_2$ adjusted.
7. Whether AutoSOME will cluster rows, columns, or both is easily specified with the "Cluster Analysis" field. If both is selected, relevant options in "Basic Fields" will be split to control rows and columns separately. "Running Mode" should be set to "Normal" for most applications and can be safely ignored. The number of "Ensemble Runs" dictates AutoSOME output stability and is analogous to a statistical bootstrapping procedure. Our empirical experiments indicate that output variance decreases only marginally after 50 iterations, and for initial analyses, this is the number we recommend. For final clustering, we recommend 100–200 runs for discrete clusters and 500 runs for creating fuzzy cluster networks (*see* **Note 12**).

Additional runs for the latter are advised since edge weights represent how often a given pair of samples clusters together, and more ensemble iterations generally yield finer network structure. The AutoSOME *p*-value threshold determines cluster resolution based on a statistical null model. The lower the threshold, the finer grained the clusters. For "No. CPUs," *see* **Note 1**.

8. Because AutoSOME ensemble iterations are parceled out among all dedicated CPU cores, the progress bar may move in larger intervals than expected. We advise waiting a minute or two if the progress bar appears to have stalled at the beginning of an AutoSOME run. If inactivity persists, try setting "No. CPUs" to 1 to see whether it progresses in smaller intervals, and if not, check memory usage (*see* **Note 4**).
9. The cluster confidence score reflects the strength of association between every data point and its assigned cluster [10]. Users can dynamically filter and restore data points using the confidence filter, located in the lower right corner of the "Output" tab. To graphically display cluster confidence, check "Show Heat Map Confidence Bar" under the "Display Options" tab in "Image Settings" (View > settings>"image settings").
10. Depending on whether rows or columns are clustered, AutoSOME outputs a corresponding TXT file, "…rows.txt" or "…columns.txt", respectively, containing cluster assignments (column 1), confidence values (column 2; *see* **Note 9**), and all pre-normalized expression data (columns 3 to *m*), ordered by decreasing cluster size and decreasing cluster confidence. To transfer these results to Java TreeView, for example, open the TXT file in Excel, remove the first row (containing data normalization settings readable by AutoSOME), and remove the first two columns (to subset the data on 1 or more particular clusters, do this before saving). Save as a tab-delimited TXT file and close Excel. Open the data in Cluster 3.0 [16] and normalize as desired and then save and open in Java TreeView [17].
11. Due the stochastic component of AutoSOME [10], data points with weak membership to a particular cluster (*see* **Note 9**) may change cluster membership in subsequent runs. The number of ensemble runs and the inherent noise in the data govern output stability, and stable cluster results are typically attained with 100–200 ensemble iterations (*see* **Note 7**). However, results may differ marginally from those presented in Fig. 2.
12. The number of genes (i.e., rows) in a typical microarray study is far larger than the number of biological samples (i.e., columns). To reduce computational load when clustering columns, AutoSOME performs an all-against-all sample comparison to

build a compact representation of each samples' relatedness, called a distance matrix. Three measures of similarity are provided (see "Distance Metric" in "Advanced Fields"): (a) *Euclidean distance* (default) is the shortest path between two sets of *n*-dimensional points (i.e., two transcriptome profiles), such that lower values represent more similar biological samples; (b) *Pearson's correlation* measures the linear relationship between two sets of *n*-dimensional points, with a range of +1 to −1, such that sets lying on the same line have a perfect correlation of +1 and sets with opposite orientations (orthogonal) have a correlation of −1; and (c) *Uncentered correlation* is the same as (b), except the magnitude of the difference between the two transcriptomes is taken into account. These metrics are reviewed in [11].

13. For readers interested in rendering fuzzy cluster networks in Cytoscape 3 (*see* **Note 3**), the following steps will largely reproduce Subheading 3.3, **steps 7–10**. To import edges, go to File > Import > Network > File, find "…Edges.txt," and then press OK. For node attributes, File > Import > Table > File, find "…Nodes.txt," and then press OK. In the VizMapper located in the Cytoscape Control Panel, press "Show All" under "Visual Mapping Browser." To color nodes by cluster assignments, double-click "Node Fill Color" in "Unused Properties" and choose column 2 (cluster ID). For mapping type, choose "Discrete Mapping," which will generate a list of cluster IDs. Right-click over the list and select "Mapping Value Generators" > Rainbow. To modify edge appearance, update the following fields with respect to column 3 (i.e., edge weights): (a) Edge Stroke Color (Unselected) {Continuous Mapping}: −0.5 = red, ≥0 = blue; (b) Edge Width {Continuous Mapping}: −0.5 = 0.5, +0.5 = 20; and (c) Edge Transparency {Continuous Mapping}: −0.5 = 0.5, +0.5 = 200. Finally, to layout the network, select Layout > Settings from the menu bar. In "Layout Settings," select "Prefuse Force-Directed Layout." Set "Weight using" to Column 3 and minimum and maximum edge weights to −0.5 and 0.5, respectively. Set "Default Spring Coefficient" to 1.0E-5, "Default Spring Length" to 40, "Default Node Mass" to 2.0, and "Force deterministic layouts" to true. Press "Execute Layout" (*see* **Note 14**). For network manipulation options (e.g., scale and rotate), go to View>"Show Tool Panel."

14. Network layout algorithms are not optimal, and the settings recommended here might not be ideal for all fuzzy cluster networks. Among available parameters, the "Default Spring Coefficient" has a major effect on node–node repulsion, and lowering this value will reduce node overlap. Moreover, user intervention may be needed to "fix" the network on occasion.

For example, two highly similar samples with a thick edge may end up too far apart. This can be fixed manually, or layout parameters can be subtly modified until a consistent network is rendered.

Acknowledgments

We thank Drs. Scott Bratman and Weiguo Feng for providing useful feedback and helpful comments on this chapter. This work was supported by a grant from the Siebel Stem Cell Institute and the Thomas and Stacey Siebel Foundation awarded to AMN.

References

1. Takahashi K, Tanabe K, Ohnuki M et al (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131:861–872
2. Vierbuchen T, Ostermeier A, Pang ZP et al (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463:1035–1041
3. Tachibana M, Amato P, Sparman M et al (2013) Human embryonic stem cells derived by somatic cell nuclear transfer. Cell 153:1228–1238
4. Reya T, Morrison SJ, Clarke MF et al (2001) Stem cells, cancer, and cancer stem cells. Nature 414:105–111
5. Barker N, Ridgway RA, van Es JH et al (2009) Crypt stem cells as the cells-of-origin of intestinal cancer. Nature 457:608–611
6. Jan M, Snyder TM, Corces-Zimmerman MR et al (2012) Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci Transl Med 4: 149ra118
7. Muller F-J, Laurent LC, Kostka D et al (2008) Regulatory networks define phenotypic classes of human stem cell lines. Nature 455:401–405
8. Ohi Y, Qin H, Hong C et al (2011) Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. Nat Cell Biol 13:541–549
9. Tang F, Barbacioru C, Bao S et al (2010) Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-seq analysis. Cell Stem Cell 6:468–478
10. Newman A, Cooper J (2010) AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinformatics 11:117
11. D'Haeseleer P (2005) How does gene expression clustering work? Nat Biotech 23: 1499–1501
12. Chin MH, Mason MJ, Xie W et al (2009) Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5:111–123
13. Newman AM, Cooper JB (2010) Lab-specific gene expression signatures in pluripotent stem cells. Cell Stem Cell 7:258–262
14. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
15. Morris J, Apeltsin L, Newman A et al (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 12:436
16. de Hoon MJL, Imoto S, Nolan J et al (2004) Open source clustering software. Bioinformatics 20:1453–1454
17. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. Bioinformatics 20:3246–3248
18. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols 4:44–57
19. Draghici S, Khatri P, Eklund AC et al (2006) Reliability and reproducibility issues in DNA microarray measurements. Trends Genet 22:101–109
20. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
21. Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32:496–501

Chapter 7

Visualization and Clustering of High-Dimensional Transcriptome Data Using GATE

Patrick S. Stumpf and Ben D. MacArthur

Abstract

The potential gains from advances in high-throughput experimental molecular biology techniques are commonly not fully realized since these techniques often produce more data than can be easily organized and visualized. To address these problems, GATE (Grid-Analysis of Time-Series Expression) was developed. GATE is an integrated software platform for the analysis and visualization of high-dimensional time-series datasets, which allows flexible interrogation of time-series data against a wide range of databases of prior knowledge, thus linking observed molecular dynamics to potential genetic, epigenetic, and signaling mechanisms responsible for observed dynamics. This article provides a brief guide to using GATE effectively.

Key words Systems biology, Gene expression dynamics, Data visualization, Transcriptome correlation, Network analysis

1 Introduction

Advances in high-throughput molecular biology techniques are allowing us to dissect regulation of cell behavior at the "systems" level with ever-increasing accuracy. However, these techniques now produce vast amounts of data and it is a considerable computational challenge to collate and analyze such datasets in a user-friendly manner. This problem is particularly acute in high-throughput comparative studies, for example, when data from different high-throughput sources are being compared (comparing genetic vs. proteomic expression profiles, for instance), when meta-analyses examining similar datasets from different experiments are being conducted, or when considering high-throughput time-series.

GATE is an integrated software platform for the analysis and visualization of high-dimensional time-series datasets that addresses these problems. GATE first uses a simple clustering algorithm to arrange time-series in a two-dimensional hexagonal grid. Each

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_7, © Springer Science+Business Media New York 2014

individual hexagon (which corresponds to a single biomolecular species) on the grid is then dynamically colored according to the expression level of the molecular component to which it is assigned, allowing the creation of easily interpreted movies that animate dynamic waves of activation and inhibition passing through the genome. GATE movies may be paused at any time by the user and are fully interactive, allowing users to highlight areas of interest, examine/save their content, and interrogate features of interest by enrichment and network analysis against a variety of background knowledge datasets. As such, GATE provides a useful tool for cell biologists interested in exploring systems-level regulatory mechanisms in complicated multilayered time-series datasets or those wishing to infer possible functional relationships from time-series correlations. GATE is specifically designed to properly utilize the ever-expanding wealth of background knowledge available, in order to aid the rational construction of testable hypotheses for validation by further experimentation.

2 Materials: GATE Download and Installation

GATE is freely available for academic use [1] and comes in two versions: (1) a Windows Installer version, which contains an executable file that guides the user through the installation process, and (2) a Unix version, which is a self-extracting archive that is ready to use upon unpacking. The GATE website [1] provides detailed documentation which can be consulted if installation questions arise. An extensive manual and detailed case study are also available from the website, which give in-depth descriptions of every aspect of the software. This protocol briefly summarizes elements of the manual and the case study.

3 Methods

3.1 Introduction

At the heart of the GATE software is a simple clustering and visualization algorithm that produces movies of expression changes on a hexagonal grid. GATE analysis therefore progresses in three stages: (1) data preprocessing and formatting; (2) clustering, annotation, and layout on the hexagonal array; and (3) interrogation of the clustered data through enrichment and/or network analysis using databases of prior knowledge. In common with most bioinformatics algorithms, GATE works best if data is preprocessed to focus only on those elements that change significantly in expression during the experiment. Therefore, prior to GATE analysis, careful selection of those genes/proteins of interest should be conducted using appropriate statistical criteria. Once elements of interest have been selected, GATE clustering and visualization may be conducted.

3.2 GATE Clustering and Visualization

The GATE clustering algorithm first uses input expression time-series to create a correlation matrix, which is then used to arrange time-series on a hexagonal grid such that genes/proteins that exhibit similar temporal patterns of expression are arranged close to each other, while those with very different expression dynamics are placed far apart (optimal arrangement of time-series on the grid is conducted using a simulated annealing algorithm; see [2] for details). Finding an optimal arrangement is achieved by building a .clu GATE cluster file from the user input data. To do this, subsequent to appropriate normalization, expression data should be saved in a simple comma-separated file, in which each row represents one gene/protein expression time-series. The first entry in each row is an official NCBI Entrez gene symbol followed by a sequence of comma-separated expression values for the corresponding time points. So, for instance, the input

NANOG,3.14,1.52,10.69,2.38

POU5F1,7.43,2.82,5.55,1.11

is an example of the appropriate way to input expression changes for Nanog and Pou5f1 (Oct3/4) at 4 different time points. For successful subsequent analysis, it is important to ensure that each row is annotated with an official NCBI Entrez gene symbol (GATE will supply a warning if this is not the case). Once expression data has been loaded into GATE, a cluster file can be created using the "Create" dialogue button (Fig. 1a). Alternatively, previously clustered files, such as the sample files supplied in the "clusterfiles" folder in the GATE installation, can be loaded immediately via the "Load" dialogue button (Fig. 1a). When creating a new cluster file, dimensions for the hexagonal grid must also be defined (Fig. 1b). GATE automatically proposes a square, or near square, grid layout, which is ideal in most cases but can be adjusted liberally. Following adjustment of setting and layout options, the actual clustering is then computed (Fig. 1c). Although relatively accurate results can be achieved in few minutes, optimal arrangements are more likely to be found with longer clustering times. It is therefore recommended to run the annealing over night. Once the annealing process has been completed, the time points of the time-series can be annotated individually and the temporal spacing between time points can be customized to account for differences in sampling intervals (Fig. 1d). Once such annotation has been completed, the main GATE window opens automatically and the animation can be played using the play button in the bottom left corner of the GATE window (Fig. 2). Note that the first frame of the movie will appear black, as this is the reference for all subsequent frames. Note also that the left- and right-hand sides of the array are associated with each other; similarly, the top and the bottom sides of the array are also associated with each other.

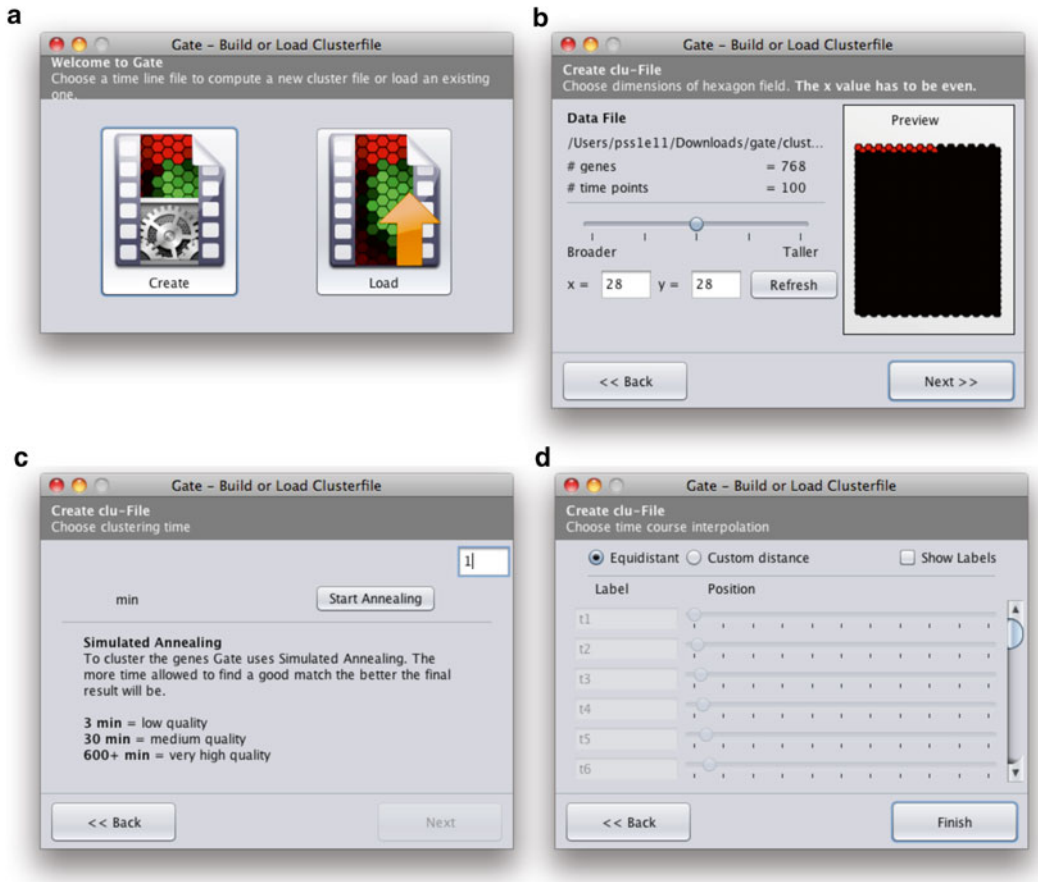


Fig. 1 The GATE clustering process. (**a**) Load and create dialogues for new or pre-clustered datasets. (**b**) Suggested grid dimensions can be customized. (**c**) Clustering is conducted using simulated annealing and is ideally run over night. (**d**) Labels and time-series intervals can be adjusted

Thus, GATE arranges input time-course data onto the surface of a hexagonally tiled torus. This ensures that there are no "special" places on the array and all molecular species are treated equally.

New cluster files may also be created according to preexisting cluster templates. This feature allows direct comparison of how dynamic changes in different regulatory "layers"—such as methylation status, transcript and protein expression—mutually affect each other [3]. Once a cluster file has been created using one dataset (for instance, changes in transcript expression), it can be used as a template for another (for instance, changes in protein expression). This function is available from the "File" tab in the top menu and requires that both datasets cover the same time points and have an identical number and annotation of objects.

When animated, GATE movies typically show concurrent waves of activation and inhibition passing simultaneously through the genome. However, these movies do not, in themselves, shed

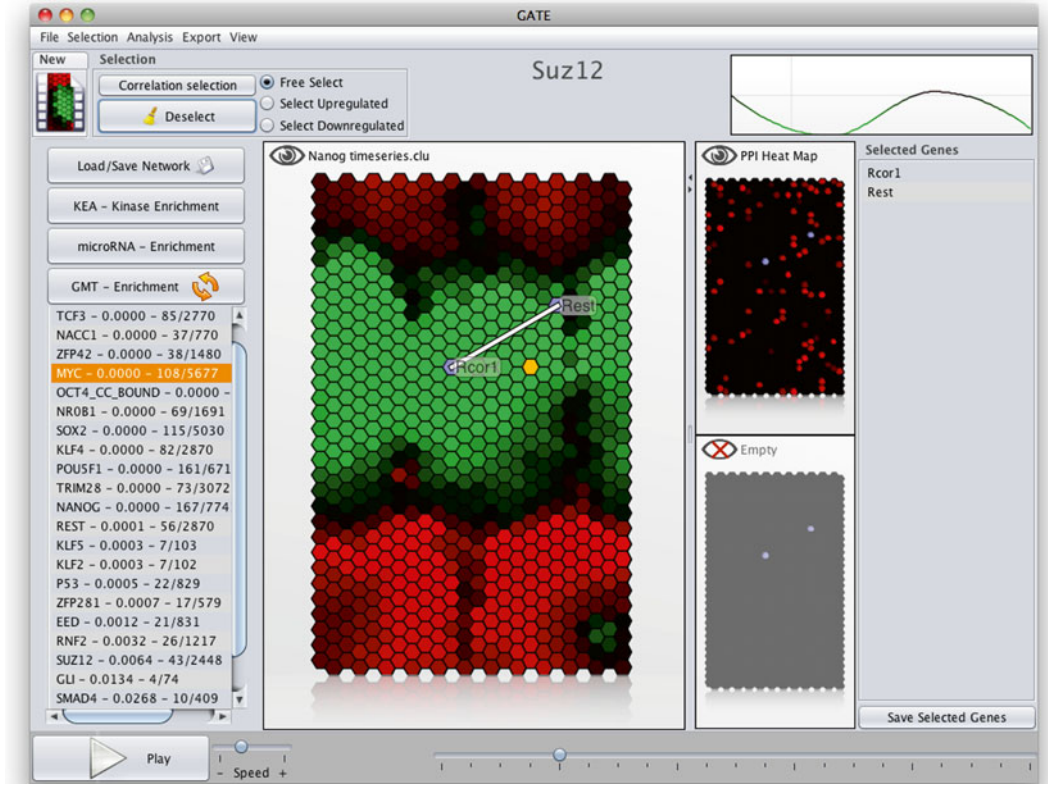


Fig. 2 The main GATE window: the hexagonal grid is in the center; selection options are in the *top panel*; network overlay and enrichment analysis options are on the *left*; selection overview is on the *right*; and movie controls are at the *bottom*

any light on the molecular regulatory mechanisms responsible for these dynamic regulatory cascades. In order to investigate further the molecular mechanisms responsible for regulatory dynamics, GATE also allows for a range of network and enrichment analyses to be conducted.

3.3 Interactive Selection of Genes

Groups of genes/proteins can be selected by encircling regions of interest on the GATE grid and left-clicking the mouse. All elements within the encircled area will then be selected. Selections may be refined by limiting to upregulated or downregulated elements by choosing the corresponding options in the upper left part of the GATE window. Additionally, the "correlation selection" button provides a more refined selection option. Using this option, elements can be selected according to the expression profile they exhibit by drawing a profile of interest in the interactive window (*see* Fig. 3).

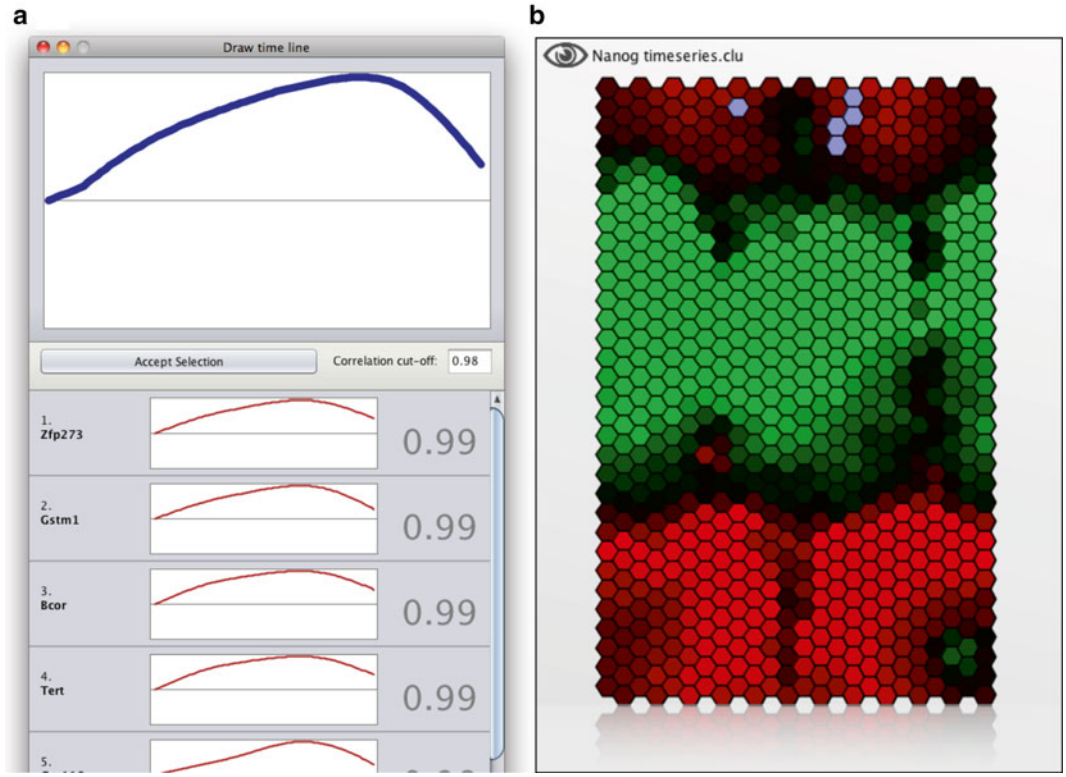


Fig. 3 Interactive selection of gene sets. (**a**) Users can sketch expression profiles of interest and GATE will return elements with expression profiles that match the sketch input. (**b**) Genes of interest are highlighted on the GATE array and can be further studied or exported

3.4 Enrichment Analysis

Enrichment analysis provides a simple way to investigate the degree to which an experimentally derived dataset intersects significantly with a collated background dataset of prior knowledge [4–6]. In order to ensure maximum compatibility with databases of prior knowledge, GATE allows enrichment analysis of experimentally derived datasets against any database as long as it is provided in gene set enrichment analysis (GSEA) gene matrix transpose (.gmt) format (formatting instructions for .gmt files are available from the GSEA documentation [7]). In particular, this feature allows enrichment analysis against both standard databases—such as gene ontology and pathway analysis—as well as against specifically tailored background datasets. A total of 18 different databases for enrichment analysis are provided with the GATE installation and many more can be downloaded online (e.g., from the molecular signatures database of the Broad Institute [8]). Additionally, user-defined enrichment files may also be uploaded in .gmt file format. Significant overrepresentation (assessed by Fisher's exact test) of genes/proteins from uploaded enrichment files can be assessed within all genes on the GATE array or among those up- or

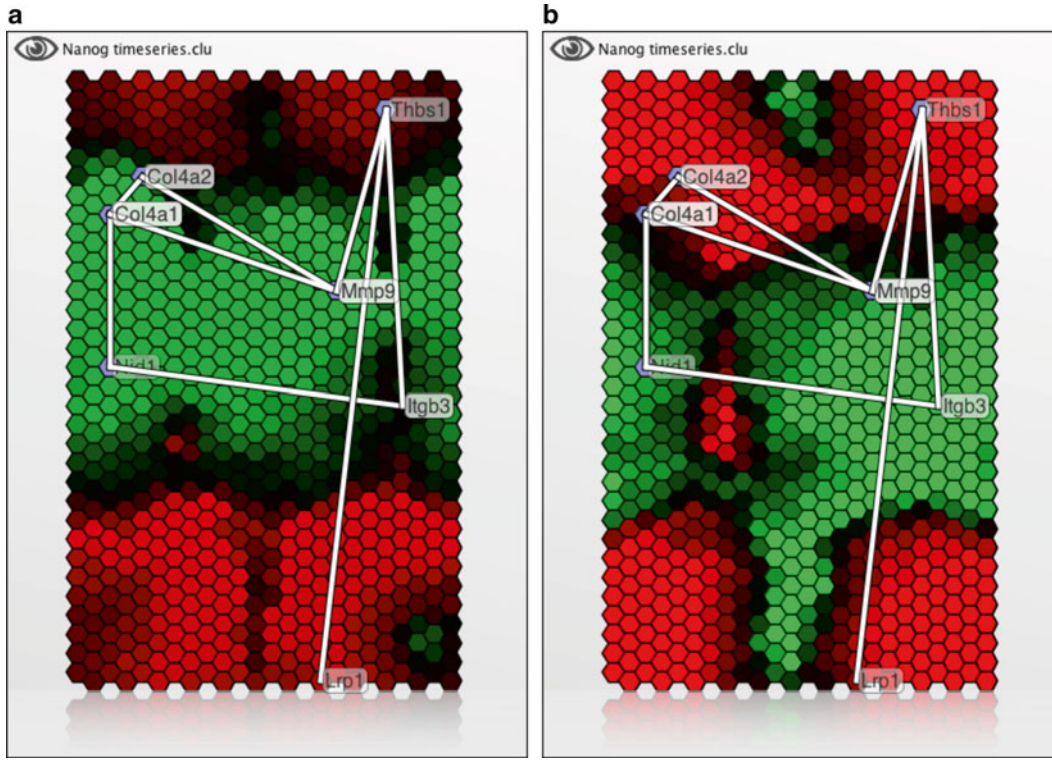


Fig. 4 GATE allows animation of expression changes of network elements. Network nodes are highlighted in *blue* and edges between nodes are shown as translucent lines for (**a**) early and (**b**) late time points from the same experiment

downregulated specifically (results are displayed in the left panel of the GATE window; *see* Fig. 2). Additionally, an unweighted GSEA-like enrichment can also be performed. The GSEA-like method is based on a running sum statistic, which takes gene sets throughout the whole list of genes into account.

3.5 Network Analysis

In addition to allowing users to investigate enrichment against background datasets, GATE also allows users to use known regulatory networks to highlight functional interactions between components directly on the GATE movie grid. When overlaid on the GATE movie, known interactions are highlighted as translucent links and Entrez gene IDs are displayed (*see* Fig. 4). This feature provides users with the ability to correlate temporal changes in expression patterns with known regulatory interactions. Four different network files are included in the GATE installation. These networks are literature curated and focus on protein-protein interactions between pluripotency-associated proteins. However, GATE allows interrogation of expression series against any background network, as long as it is in Cytoscape [9]-compatible

simple interaction file (.sif) format. Once an interaction network has been loaded, interactions (edges) between elements of interest may be visualized by simply highlighting the appropriate areas on the GATE array. In order to ensure maximum compatibility with existing network analysis software, networks between elements of interest can be exported to Cytoscape [9] from the "Load/Save" button on the left-hand side panel of the GATE window.

3.6 Exporting Data and Movies

Once an interesting set of genes has been identified and selected, exporting the list as a text file can be achieved through the "Save Selected Genes" button at the very bottom of the right panel of the GATE window (*see* Fig. 2). Snapshots of individual movie frames can also be saved. The top menu provides the appropriate "Take Screenshot" function under the "Export" tab. Additionally whole GATE movies may be exported in either QuickTime movie-format .mov (which enables playback in most video players and the web) or Adobe ActionScript 3 format (which can be used to create vector graphics movies that can be efficiently embedded in websites or PDF documents). Both options are available from the "Export" menu.

3.7 Conclusions

A particular problem encountered when analyzing high-dimensional time-series is their representation in a format that is appropriate, accessible, and amenable to further analysis. GATE is a computational software platform for integrated visualization and analysis of complex expression time-series that addresses these issues. Given a high-dimensional time-series, GATE employs a simple clustering algorithm that creates movies of expression dynamics by assigning individual time-series to hexagons on a hexagonal array and dynamically coloring each hexagon according to the expression of the molecular species to which it is associated. This procedure allows presentation of complex datasets as movies in a format that is immediately accessible. Furthermore, GATE movies are interactive and allow flexible interrogation of experimentally derived time-series data against databases of prior biological knowledge, thus linking observed molecular dynamics to known or putative regulatory mechanisms.

References

1. GATE Website (2009) http://amp.pharm.mssm.edu/maayan-lab/gate.htm. Accessed 28 May 2013
2. MacArthur BD, Lachmann A, Lemischka IR et al (2009) GATE: software for the analysis and visualization of high-dimensional time series expression data. Bioinformatics 26:143–144
3. Lu R, Markowetz F, Unwin RD et al (2009) Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature 462: 358–362
4. Huang DW, Sherman BT, Tan Q et al (2007) DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35:W169–W175
5. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a

knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550

6. Subramanian A, Kuehn H, Gould J et al (2007) GSEA-P: a desktop application for gene set enrichment analysis. Bioinformatics 23: 3251–3253
7. The Broad Institute of MIT and Harvard (2013) Gene set enrichment analysis wiki—data formats instructions. http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats. Accessed 28 May 2013.
8. The Broad Institute of MIT and Harvard (2013) Molecular signatures database collections. http://www.broadinstitute.org/gsea/msigdb. Accessed 28 May 2013
9. Shannon P (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504

Chapter 8

Interpreting and Visualizing ChIP-seq Data with the seqMINER Software

Tao Ye, Sarina Ravens, Arnaud R. Krebs, and Làszlò Tora

Abstract

Chromatin immunoprecipitation coupled high-throughput sequencing (ChIP-seq) is a common method to study in vivo protein–DNA interactions at the genome-wide level. The processing, analysis, and biological interpretation of gigabyte datasets, generated by several ChIP-seq runs, is a challenging task for biologists. The seqMINER platform has been designed to handle, compare, and visualize different sequencing datasets in a user-friendly way. Different analysis methods are applied to understand common and specific binding patterns of single or multiple datasets to answer complex biological questions. Here, we give a detailed protocol about the different analysis modules implemented in the recent version of seqMINER.

Key words Protein–DNA interactions, Chromatin immunoprecipitation coupled high-throughput sequencing (ChIP-seq), seqMINER, Genome-wide, Visualization, Multiple datasets, k-means clustering

1 Introduction

Chromatin immunoprecipitation (ChIP) is a powerful technique for mapping of protein–DNA interactions inside a cell [1]. In combination with high-throughput sequencing (ChIP-seq), the localization of posttranslationally modified histone proteins, histone variants, transcription factors, or histone modifying enzymes can be determined at a genome-wide scale [2, 3]. The method is based on formaldehyde cross-linking of protein–DNA complexes in living cells, following cell lysis and shearing of DNA into 200–700 base pair (bp) fragments. In case the protein of interest (POI) is very stably associated with the DNA, such as histone proteins, the cross-linking step can be skipped. Next, an antibody targeting the POI will pull down the actual DNA binding sites. After reverse cross-linking, the ChIP-ed DNA fragments are purified and quantitatively analyzed by high-throughput sequencing. For further bioinformatics analysis of all experiments, it is important to include a control sample. This can

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_8, © Springer Science+Business Media New York 2014

be either DNA, treated like the immunoprecipitated DNA (Input), or "mock" ChIP-ed DNA, using a nonspecific antibody for the ChIP (i.e., IgG control).

Each ChIP-seq run leads to millions of short reads (tags) and it is challenging to process, analyze, and interpret the large amount of data. The first part of high-throughput sequencing analysis uses common processing pipelines, which involves the alignment of raw reads to the genome, data normalization, and identification of enriched signal regions (Peak calling) [4]. In the second stage, individual programs allow detailed analysis, biological interpretation, and visualization of ChIP-seq results.

To provide a more complex picture of biological processes in a cell, many studies aim to compare different datasets obtained by ChIP-seq. Since the protein–DNA interactions studied by ChIP represent only POI binding at the moment of cross-linking, studies apply cellular differentiation models to compare different factors or chromatin modifications in a system with dynamic transcriptional changes. All these require analytical and computational modeling techniques, which compare multiple sequencing datasets in one cell type or the differential binding of factors in various cell types.

seqMINER has been designed to analyze multiple, or single, ChIP-seq datasets of different factors like transcription factors, chromatin-modifying enzymes, or histone modifications [5]. It is a user-friendly software, which can be used to analyze specific and common binding patterns of different factors. In addition, seqMINER helps to understand differential binding patterns of one factor in more than one cell type. Here, we provide a detailed protocol for the usage of the seqMINER platform.

1.1 Overview of Software Tools Required for the Generation of Appropriate seqMINER Input Files

To successfully run the seqMINER software, we will first give a small overview about available software tools required for the generation of appropriate seqMINER input files. The mapping of short reads against the reference genome is typically the first step to analyze the obtained ChIP-seq data. The most popular software tools are BOWTIE [6] and BWA [7]. For the manipulation and storage of read alignments or the generation of SAM/BAM formats, SAMtools [8] are recommended. To identify high-confidence binding sites of a ChIP sample, peak calling algorithms calculating the enriched tag densities over the background noise are applied. The peak calling methods normally include the normalization between the ChIP and control samples. Since the identification of signal peaks is a central task in interpreting ChIP-seq results, many peak calling algorithms have been established. The most common methods are MACS [9], SICER [10], or FindPeaks 4.0 [11].

The seqMINER software applies different analysis methods to highlight general as well as specific patterns in a given dataset.

All methods require a set of reference coordinates, which can be either a list of ChIP-seq enrichment clusters (peaks) of a particular factor. Alternatively, transcription start sites (TSSs) of genes, transcription termination sites (TTSs), or whole gene coordinates can be used. This might be of interest to analyze, for instance, the binding patterns of RNA polymerase II (Pol II) at the start or the end of genes. For the data collection, all middle points of the reference coordinates are calculated and the read densities of multiple aligned read dataset are collected in a defined window around the reference coordinates. The signal enrichment status of these multiple tracks can be analyzed through two different algorithms. In the Density Array Method, the created matrix of tag densities around the reference coordinates is reorganized by k-means clustering. This will create different groups with similar genomic features, which can be visualized and further analyzed in a heatmap graphical interphase. However, this clustering method does not allow the comparison of quantitative changes between multiple datasets. The enrichment-based method allows the one to one comparison of enrichment values between two different datasets. In this case the enrichment values are presented in a scatter plot and a table, which can be exported for further analysis. Besides, additional options have been integrated into the platform to allow individual analysis of datasets or clustering results. Since these options have been described previously, the newest seqMINER release implements an Annotation system.

2 Materials

2.1 Operating System and Software Requirements

The seqMINER software is suitable for any operating system which has a Java Runtime Environment (version 6 or above) installed, such as Linux, OS X (≥10.5), or Microsoft Windows.

For 4–5 datasets of ten million reads, seqMINER can be used on a local computer with a 32-bit operating system having 2–4GB random access memory (RAM) or on a 64-bit operating system, which will have almost no limit for memory usage.

Still it is recommended to run seqMINER over a server since it equipped normally with more RAM. This will decrease the analysis time and allow increasing the amount of data to be analyzed. In this case, a local X-windows service should be installed (e.g., Xming for Windows).

2.2 Installation

Users should first check or download the Java Runtime Environment from http://www.java.com/. The last version of seqMINER can be downloaded from http://sourceforge.net/projects/seqminer/. More information is available at http://bips.u-strasbg.fr/seqminer/, which is under General Public License (GPL3). The downloaded file needs to be unzipped.

To launch seqMINER under windows, double-click the seqMINER.bat file (*see* **Note 1**).

To launch seqMINER by a terminal:

Go to the seqMINER folder

>cd seqMINER_folder

Launch the application

>java -Xmx1500m –jar seqMINER.jar

2.3 Files and Formats

seqMINER supports BED, SAM/BAM, or default Bowtie output file formats as input files (for file format information: https://genome.ucsc.edu/FAQ/FAQformat.html). Two different types of input files are required: (1) a reference coordinate file and (2) aligned read files. The reference coordinate file can be generated with any peak detection algorithm, e.g., MACS [9], SICER [10], or FindPeaks 4.0 [11]. It is recommended to directly use the summits file of the peak detection analysis, which represents the summit points of identified peaks. On the other hand, it is possible to take RefSeq or Ensembl transcription start sites or whole RefSeq or Ensembl genes as reference coordinates. These files can be extracted from annotation databases or downloaded from http://sourceforge.net/projects/seqminer/files/Reference%20coordinate/. The second input files are the aligned raw read files of all datasets, which can be analyzed through different methods described in the following section.

3 Method

seqMINER is designed to carry out basic ChIP-seq data analysis in an easy and fast way to answer biological questions. The platform allows the comparison between a reference set of genomic positions (reference coordinate file) and multiple ChIP-seq datasets (aligned raw read file).

3.1 Standard Analysis

The standard analysis is separated in three steps, which can be followed at the java interface from the left to the right. A screenshot of the java interface with already uploaded datasets is presented in Fig. 1.

3.1.1 Step 1: File Selection

Over the browser button or under “File,” the files are chosen.

1. One reference coordinate file is required. As an example, the refGene mm9 seqMINER files is taken (Fig. 1).
2. Several aligned read files can be selected. Here we used already published datasets for the histone marks H3K27me3 (GSM307619) and H3K36me3 (GSM307619).

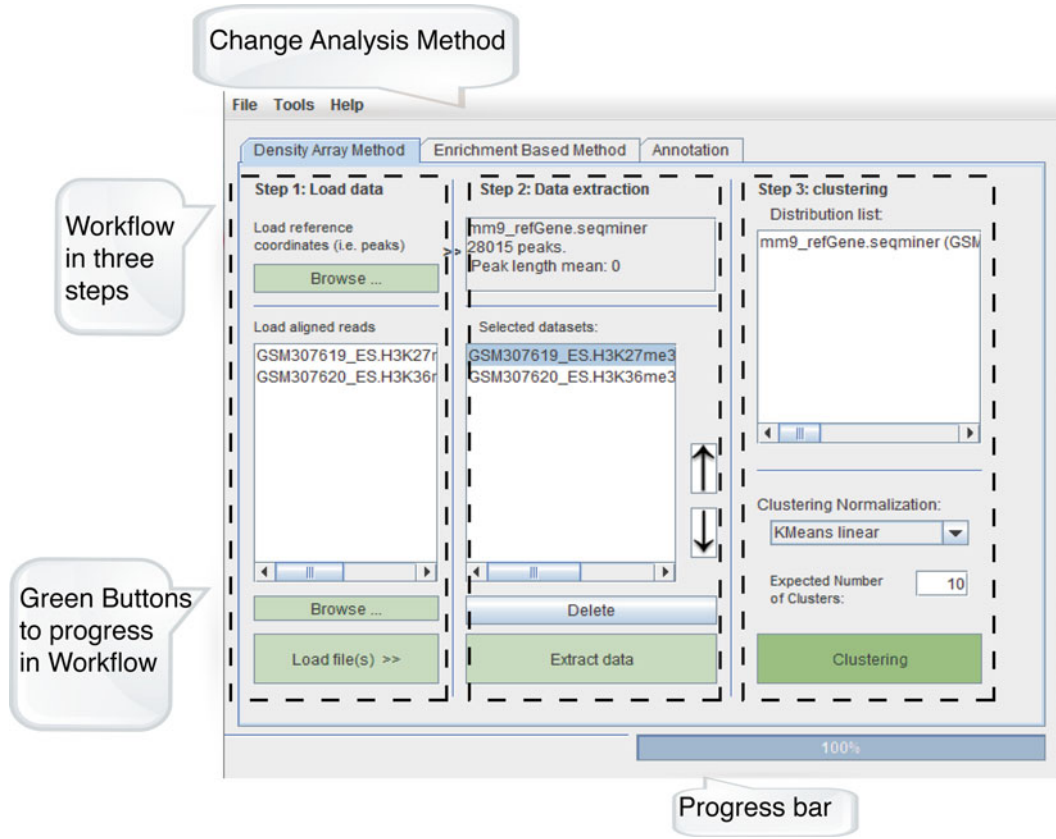


Fig. 1 seqMINER interface with analysis methods and workflow. The *green* buttons are required for data upload and progressing in the workflow from the *left* to the *right*. In the *upper* part the analysis methods can be changed to Density Array Method, Enrichment-Based Method, or Annotation. As reference coordinate, the mm9_refGene.seqminer file with 28015 peaks is uploaded. Aligned read files are taken from the Gene Expression Omnibus database (GSM307619, GSM307620)

All required files can be loaded simultaneously or one by one through the "Load file(s)" button (*see* **Note 2**).

3.1.2 Step 2: Extract Data

After loading all files, it is possible to change their order, which will be later presented in the heatmap, through the flash buttons. In addition, the "Delete" button erases already loaded datasets. At this step optional parameters for the different methods should be defined (*see* Subheadings 3.3 and 3.4). Finally, the green "Extract data" button extracts the tag densities of the datasets according to the reference coordinates for further analysis (*see* **Note 3**).

3.1.3 Step 3: Clustering and Data Visualization

The 3rd step varies depending on the method to be used to analyze the signal enrichment in multiple tracks. They can be chosen in the upper part of the java interface (Fig. 1). The methods are discussed in the following sections.

3.2 Density Array Method

This method collects the read densities over a window around a reference coordinate. The identification of groups with similar features is conducted through k-means clustering.

seqMINER proposes two normalization methods (linear and ranked normalization), which are directly applied in the clustering step. Of note, the non-normalized data is represented in the final visualization step. The k-means clustering method of interest can be chosen under clustering normalization. The default algorithm is set to k-means raw (*see* **Note 4** below). It is recommended to use k-means raw for a single datasets and k-means linear or ranked for multiple datasets. The number of expected clusters is 10 by default. It is possible to define a higher or lower value.

By clicking the "Clustering" button, the analysis is started (*see* **Note 5** below). Visualization of the results is achieved through a heatmap (Fig. 2), which will automatically appear in a separated interface (*see* **Note 6**). Based on their biological meaning, the generated clusters can be reorganized. In Fig. 2 for this the given cluster needs to be selected and shifted with the flash buttons.

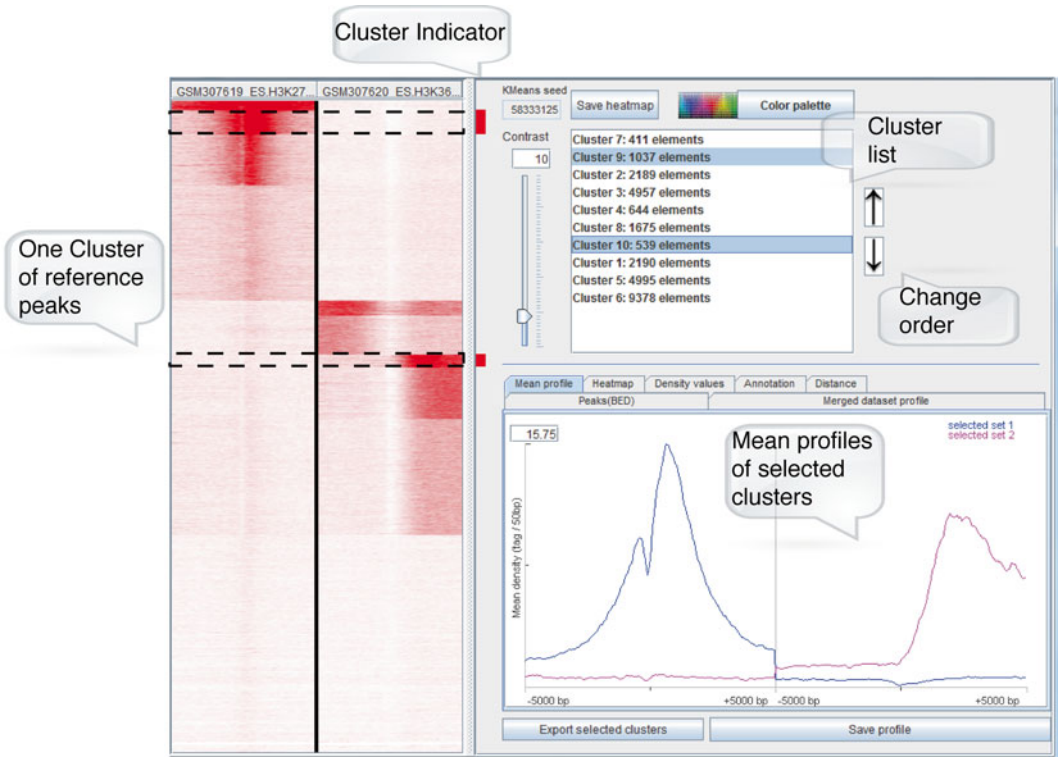


Fig. 2 Visualization of data after clustering with the Density Array Method. The *left* site represents a heatmap with different clusters of reference peaks. These clusters are listed on the *right*, whereas the order can be changed according to the biological relevance. In addition, the mean profiles of the selected clusters (two) are shown on the lower *right panel*

By using the shift or control (Ctrl) key, many clusters can be selected at the same time. The generated heatmap can be exported with "Save heatmap" as a png file.

It is possible to investigate each generated cluster or functional group. Under "Peaks" (bed) the reference coordinated for the selected cluster is found and can be saved with export selected cluster as a bedfile. Following the lower right panel in the Cluster Heatmap interface, a merged dataset profile of all selected clusters is depicted. This average profile represents the mean tag density for one, or multiple, selected cluster over the defined analysis window around the peaks. Mean profiles compare distributions around the peak middle points of the selected cluster(s). Different clusters will be presented in different curves. In Fig. 2, an example of the H3K27me3 and H3K36me3 mean profiles is depicted. H3K27me3 has a peak around the promoter, whereas H3K36me3 localizes more downstream of the promoters. In addition, the average profile can be shown as a heatmap. The heatmap represents the raw tag densities in the defined window around the reference coordinates of the selected cluster(s). "Save profile" will export the given graph as a png file. Additionally, the density values used for the generation of profiles are presented in a table (*see* **Note 7**). After copying these values into an Excel or a text file, they can be used for further analysis or plotting. The order of the density values corresponds with the clustered Peak (bed) coordinates (*see* **Note 8**).

A ChIP-seq peak annotation system is implemented in the seqMINER platform. However, it is necessary to select a genome assembly under Annotation panel (*see* **Note 9** below). For each reference coordinate, the closest gene will be annotated. Moreover, the distance of peaks to the closest TSS is represented in a bar chart under "Distance." The window size and bin size are configurable.

3.3 Enrichment-Based Method

This method allows the one to one comparison of datasets in a quantitative way. It calculates the total number of reads for each dataset at the reference coordinates. By default the analysis window is set to the peak interval. It can also be defined as a fixed interval around the calculated middle points of the reference coordinates. It is optional to load a control dataset (*see* **Note 10**). The analysis is launched with the "Calculate density array" command and a Dot Plot interface will appear. The files are compared in a Scatter Plot. The data presented at the x and y axis, as well as the coordinates can be modified manually. In the right panel, a table with all the calculated values is generated. The scatter plot and the table could be exported with "Save Image" or "Export table" commands for further analyses.

3.4 General Options

The default parameters for the peak extensions, read and clustering options are found in Tools with the shortcut Alt –O or under Tool → Options → General.

1. Peak extension: The read densities are collected around each calculated middle point of the reference coordinates, peak summits, or TSSs in a defined analysis window. The default value is 5,000 base pairs up and downstream from the middle point.
2. Read options: If there is strand information in the reference peak file, the reverse strand reference genes are automatically turned to forward strands. After calculating the peak middle point and prior to analysis, all reads are extended to 200 bp by default. In the read options, "Enable reads extensions" can be inactivated or the size of extendable reads can be modified.
3. Clustering options: These options are applied in the different algorithms and normalization procedures applied by the Density Array Method. The Wiggle step defines the clustering resolution. By default the Wiggle step will generate, for example, 10,000(bp)/50(step length) = 200 values per reference coordinate. In case there is not enough memory, the "Wiggle step" can be increased. An increase in the "Wiggle step" to, for example, 100 will result in half of the memory consuming of the clustering step. The "Max runs" indicates the numbers of clustering steps in the algorithms. The "Percentile threshold" is used in the k-means linear normalization method. All intensity values are divided by the percentile of the threshold. The "Percentile threshold" is by default 75 %, which is the third quartile of the distribution. The background values which are smaller than the T threshold will be excluded from the previous distribution. The T threshold is also applied in the ranked-based normalization method. All intensity values are sorted in ascending order. The T threshold defines the rank of the sorted intensity values, which will then be replaced by 0. Thus, all background values are considered as equal during the clustering process.

3.5 Gene Profile Option

seqMINER conducts the Density Array Method with the reference gene body. The gene profile options are found under Tools → Options → Gene profile. It is important to activate these setting before Subheading 3.1.2 (Extract data) (*see* **Note 11**).

As an example dataset, we used a Pol II dataset (GSM307823), whereas the seqMINER refGene mm9 was uploaded as a reference coordinate in Subheading 3.1.1. Usually, the reference file should contain the gene start and end coordinates. The annotation data can be downloaded from annotation databases or http://sourceforge.net/projects/seqminer/files/Reference%20coordinate/. By default each gene (reference gene body) is equally divided in 160 bins, whereas 20 extra bins are added to the upstream and downstream regions. These extra bins are by default 5,000 bp long, which can be modified through peak extensions. Then, the read

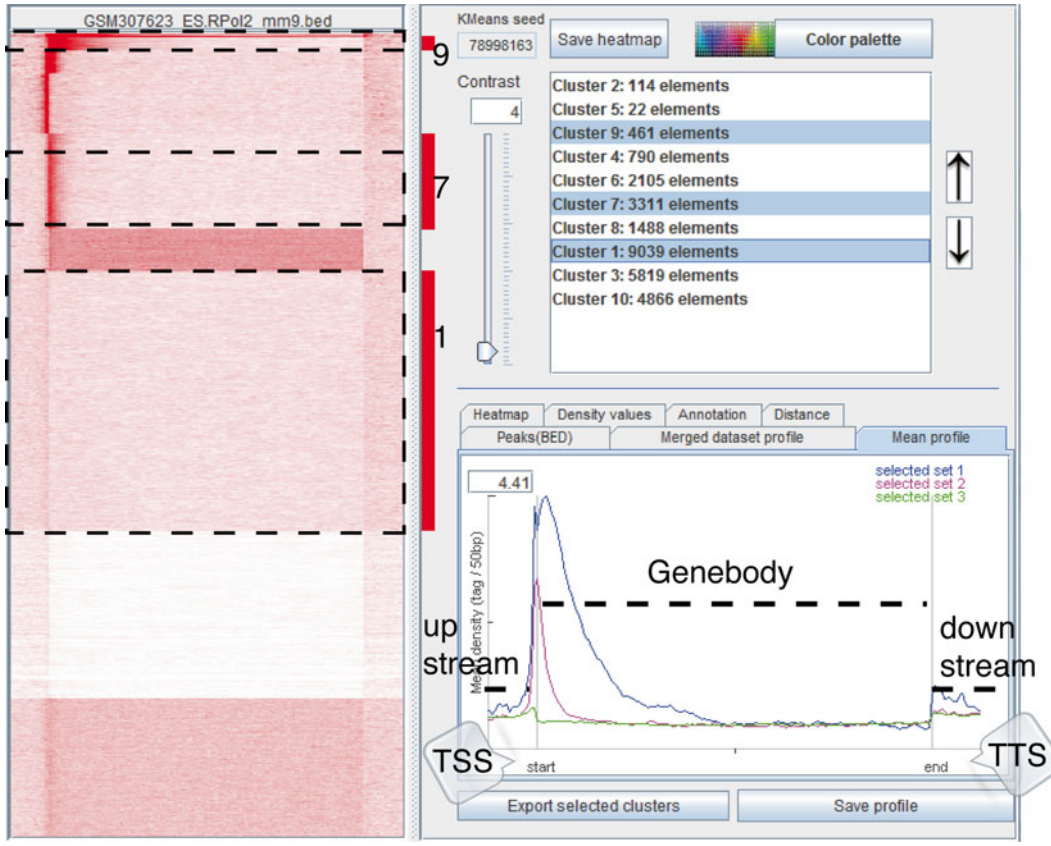


Fig. 3 Visualization of data after gene profile analysis through the Density Array Method. An RNA polymerase II aligned read file is used as an example file (GSM307623) to analyze the gene profile at the reference genes (mm9_refGene.seqminer file). On the *right panel*, the resulting heatmap of the different clusters is shown. The reference file is divided into an upstream region, the TSS, gene body, TTS, and downstream region

densities can be collected within these bins. After the clustering step, the results of the gene profile analysis are shown in a separated java interface (Fig. 3).

3.6 Re-clustering of Data

There are two possibilities of re-clustering data with the Density Array Method:

1. In case the user would like to get the same clustering results through re-clustering of the same dataset, the "Run k-means with a given seed" option under Tools → Options → General should be activated or defined. By default this k-means seed value is randomly generated by seqMINER.
2. For a better resolution of genomic features in particular clusters, obtained through the Density Array Method, seqMINER proposes a re-clustering of data. First, the reference peak bedfile of the generated cluster (s) in Subheading 3.1.3 needs to

be exported as described under Subheading 3.3. Afterward the exported bedfile can be loaded under Subheading 3.1.1 (Load) data as the reference coordinate file, following data extraction. The clustering normalization method should be set to k-means enrichment (linear) before the "Clustering" button is pressed. The data can be analyzed as described before.

3.7 Visualization of Data Without Clustering

seqMINER provides the possibility to visualize and analyze the data using the Density Array Method without prior clustering. For this the heatmap interface can be activated with the Visualization button after data extraction (Subheading 3.1.2). The "Visualization of HeatMap" button is found through a right-mouse click at the highlighted dataset in the Distribution list. In addition, all extracted read densities can be saved as a text file with "Export data." Another advantage of skipping the clustering step is that already analyzed datasets, or existing results, can be uploaded for visualization and reanalysis. In addition, it is possible to annotate all identified peaks (reference coordinates) of a dataset with a peak annotation method.

4 Notes

1. Depending on the quantity of data and the computer configuration, the maximum RAM memory, which is attributed to the Java virtual machine, requires to be increased or decreased using the option –Xmx. For a 32-bit operation system, the maximum available memory is 1.5GB, so the default parameter is -Xmx1500m. There is almost no limitation for memory usage, and it is possible to set a value higher than the physical memory for a 64-bit operating system. To change the memory usage under windows, the seqMINER.bat file needs to be opened with the text editor. Thus, the value in red can be modified: java -Xmx1500m -jar seqMINER.jar.
2. This step takes the most of the analysis time. Therefore, it is recommended to load only the required files. In addition, it is better to load the files one by one or at most two at the same time for a PC with few RAM. Under Tools → "Statistic": Information about the reference coordinate file, loaded and aligned read files, and memory usage can be found. If the button "Garbage collection" is pressed, the memory usage can be reduced.
3. Under Tools → "Statistic all extracted": Distributions with information about the elements per line and estimated memory usage are listed.
4. Different clustering normalization methods can be applied. It is recommended to use k-means raw for single dataset clustering, since there is no normalization between datasets

included. The k-means linear and ranked clustering methods should be used for the analysis of multiple ChIP-seq datasets. These methods implement normalization between datasets as described by Ye et al. [5]. The first method applies linear normalization, whereas the percentile (P) is chosen by the user. The second method is a ranked-based normalization method. The minimum threshold (T) is by default 10 and can be modified. Both parameters (P and T) can be defined under Tools → Options → General, before data extraction.

5. In case there is not enough available memory, the clustering step will take a long time or an error will be indicated. To overcome this problem, it is recommended to clear non-used datasets at the Distribution list (Subheading 3.1.3). The "Delete" option is found at the highlighted/activated datasets with the right-mouse click. In addition, the memory usage can be reduced under Tools → Statistic and "Garbage collection."
6. seqMINER normalizes the datasets before the k-means clustering. The obtained clustering results (Heatmap, density profiles) are presented with the raw sequencing files, which are not normalized.
7. To export the extracted read densities of all generated clusters, it is recommended to highlight the given dataset at the Distribution list in the analysis Subheading 3.1.3. Using the right-mouse button, "Export data" is found. This will generate a text file.
8. Since the order of lines corresponds to the Peak(bed) file, we suggest to create an excel table including the reference coordinates and density values. The columns of the density values represent the defined bins of the dataset(s). In case there are multiple datasets, five columns with the value −1 separate the datasets one by one.
9. From the seqMINER 1.3, we have added a new panel named "Annotation." This function helps us to do the peak annotation with the public databases as Refseq or Ensembl directly in seqMINER. A genome assembly could be selected before the clustering analysis. Recent human and mouse assembly annotations are already provided as a combo box. Customized annotation table could be extracted with Ensembl-biomart tool: (http://www.ensembl.org/biomart/martview/). After selecting the assembly, the attributes should be selected in the following order:

   Chromosome Name
   Gene Start (bp)
   Gene End (bp)
   Ensembl Gene ID
   Associated Gene Name
   Strand

Finally the result can be exported in text format. User can browse the file within the "Advanced" button popup panel or copy the file into the "lib" folder under the unzipped seqMINER folder by adding an extension of ".seqminer" manually for permanent usage.

10. When a control dataset (c) is added, the enrichment values (q) of the analyzed dataset (d) are calculated as described by Ye et al. [5]. This q value can be log(2) transformed.
11. seqMINER does not normalize the average gene profile to the length of the genes. seqMINER divides the distance before and after the gene body in 20 bins, which is fixed for each reference coordinate. In contrast, the distance in the gene body is not in a fixed interval. Thus, the 160 bins of each reference coordinate differ in the number of base pairs. Therefore, small genes like ribosomal or histone genes will appear with artificial high values in the Heatmap.

References

1. O'Neill LP, Turner BM (1996) Immunoprecipitation of chromatin. Meth Enzymol 274:189–197
2. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680
3. Ku CS, Naidoo N, Wu M et al (2011) Studying the epigenome using next generation sequencing. J Med Genet 48:721–730
4. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6(11 Suppl):S22–S32
5. Ye T, Krebs AR, Choukrallah MA et al (2011) seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 39(6):e35
6. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
7. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
8. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
9. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
10. Zang C, Schones DE, Zeng C et al (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25: 1952–1958
11. Fejes AP, Robertson G, Bilenky M et al (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24:1729–1730

Chapter 9

A Description of the Molecular Signatures Database (MSigDB) Web Site

Arthur Liberzon

Abstract

Annotated lists of genes help researchers to prioritize their own lists of candidate genes and to plan follow-up studies. The Molecular Signatures Database (MSigDB) is one of the most widely used knowledge base repositories of annotated sets of genes involved in biochemical pathways, signaling cascades, expression profiles from research publications, and other biological concepts. Here we provide an overview of MSigDB and its online analytical tools.

Key words Bioinformatics, Database, Genomics, Molecular sequence annotation, Gene expression profiling

1 Introduction

Many genomic and molecular biology studies report their findings in the form of gene lists. For example, the list could consist of genes belonging to a biochemical pathway, or of genes associated with a disease phenotype, etc. Png genes in such lists helps to figure out what they are doing in the cell. In addition, considering many genes as a group increases power of many popular analytical methods, as has been pioneered in Gene Set Enrichment Analysis (GSEA) [1]. Tools like GSEA rely on well-annotated collections of gene sets. In fact, we have originally developed the Molecular Signatures Database (MSigDB) to supply gene sets for GSEA. MSigDB quickly gained popularity as a stand-alone knowledge-base resource that can also be used independently of GSEA. Here, we will review main features of MSigDB and the accompanying suite of Web tools.

MSigDB is one of the largest and most widely used databases of gene sets [2]. Its most recent version, v4.0, released on May 31, 2013, contains 10,925 gene sets. Gene sets in MSigDB are lists of genes (in no particular order, each gene occurs only once in the set) with annotations and links to external sources.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_9, © Springer Science+Business Media New York 2014

GSEA/MSigDB Web site has several key components listed below. In this manual, our primary focus will be on the MSigDB home page.

1.1 Login/Register

The link is located in the upper right corner, near logo of the Broad Institute. If you have not registered before, you can do it by clicking this link. Registration is free for noncommercial users—its only purpose is to track usage for reports to our funding agencies. You can also register by following the corresponding link in Subheading 3.1 on GSEA or MSigDB home pages.

1.2 Navigation Tools

1.2.1 GSEA/MSigDB Navigation Strip

The horizontal navigation strip serves to quickly move throughout the entire site. Use links in this strip to move between these pages: GSEA home, Downloads, MSigDB home, Documentation, and Contact information. This navigation strip runs through all pages, except for the Documentation wiki. The navigation strip has links to the following pages:

GSEA Home

http://www.broadinstitute.org/gsea/index.jsp

This page provides a registration link and displays latest news about GSEA and MSigDB; information about members of the GSEA team, members of the scientific advisory board, and about funding agencies; and contact and acknowledgements to our contributors. Finally, there is also a note explaining how to cite use of GSEA.

Downloads

From this page, you can download GSEA software and MSigDB database files.

Molecular Signatures Database

http://www.broadinstitute.org/gsea/msigdb/index.jsp

This page provides an overview of MSigDB Web site with links to its Web analytics. It has an additional navigation menu on the upper left side to move around MSigDB pages. This page also includes directions for registration; notes about current version of MSigDB database and the Web site; information about members of the GSEA team, members of the scientific advisory board, and about funding agencies; contact and acknowledgements to our contributors; overview of MSigDB collections; contact information; and a note explaining how to cite use of MSigDB (*see GSEA Home* in Subheading 1.2.1 above).

Documentation

This link leads to pages with detailed documentation about the GSEA/MSigDB resource. It contains GSEA User Guide and Tutorial, FAQs, descriptions of data formats, release notes, and other documents.

Contact

Here you will find how to send us questions, suggestions, and other feedback information. For convenience, this information is reproduced at the end of GSEA and MSigDB home pages.

1.2.2 MSigDB Side Bar

This navigation menu appears on every MSigDB Web page, is restricted to these pages only, and allows you to quickly move around the entire MSigDB web site.

MSigDB Home

Clicking this link will bring you to the MSigDB home page described *Molecular Signatures Database* in Subheading 1.2.1 above.

About Collections

This link will bring up detailed information about individual collections and subcollections of MSigDB. The page is organized as a table with three columns, where the first column contains collection names and a link showing total number of sets in the collection. The middle column contains detailed description of each collection and recommendations for its use. The third column contains links to download gene sets as GMT files (*see* **Note 1**).

Browse Gene Sets

Clicking this link will let you browse gene sets by their name or collection (*see* Subheading 3.2 for details).

Search Gene Sets

Clicking this link will bring you to the search tool (*see* Subheading 3.3 for details).

Investigate Gene Sets

Click this to navigate to a suite of gene set analysis tools (*see Gene families* in Subheading 3.4).

View Gene Families

This link will direct you to a table with a functional overview of all MSigDB sets categorized into a small number of gene families. Gene families are special collections of gene products that share a common feature such as homology or molecular function.

Help

Follow this link to learn more about features listed in the MSigDB navigation side menu.

2 Materials

To explore MSigDB, you will need a computer with the Internet connection and a Web browser. To access MSigDB, registration is required. Registration is free for noncommercial users. Its only purpose is to help us track usage for reports to our funding agencies. After registering, you can log in at any time using your e-mail address.

3 Methods

3.1 Registration

Register to view and explore MSigDB contents. To register, go to the MSigDB home page (http://www.broadinstitute.org/gsea/msigdb/index.jsp) or GSEA home page (http://www.broadinstitute.org/gsea/index.jsp) and click on the word **register**

in the Registration section. Alternatively, you can click on the word **register** at the top of these pages, near the Broad Institute logo (*see* Subheading 1.1).

3.2 Browse

Here you can browse gene sets by their name or collection. You can get there by clicking **Browse** in the **Overview** section of the MSigDB home page or by clicking **Browse Gene Sets** at the side bar navigation menu. This page has three parts.

In the top part, you can search by **gene set name**. For example, to search for sets with names containing word **stem**, type it in the form and click the **search** button. To see the results, scroll down to the end of the page. There are Web links to pages matching the search term in set names.

Alternatively, you can browse sets by first letter or number. For example, click on the letter **B** and then scroll down to the end of the page. There will be Web links to pages of sets with names starting from letter **B**. Note that the second set there is named **B_CELL_DIFFERENTIATION**.

Finally, you can browse sets by collection. To see a short info about each (sub) collection, move mouse over the question mark icon near it. For example, go to **C5** and click on **BP**. In the long list of links displaying all sets in the C5 BP subcollection, locate the link to **B_CELL_DIFFERENTIATION**.

3.3 Search

Here you can find gene sets by keyword, gene set name, collection, organism, or contributor. You can get to this tool by clicking the **Search** link in the Overview section of MSigDB home page. Alternatively, you can get there by clicking **Search Gene Sets** at the side bar menu. Accessing this tool requires registration. You can do the search by typing a keyword in the form below the word **Keywords** and click the **search** button. For details about keyword search, move your mouse over the question mark icon. You can apply a number of filters to your search using the scroll bar forms under **Search Filters**. For example, to search for sets containing word "stem," type it in the keywords form and click "search." Search results appear below the red line which reads "found 439 gene sets" in this case. The results appear in a table with up to 10 rows by default and six columns. To see next pages of the results, you can use navigation page numbers above the table. You can vary the number of hits per page from 10 (default) to 20, 50, or 100. For each gene set, the columns display its **name**, number of genes (**#genes**), **description**, its collection code (**collections**), **organism**, and **contributor** organization. Click on rows to select gene sets for subsequent action. For example, click on the row with **B_CELL_DIFFERENTIATION**. This set has 12 genes and comes from C5 BP subcollection. Notice that the results now show **1** gene set selected. To un-select, click the row again. Note that the number of sets selected reverts to 0. Click the set to select it again.

At the top of the table, there is a menu **Select An Action…** Clicking and holding it shows a number of action options available. Top five choices allow you to export selected gene set in a variety of standard GSEA formats.

3.4 Examine

To get detailed information about a single gene set, click a link with the set's name. You can come across these during your explorations in Subheadings 3.2 or 3.3. For example, go to the **Browse** tool by clicking **Browse Gene Sets** on the side menu. Then type **B_CELL_DIFFERENTIATION** in the form **Gene set name** and click search button. Scroll down the page to see the results—there should be only one link to set named **B_CELL_DIFFERENTIATION**. To examine the set, click on this link. This will bring you to a standard gene set page that has detailed description of the individual set. The page has four major sections with all fields described in detail in the documentation (*see* **Note 2**).

The first section consists of the set's annotations. All sets have unique database identifiers and names and include brief and full descriptions. We use HUGO gene symbols and human Entrez Gene IDs as universal gene identifiers. We also preserve original identifiers as they appeared in the gene set source. Other annotations depend on the type of gene set. Annotations linking to external resources are particularly important as they allow researchers to place the sets in the context of their origins [2].

The second section describes how to use MSigDB analysis tools to further investigate the set.

3.4.1 Download Gene Set

It allows you to download the set in one of our standard file formats. Gene set file formats are described in the **Documentation** section here: http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#Gene_Set_Database_Formats. The file contains a single gene set made of human gene symbols (*see* **Note 3**). Navigate to the gene set page for **B_CELL_DIFFERENTIATION** as described in Subheading 3.4 and download the set in GRP format by clicking on "format: grp." This should have downloaded the **B_CELL_DIFFERENTIATION** set in the GRP format. The specific location depends on the download settings of your browser.

3.4.2 Compute Overlaps

Examining genes shared by two sets can highlight common processes and reveal other useful relationships. This tool evaluates the overlap of a query gene set and an estimate of the statistical significance of the overlap with one or more MSigDB collections or subcollections (*see* **Note 4**). Overlaps can only be done against current version of MSigDB gene sets. Overlaps are computed using human gene symbols and the tool does any required conversion automatically.

Follow the steps in the first paragraph of Subheading 3.4 to navigate to the gene set page for **B_CELL_DIFFERENTIATION**. To display the results, click on a link to a gene set collection, e.g., CP:canonical pathways (*see* **Note 5**).

The results have three parts. At the top, there is a table summarizing conversion details from input gene identifiers to human gene symbols. The next part lists summary statistics for the overlaps as a table with rows corresponding to the number of MSigDB collections chosen for the analysis and the following columns:

Collections—indicates collection of sets selected

Overlaps—lists the number of overlapping gene sets (*see* **Note 6**)

Gene sets in collections—lists the total number of sets tested

Genes in comparison (n)—lists the number of genes in your set

Genes in Universe (N)—list the number of all known human gene symbols

The following table reports detailed overlap statistics for each gene set. A link above this table allows you to export the results as an Excel file. The table lists one set per row and has the following columns:

Gene Set Name [# Genes (K)]—link to the gene set page [number of genes]

Description—brief description of the set

Genes in overlap k/k (k)—number of genes in the overlap

p-value—the significance of the overlap according to the hypergeometric distribution [3]

FDR *q*-value—the significance estimate after correcting for multiple hypothesis testing [4]

The final part contains the overlap matrix where rows are genes from the query set and columns are links to the overlapping sets.

3.4.3 Compendia Expression Profiles

This tool displays a profile of the gene set based on a selected compendium of expression data, such as human tissue compendium (Novartis) [5], global cancer map (Broad Institute) [6], or NCI-60 panel of cell lines (National Cancer Institute) [7].

Follow the steps in the first paragraph of Subheading 3.4 to navigate to the gene set page for **B_CELL_DIFFERENTIATION**. To display the results, click on a link to a compendium.

Alternatively, you can submit the set to this tool by clicking the further investigate link near the **Advanced query**. This will lead you to the **Investigate Gene Sets** page and your set will be pasted to the **Gene Identifiers** box. Choose one of the available compendia and

click on "display expression profile." The resulting heat map includes dendrograms clustering gene expression by gene and samples. Genes are indicated by probe set id, gene symbol, description, and gene family.

3.4.4 Gene Families

Gene families are special collections of gene products that share a common feature such as homology or molecular function. This feature highlights particularly interesting members of a set.

Follow the steps in the first paragraph of Subheading 3.4 to navigate to the gene set page for **B_CELL_DIFFERENTIATION**. Click on the Categorize link to retrieve an overview of the set with its members categorized into the gene families.

Alternatively, you can submit the set to this tool by clicking the further investigate link near the **Advanced query**. This will lead you to the **Investigate Gene Sets** page and your set will be pasted to the **Gene Identifiers** box. Click on "show gene families" to retrieve an overview of the set with its members categorized into the gene families.

3.4.5 Advanced Query

This action submits the set to a suite of analysis tools at **Investigate Gene Sets** page. See Subheading 3.4 for details.

4 Notes

1. We provide three kinds of GMT files depending on what type of gene identifiers is used to make gene sets. Thus, "original identifiers" correspond to whatever gene identifiers were used in the original source of the set. We provide these files for reference only and do not recommend using them for standard analyses because various sets can have different types of original identifiers. On the other hand, "gene symbols" correspond to the GMT file with sets made of human gene symbols. We use our own system to map all kinds of original identifiers to the space of human gene symbols. When the original identifiers stand for genes from species other than human, we map them to the corresponding orthologous human gene symbols. Biologists are familiar with human gene symbols and we thus recommend using GMT built from human genes symbols for most purposes. Computationally, working with gene symbols can present certain challenges because different genes can have the same symbols and the same gene can have a number of alternative gene symbols. For programmatic access, it is safer to work with more robust gene identifiers, such as NCBI Entrez Gene IDs. We thus provide a version of GMT files with genes made of human Entrez Gene IDs as well.

2. Detailed description of this page and its features is here: on the top horizontal navigation strip, click on **Documentation**. This will bring you to the documentation wiki pages. On the left side of this wiki page, go to section **msigdb** and click on **Guide to a GeneSetCard**.
3. To download many gene sets, click **Downloads** on the navigation strip and then locate a desired GMT file that contains your sets of interest. There you can also choose what kind of gene identifiers to have in the GMT file. Alternatively, click **Molecular Signatures Database** on the navigation strip and then go to **Search**. After the search is done, select as many sets as needed and then click on **Select An Action…** and export the sets in desired file format. This option will export sets as human gene symbols only.
4. To compute overlaps between several collections of gene sets, you should submit the set to this tool by clicking the further investigate link near the **Advanced query**.
5. From the **Investigate Gene Sets** page, click on the **compute overlaps** button to display the results.
6. By default, the report displays the 10 gene sets in the collection that best overlap with your gene set. If you compute overlaps from the **Investigate Gene Sets** page, you can choose the number of overlapping gene sets to display in the report by varying the top number of sets in the pull-down menu (10, 20, 50, or 100) or by changing the FDR q-value threshold.

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550
2. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27(12):1739–1740
3. Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4):401–417
4. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57(1):289–300
5. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101(16):6062–6067
6. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98(26):15149–15154
7. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginow D, Cossman J, Kaldjian EP, Scudiero DA, Petricoin E, Liotta L, Lee JK, Weinstein JN (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol Cancer Ther 6(3):820–832

Part III

Transcriptional Networks in Embryonic and Adult Stem Cell

Chapter 10

Use of Genome-Wide RNAi Screens to Identify Regulators of Embryonic Stem Cell Pluripotency and Self-Renewal

Xiaofeng Zheng and Guang Hu

Abstract

Embryonic stem cells (ESCs) are characterized by two defining features: pluripotency and self-renewal. They hold tremendous promise for both basic research and regenerative medicine. To fully realize their potentials, it is important to understand the molecular mechanisms regulating ESC pluripotency and self-renewal. The development of RNA interference (RNAi) technology has revolutionized functional genetic studies in mammalian cells. In recent years, genome-wide RNAi screens have been adopted to systematically study ESC biology, and have uncovered many previously unknown regulators, including transcription factors, chromatin remodelers, and posttranscriptional modulators. Here, we describe a method for the identification of regulators of ESC pluripotency and self-renewal using RNAi screens, as well as assays for further validation and characterization of the identified candidates. With modifications, this method can also be adapted to study the fate specification events during ESC differentiation.

Key words Embryonic stem cell, Self-renewal, Pluripotency, RNAi, Genome-wide, Genetic screen, Reporter assay

1 Introduction

ESCs are derived from the inner cell mass of the blastocyst stage embryos. They have two distinctive characteristics: the ability to differentiate into any cell type of the three germ layers, known as pluripotency, and the ability to proliferate while maintaining the pluripotent state, known as self-renewal. Because of these properties, ESCs present a unique opportunity to advance many aspects of biology and medicine, such as mammalian development, disease modeling, drug screening, and stem cell therapies [1–3]. To successfully use ESCs for these research and clinical applications, it is critical to understand the mechanisms that control their pluripotency and self-renewal. A mechanistic model of pluripotency and self-renewal will help to elucidate the molecular pathways involved in early embryonic development. It will also facilitate the efficient

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_10, © Springer Science+Business Media New York 2014

derivation, maintenance, and expansion of pluripotent stem cells, and provide guidance to the generation of desired cell types for therapeutic purposes.

It is known that ESC pluripotency and self-renewal is governed by the combination of signal-transduction pathways, transcription factors, epigenetic modifiers, and posttranslational regulators [4]. However, until recently, most of the current knowledge came from early studies that relied on candidate approaches or bootstrapping strategies, and novel classes of important self-renewal genes are continued to be discovered. Recent advances in RNAi technologies made it possible to interrogate gene function by loss-of-function screens on a genome-wide scale. To systematically study pluripotency and self-renewal, we and others have carried out large-scale RNAi screens in ESCs using various readouts from cell morphology, proliferation, to reporter assays [5–11]. These screens identified many novel genes that play critical roles in self-renewal, as well as several protein complexes such as the Ccr4-Not, Tip60-p400, Paf1, and cohesion-mediator complexes [6–8, 11, 12]. The success of these screens illustrates the power of RNAi and forward genetics in the study of pluripotency and self-renewal, and paves a path for functional genetic studies of ESC fate specification in the future.

Here, we describe the method we employed for the genome-wide RNAi screen in ESCs [8]. In the initial screen, a fluorescence reporter assay based on the Oct4GiP cells is used to determine the consequence of silencing individual genes on self-renewal. The Oct4GiP cells express the enhanced green fluorescent protein (EGFP) under the control of the Oct4 gene promoter. Oct4 is highly expressed in ESCs and quickly downregulated during differentiation. As a result, EGFP expression in the Oct4GiP cells faithfully correlates with the ESC state, and fluorescence-activated cell sorting (FACS) analysis can be used to determine the self-renewal status of the cells at the single cell level. With this assay, the function of each gene in self-renewal and pluripotency can be quickly assessed by RNAi using a genome-wide siRNA library. Genes whose silencing leads to the loss of the reporter expression are likely to have important roles in ESCs and can be quickly identified. Finally, additional assays such as the alkaline-phosphatase (AP) staining and reverse-transcription quantitative polymerase chain reaction (RT-qPCR) of lineage markers are used to further confirm and characterize the screen hits. The method described here can be carried out either manually or with automation by the liquid-handling instruments for increased throughput. With modifications, such as using different reporter cell lines and culture conditions, it can also be adapted to screen for genes involved in other aspects of ESC fate specification.

2 Materials

2.1 Mouse ESC Culture

1. Water-jacketed CO_2 tissue culture incubator (Thermo).
2. Tissue culture hood (biological safety cabinet).
3. Tissue culture plates or flasks.
4. 0.1 % gelatin (Sigma) in water.
5. PBS without Ca or Mg.
6. 0.05 % trypsin.
7. Hemocytometer.
8. M15 medium: DMEM (Invitrogen) supplemented with 15 % ES-qualified fetal bovine serum (FBS), 1,000 U/ml ESGRO (Millipore), 1x nonessential amino acids (Invitrogen), 1× EmbryoMax Nucleosides (Millipore), and 10 μM β-mercaptoethanol.

2.2 Genome-Wide RNAi Screen in ESCs

1. 384-well PCR plates (Thermo Scientific).
2. 384-well pipette tips (Thermo Scientific).

2.2.1 siRNA Transfection

3. Flat-bottom 384-well tissue culture plates (Corning).
4. 2–10, 5–50, 50-300 μl multichannel pipette (Thermo Scientific)
5. Multichannel aspirator (Corning) or vacuum wand (VP-Scientific).
6. Microplate dispenser (Thermo Scientific Wellmate microplate dispenser or equivalent).
7. Liquid-handling system (Agilent Velocity 11 or equivalent).
8. OptiMEM (Invitrogen).
9. Lipofectamine 2000 (Invitrogen).
10. Genome-wide siRNA library (Thermo Scientific mouse siGenome library or equivalent).
11. M15 medium.

2.2.2 FACS Analysis

1. Multichannel aspirator (Corning) or vacuum wand (VP-Scientific).
2. 2–10, 5-50, 50-300 μl multichannel pipette (Thermo Scientific).
3. BD LSRII FACS analyzer with the HTS unit or other similarly equipped FACS analyzer.
4. Reagent reservoirs (Thermo Scientific).
5. 0.25 % trypsin.
6. PBS + 10 % FBS

2.3 Hit Validation

2.3.1 Alkaline-Phosphatase Staining

1. Zeiss Axiovert 40 CFL with Axiocam MRC Camera or equivalently equipped microscope.
2. 96-well and 24-well tissue culture plates (Corning)
3. AP staining kit II (STEMGENT).
4. PBST: 1× PBS, 0.05 % Tween 20.

2.3.2 RT-qPCR Analysis

1. CFX384 real-time thermal cycler (Bio-Rad or equivalent).
2. 96-well and 24-well tissue culture plates (Corning).
3. 96-wells or 384-wells PCR plate (Bio-Rad).
4. Aurum Total RNA Mini Kit (Bio-Rad or other RNA extraction kits).
5. iScript cDNA synthesis kit (Bio-Rad or other reverse-transcription kits).
6. SsoFast EvaGreen Supermix (Bio-Rad or other SybrGreen qPCR mix).

3 Methods

The screen is carried out in three phases: the primary screen, the secondary screen, and hit validation (Fig. 1a). In the primary screen, the Oct4GiP ESCs are transfected with the SMARTpool siRNAs from the mouse siGenome library in 384-well plates (Fig. 1b). The effect of individual gene silencing is examined by the Oct4GiP reporter assay to identify candidate genes that are important for pluripotency and self-renewal. In the secondary screen, the four individual siRNAs of the SMARTpools against the primary hits are rescreened with the reporter assay, and only genes that score again are considered positive hits. Finally, in hit validation, the gene silencing efficiency for the siRNAs identified in the secondary screen is verified, and the positive hits are tested with two additional assays, the AP staining and the lineage marker expression analysis, to further confirm their roles in ESCs.

In all the steps described below, follow the general guidelines for good cell culture practice to avoid contaminating the cells.

3.1 Oct4GiP ESC Culture

1. Coat tissue culture plates with 0.1 % gelatin: Add the 0.1 % gelatin solution (0.1 ml gelatin solution/cm^2) to 10-cm tissue culture plates or other tissue culture vessels of choice. Incubate at room temperature for 30 min or longer.
2. Plate Oct4GiP ESCs: Remove the gelatin solution and plate ~$2 \times 10^4/cm^2$ Oct4GiP cells in 10 ml of the M15 medium in each 10-cm plate. Culture the cells in a tissue culture incubator at 37 °C, 5 % CO_2. This step can be scaled up or down based on the number of cells needed.

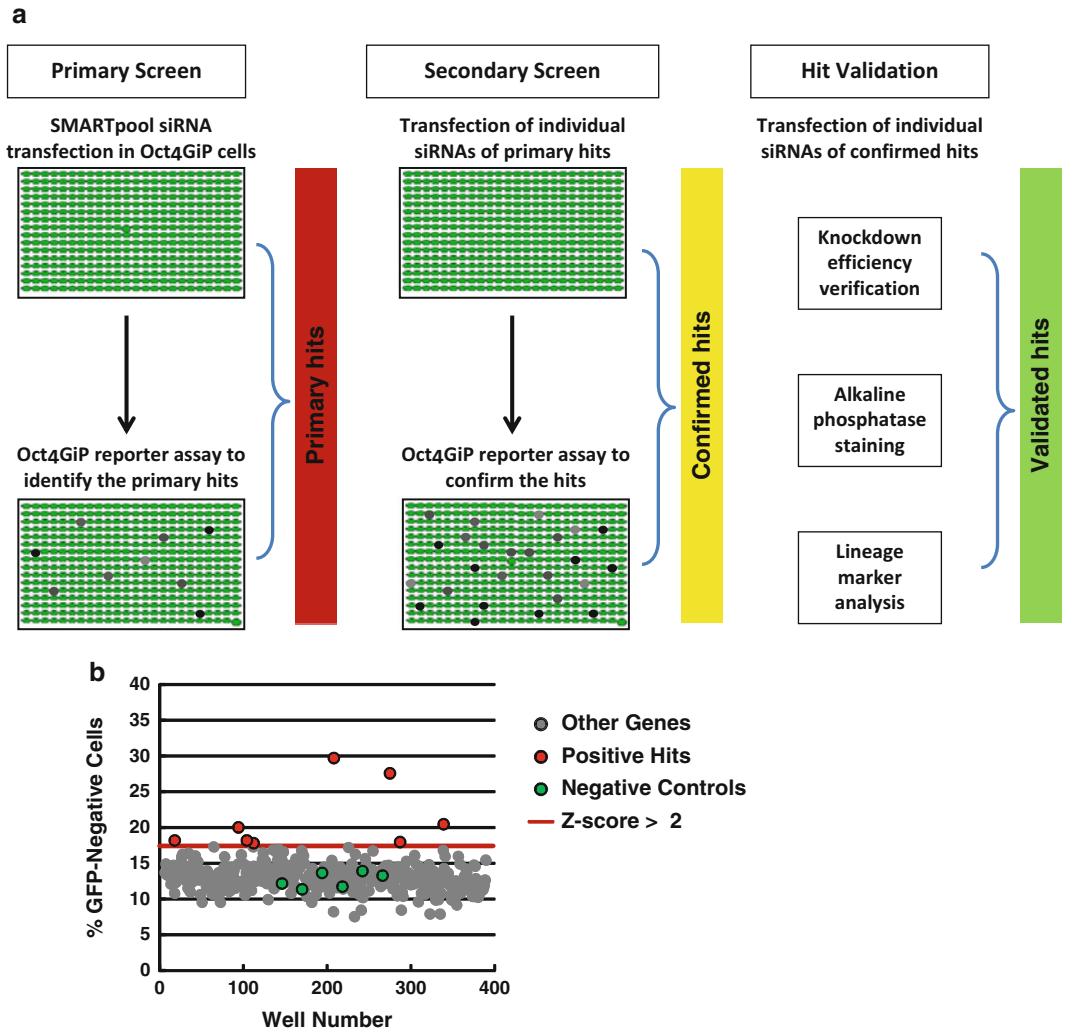


Fig. 1 The overall procedure for the genome-wide RNAi screen in ESCs. (**a**) In the primary screen, the Oct4GiP ESCs are transfected with the siGenome SMARTpool siRNAs in 384-well plates. Four days after transfection, the primary hits are identified by FACS using the Oct4GiP reporter assay. In the secondary screen, the cells are transfected with individual siRNAs against the primary hits, and genes corresponding to the siRNAs scored again are considered confirmed hits. Additional assays such as knockdown efficiency verification, alkaline-phosphatase staining, and lineage marker expression analysis are carried out to further validate the hits. (**b**) A representative result of the Oct4GiP reporter assay for one 384-well plate from the primary screen. *Red*: primary hits; *Green*: negative controls included on the plate (nontargeting siRNAs); *Gray*: other genes on the plate. *Red line*: two standard deviations from the plate mean

3. Change the medium every day.
4. Split the cells every 2–3 days (*see* **Note 1**): Remove medium. Rinse cells with PBS, and dissociate the cells with 1 ml 0.05 % trypsin for each 10-cm plate at room temperature for ~5 min. Neutralize the trypsin by adding fresh M15 medium, 5 ml for

each 10-cm plate, and dissociate cells into single cell suspension by repeated pipetting (*see* **Note 2**). Collect cells in a 15-ml conical tube by centrifugation at 1,000 rpm for 5 min. Remove supernatant and resuspend the cells in fresh M15 medium. Count the cells on a hemocytometer or with an automated cell counting device and plate in 10-cm plates as described above.

3.2 Genome-Wide RNAi Screen in Oct4GiP ESCs

3.2.1 siRNA Transfection

1. Transfection mixture assembly: Prepare a master mix of OptiMEM and Lipofectamine 2000 at a ratio of 80:1 (vol/vol). Aliquot 10 μl of the master mix to each well of the 384-well PCR plates and incubate at room temperature for 5 min. Add 2 pmol siRNA from the siRNA library working stock to the OptiMEM-Lipid mixture in each well (*see* **Notes 3** and **4**), and incubate for another 15 min to allow the formation of the siRNA-Lipid complexes. This step can be carried out either manually or with robotic automation depending on the scale of the screen (*see* **Notes 5** and **6**).
2. Transfection of the Oct4GiP cells: Coat 384-well tissue culture plates with 0.1 % gelatin before assembling the transfection mixture. Harvest the Oct4GiP cells and resuspend in fresh M15 medium to 7×10^4 cells/ml. Remove gelatin from the 384-well plates, and aliquot 30 μl of the cell suspension to each well (*see* **Note 7**). Transfer the siRNA-Lipid mixture to the cell suspension in each well and mix by pipetting up and down 3–5 times. Use new pipette tips for each plate of transfections. Move the plates to the tissue culture incubator.
3. Medium change: Change medium every 24 h. Remove medium using the multichannel aspirator or vacuum wand, and add 40 μl fresh M15 medium. Be careful not to scratch the cells at the bottom of the plates. Return the plates to the incubator.

3.2.2 FACS Analysis

1. Cell dissociation: Remove medium and rinse the cells with PBS. Add 10 μl 0.25 % trypsin to each well in 384-well plate and incubate at room temperature for ~5 min with occasional agitation. Visually inspect to ensure complete detachment of the cells. Add 30 μl of PBS with 10 % FBS to neutralize the trypsin and dissociate cells into a single cell suspension by repeated pipetting. Prepare one plate at a time to avoid cell aggregation and keep the plate on ice until ready for FACS analysis.
2. FACS analysis: Analyze the cells on the BD LSRII FACS analyzer using the HTS unit. Adjust PMT voltage and threshold to correctly capture the cells in the forward vs. side-scatter plot. Set the HTS unit in high-throughput mode and analyze 10 μl of cell suspension from each well (*see* **Note 8**). Prepare the next plate during each run.

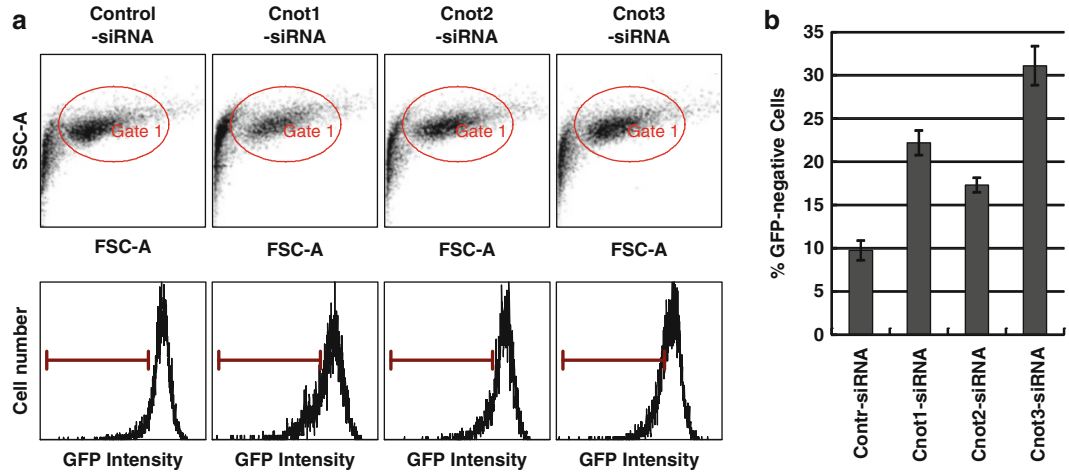


Fig. 2 The Oct4GiP reporter assay. Oct4GiP ESCs are transfected with the Control-, Cnot1-, Cnot2-, or Cnot3-siRNAs in 96-well plates and FACS analyzed 4 days after transfection. (**a**) Live cells are first gated in the forward vs. side-scatter plots, and % GFP-negative cells are then determined from the histograms of the GFP channel. (**b**) % GFP-negative cells from independent biological replicates are plotted as mean ±SEM ($n=6$)

3. Data analysis: Use the lipid-only or control-siRNA transfected cells as controls to set the gates for the analysis. First, gate for the live cell population in the forward vs. side-scatter plot. Next, gate for the GFP-negative cells in the GFP channel: set the gate so that ~10 % of the cells appear to be GFP-negative in the controls (Fig. 2a). Determine the percentage of GFP-negative cells in each well (Fig. 2b). In the primary screen, genes are scored as positive hits if the corresponding SMARTpool siRNA increased the percentage of GFP-negative cells by two standard deviations from the plate average (Fig. 1b). In the secondary screen, genes are scored as positive hits if any of the individual siRNAs increased the percentage of GFP-negative cells by two standard deviations from the average of the control wells (*see* **Note 9**).

3.3 Hit Validation

3.3.1 Analysis of the Gene Silencing Efficiency

1. siRNA transfection: Coat 24-well tissue culture plates with 0.1 % gelatin. Assemble the siRNA-Lipid complexes similarly as described above, by mixing 50 µl of OptiMEM, 1.5 µl Lipofectamine 2000, and 25 pmol siRNA (siRNAs against the positive hits from the secondary screen) in 96-well U-bottom plates. Include lipids-only and nontargeting siRNAs as controls. Aliquot 0.5 ml Oct4GiP cells at 2×10^5 cells/ml in M15 medium in each well of the gelatin-coated 24-well plates, and add the siRNA-Lipid mixture to the cells. Mix well and transfer the plates to the incubator. Change medium the next day.
2. RNA extraction and reverse transcription: 48 h after transfection, remove the medium from the cells, and lyse cells directly with the lysis buffer from the Aurum Total RNA Mini Kit.

Extract total RNAs from the transfected cells according to the instructions of the kit, and use 1 μg of the total RNA to set up reverse transcription with the iScript cDNA synthesis kit.

3. Real-time quantitative polymerase chain reaction (RT-qPCR): Design RT-qPCR primers for the target genes using online tools such as the Primer3Plus. Add the cDNAs, primers, and the Ssofast Evagreen supermix in 384-well RT-PCR plates and perform PCR in the real-time thermocycler. Use housekeeping genes such as *β-Actin* or *Gapdh* for normalization, and determine the gene silencing efficiency by comparing the relative expression of the target genes in cells transfected with lipids-only or nontargeting siRNAs to those transfected with gene-specific siRNAs. Effective siRNAs should lead to greater than 60 % reduction in target gene expression (Fig. 3a).

3.3.2 AP Staining and Lineage Marker Expression Analysis

1. siRNA transfection: Coat 96-well tissue culture plates with gelatin. Assemble the siRNA-Lipid complexes similarly as described above using 10 μl of OptiMEM, 0.6 μl Lipofectamine 2000, and 10 pmol siRNA in 96-well U-bottom plates, and include lipids-only and nontargeting siRNAs as controls. Aliquot 100 μl Oct4GiP cells at 4×10^5 cells/ml in M15 medium in each well of the gelatin-coated 96-well plates, and add the siRNA-Lipid mixture to the cells. Mix well and transfer the plates to the incubator.
2. Replating the cells: The next day, remove the medium and rinse the cells in the 96-well plates with PBS. Add 25 μl 0.25 % trypsin to each well and incubate at room temperature for ~5 min. Neutralize trypsin with 100 μl M15 medium in each well, and dissociate cells into single cell suspension with repeated pipetting. From each well, transfer 60 μl of the cell suspension to one well of a gelatin-coated 24-well plate, and transfer the other 60μl to another 24-well plate, generating two replicates from each transfection (*see* **Note 10**). Add 0.5-ml/well M15 medium to the 24-well plates. Mix well and transfer the plates to the incubator. Culture the cells for a total of 3 days from the day of replating in the M15 medium with medium change every day.
3. AP staining: 4 days after transfection, use one set of the replicate 24-well plates for AP staining. Carry out the staining with the AP staining kit II from STEMGENT. Take pictures of the stained cells using the Zeiss Axiovert 40 CFL microscope equipped with the Axiocam MRC Camera. Cells transfected with lipids-only or nontargeting siRNA should show typical ESC morphology as compact and dome-like colonies with strong AP staining. Those transfected with siRNAs against pluripotency genes should differentiate and appear flat and dispersed with much reduced staining (Fig. 3b).

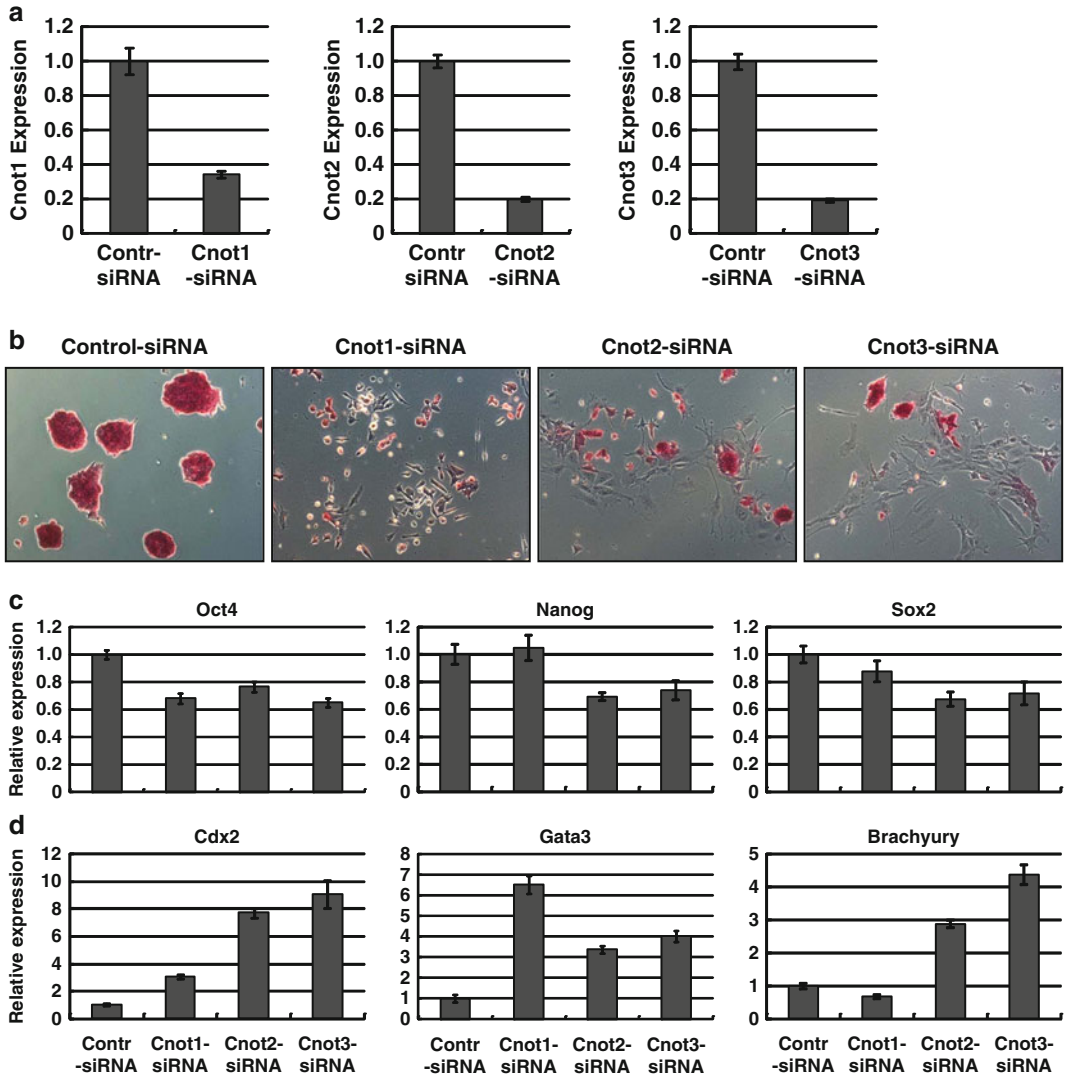


Fig. 3 Validation of the screen hits. (**a**) Oct4GiP cells are transfected with the indicated siRNAs, and Cnot1, Cnot2, and Cnot3 mRNA expressions are determined by RT-qPCR 2 days after transfection. (**b**) Oct4GiP cells are transfected with the indicated siRNAs and replated the next day. Cells are stained by AP staining 4 days after transfection (**c**), and the expression of lineage markers are determined RT-qPCRs (**d**)

4. Analysis of lineage marker expression: Extract total RNA from the cells in the other set of the replicate 24-well plates using the Aurum Total RNA Mini Kit. Prepare cDNAs from 1 μg total RNA with the iScript cDNA synthesis kit. Carry out RT-qPCR for pluripotency markers such as *Oct4*, *Nanog*, and *Sox2*, as well as differentiation markers such as *Cdx2*, *Gata3*, and *Brachyury*, using housekeeping genes such as *β-Actin* or *Gapdh* for normalization. Cells transfected with siRNAs against pluripotency genes should show reduced expression of the pluripotency markers and/or increased expression of the differentiation markers (Fig. 3c).

4 Notes

1. Do not let ESC culture become over-confluent, as over-confluency can result in increased ESC differentiation.
2. Do not over-trypsinize ESCs to avoid clumping and loss of cell viability. Completely dissociate the cells into single cell suspension before plating to avoid heterogeneity in colony size and quality during subsequent cultures.
3. To reduce the edge effect, fill the edge wells in the 384-well plates with PBS and do not use them for the screen.
4. In each 384-well plate, assign designated wells for the following controls (two wells for each control): lipids-only, nontargeting siRNA, *Oct4* siRNA, *Plk1* siRNA. The lipids-only and nontargeting siRNA wells serve as negative controls. The *Plk1* well serves as a positive control for the transfection, as *Plk1* is essential and its downregulation causes cell death that can be easily detected. The *Oct4* well serves as a positive control for the screen, as *Oct4* is required for ESC pluripotency and self-renewal.
5. For large-scale screens, the transfection step is usually carried out with liquid-handling systems and dispersers such as the Velocity 11 and the Wellmate. For small- to medium-scale screens, it can be performed by manual pipetting with multichannel pipettes.
6. It is recommended that the siRNA screen is carried out in duplicate or triplicate to reduce the false-positive rate.
7. For siRNA transfections, the optimal cell plating density is $\sim 2 \times 10^4$ cells/cm^2 in general. But this number may require additional optimization by titration. Low plating density usually leads to poor cell survival during transfection, and high plating density causes high background due to increased spontaneous differentiation.
8. It takes about 1 h to complete the FACS analysis for each 384-well plate.
9. Reporter assays based on other reporter ESCs, such as the Nanog-GFP cells, may be used after the secondary screen to quickly validate and further narrow down the positive hits.
10. For genes that dramatically affect cell growth or viability, set up the initial transfections in multiple wells and pool them for replating to compensate for cell loss.

References

1. Evans M (2011) Discovering pluripotency: 30 years of mouse embryonic stem cells. Nat Rev Mol Cell Biol 12:680–686
2. Murry CE, Keller G (2008) Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell 132:661–680
3. Smith AG (2001) Embryo-derived stem cells: of mice and men. Annu Rev Cell Dev Biol 17:435–462
4. Young RA (2011) Control of the embryonic stem cell state. Cell 144:940–954
5. Chia NY, Chan YS, Feng B, Lu X, Orlov YL, Moreau D, Kumar P, Yang L, Jiang J, Lau MS, Huss M, Soh BS, Kraus P, Li P, Lufkin T, Lim B, Clarke ND, Bard F, Ng HH (2010) A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468:316–320
6. Ding L, Paszkowski-Rogacz M, Nitzsche A, Slabicki MM, Heninger AK, de Vries I, Kittler R, Junqueira M, Shevchenko A, Schulz H, Hubner N, Doss MX, Sachinidis A, Hescheler J, Iacone R, Anastassiadis K, Stewart AF, Pisabarro MT, Caldarelli A, Poser I, Theis M, Buchholz F (2009) A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. Cell Stem Cell 4:403–415
7. Fazzio TG, Huff JT, Panning B (2008) An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. Cell 134:162–174
8. Hu G, Kim J, Xu Q, Leng Y, Orkin SH, Elledge SJ (2009) A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev 23:837–848
9. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR (2006) Dissecting self-renewal in stem cells with RNA interference. Nature 442:533–538
10. Zheng X, Hu G (2012) Oct4GiP reporter assay to study genes that regulate mouse embryonic stem cell maintenance and self-renewal. J Vis Exp Pii:3987
11. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA (2010) Mediator and cohesin connect gene expression and chromatin architecture. Nature 467:430–435
12. Zheng X, Dumitru R, Lackford BL, Freudenberg JM, Singh AP, Archer TK, Jothi R, Hu G (2012) Cnot1, Cnot2, and Cnot3 maintain mouse and human ESC identity and inhibit extraembryonic differentiation. Stem Cells 30:910–922

Chapter 11

Correlating Histone Modification Patterns with Gene Expression Data During Hematopoiesis

Gangqing Hu and Keji Zhao

Abstract

Hematopoietic stem cells (HSC) in mammals are an ideal system to study differentiation. While transcription factors (TFs) control the differentiation of HSCs to distinctive terminal blood cells, accumulating evidence suggests that chromatin structure and modifications constitute another critical layer of gene regulation. Recent genome-wide studies based on next-generation sequencing reveal that histone modifications are linked to gene expression and contribute to hematopoiesis. Here, we briefly review the bioinformatics aspects for ChIP-Seq and RNA-Seq data analysis with applications to the epigenetic studies of hematopoiesis and provide a practical guide to several basic data analysis methods.

Key words Hematopoiesis, Epigenetics, Histone modification, RNA-Seq, ChIP-Seq

1 Introduction

Hematopoietic stem cells give rise to all blood cell types while maintaining a capacity of self-renewal [1]. It is known that a core set of transcription factors form a tightly regulated network that controls the spatiotemporal regulation of lineage-specific genes during hematopoiesis [2]. However, the precise molecular mechanisms governing HSC self-renewal and differentiation remain unclear.

There is an increasing awareness of epigenetic mechanisms in controlling the developmental hierarchy of hematopoietic system [3]. Genomic DNA within the eukaryotic nucleus is packaged with histones into a compact form called chromatin. The N-terminal tails of histones are subjected to a variety of posttranslational modifications including acetylation and methylation. Our previous works revealed that histone modifications correlate with gene activities, contribute to T-cell specificity/plasticity, and set stages for hematopoiesis [4–8].

Our knowledge about epigenetic regulation has been greatly advanced by recent development of genome-wide techniques

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_11, © Springer Science+Business Media New York 2014

such as ChIP-Seq and RNA-Seq. While ChIP-Seq charts genomic landscapes of transcription factor binding and histone modifications, RNA-Seq quantifies gene expressions. A combinational use of ChIP-Seq and RNA-Seq has been widely applied to the epigenetic studies of hematopoiesis [3]. Here, we have briefly reviewed the bioinformatical steps commonly used to address epigenetic questions in hematopoiesis by using ChIP-Seq and RNA-Seq.

2 Materials

2.1 Genome Annotation

Genome annotations were downloaded from the online UCSC genome browser [9]:

1. Go to http://genome.ucsc.edu/cgi-bin/hgGateway, choose genome and assembly version.
2. Click "tools" at the top of the browser, and then choose "Table Browser."
3. Specify the source of genome annotation. The "Table Browser" by default provides download of annotation for the UCSC known genes. One may choose other sources of annotation such as RefSeq and Ensembl from the "track" down-drop list.
4. Type a file name in the "output file" text filed, click the "get output" button, and save the annotation to a local drive.

2.2 Public ChIP-Seq/RNA-Seq Data

Raw fastq sequence files and/or processed BED6 files included in this review were downloaded from Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo/) [10].

2.3 Software

1. In-house C++ programs (executable files and source code are available upon request):
   (a) Sam2Bed6_Bowtie2: Convert a sam file from Bowtie2 to a BED6 file.
   (b) RemoveRedundantRead: Remove redundant reads from a BED6 file.
   (c) GenerateRPBMBasedSummary: Convert a BED6 file to a BEDGraph file.
   (d) AverageDensityAcrossGenes: Generate average read density across gene features.
   (e) RPKMCalculator: Calculate RPKM at gene level.
   (f) Cat_expr_file: Concatenate gene expression files from different samples.
   (g) SortGeneAnnoByExpr: Sort genome annotations by gene expression level.
   (h) DensityCalculatorPromoters: Calculate normalized read density for promoters.

2. Microsoft Excel: a spreadsheet editing software.
3. Microsoft Access: a database management system.
4. MeV (Multiple Experiment Viewer): a JAVA application for statistic analysis, clustering, and visualization of gene expression data [11].

3 Methods

We first reviewed several common steps for ChIP-Seq and RNA-Seq data analysis, including (1) inspection of data quality, (2) sequence alignment, (3) data visualization, (4) identification of read-enriched regions, and (5) quantification of gene expression. We then introduced a combinational usage of in-house C++ programs, Excel (Microsoft), Access (Microsoft), and MeV (Dana-Farber Cancer Institute) to correlate histone modification with gene expression. We showed examples by using public ChIP-Seq and RNA-Seq data sets generated for epigenetic studies on hematopoiesis.

3.1 Initial Data Quality Inspection

The first step in processing next-generation data is to check the sequence quality. FastQC (Babraham Institute) is a standalone Java application that outputs a summary statistics for fastq files (*see* **Note 1**). It issues warnings for bases with low quality, for bases with abnormal sequence contents, and for primer/adapter contaminations.

3.2 Sequence Alignment

There are dozens of algorithms to map short reads to a reference genome. Low-quality bases may be clipped off to increase mapping rate (the % of reads mapped to the reference; *see* **Note 2**). To minimize ambiguity, reads that are mapped to multiple positions (called multireads) are frequently discarded. Consequently, read enrichments within repetitive regions are underestimated; repetitive sequences are found in constitutive heterochromatin and segmental duplications, both with functions implicated in hematopoiesis [12]. The exclusion of multireads will also underestimate the expression of genes with multiples copies. To address this challenge, several sophisticated probabilistic methods with user-friendly tools have been proposed, for instance, as described in [13].

3.3 Data Visualization

A visualization of the ChIP-Seq and RNA-Seq data in a genome browser such as the UCSC genome browser [9] helps to further inspect the data quality. A local mirror of the UCSC genome browser may be installed. But its maintenance usually requires substantial computational resources. Thus, for a small number of samples, the "custom track" feature from the online UCSC genome browser is recommended (*see* **Note 3**). An example is illustrated below about how to upload a ChIP-Seq data set (for transcription factor GATA1)

and an RNA-Seq data set (for human CD36+ erythrocyte precursor cells 14) to the online UCSC genome browser:

1. Download the two data sets (GSM651547 and GSM651555) from GEO. The files are in sra format (short read archive). A sub-module called "fastq-dump" from the SRA Toolkit (NCBI) extracts fastq files from sra files (*see* **Note 4**).
2. Map the sequences to human genome (hg18) by using Bowtie2, which reports the alignments in sam format [15]. Bowtie2 by default reports the best hit for multireads.
3. Convert sam to BED6 file by using the in-house C++ program "Sam2Bed6_Bowtie2" (*see* **Note 5**).
4. Remove redundant reads with "RemoveRedundantRead" (*see* **Note 6**). Since the probability that two reads are mapped to the same genomic position is small for ChIP-Seq data, only one read is retained for each genomic position to minimize biases from amplification (*see* **Note 7**).
5. Generate genomic distribution of reads with "Generate RPBMBasedSummary." The program outputs a BEDGraph file, with the first three columns denoting chromosome, starting position and ending position, and the last column denoting the number of reads mapped to the genomic region (*see* **Note 8**).
6. Customize the BEDGraph file for the UCSC genome browser. The BEDGraph file acceptable by the online UCSC genome browser reserves the first line for parameters of the track (http://genome.ucsc.edu/goldenPath/help/bedgraph.html). One needs to edit the BEDGraph file from **step 5** to accommodate this requirement (*see* **Note 9**).
7. Upload the BEDGraph files to the online UCSC genome browser: (1) Go to http://genome.ucsc.edu/cgi-bin/hgGateway, choose genome (Human) and assembly version (hg18), and click "add custom tracks"; (2) click "Browse" and choose the BEDGraph file and click "Submit" (*see* **Note 10**); (3) after uploading a file, the user will be redirected to a page called "Manage Custom Tracks"; from there one may choose to upload more files or go to the genome browser.
8. Save your session (*see* **Note 11**). Click the "My Data" option on the top of the genome browser and then choose "sessions." One needs to register, if not have done, to save the session.

Figure 1 shows a screenshot from the online UCSC genome browser for the two data sets. GATA1 not only occupies the promoters of genes *HBB*, *HBD*, and *HBBP1* but also binds to the enhancer sites downstream to gene *HBE1* (top track). Consistent with their important functions in hematopoiesis, both HBB and HBD are highly expressed (second track). One great advantage of using the

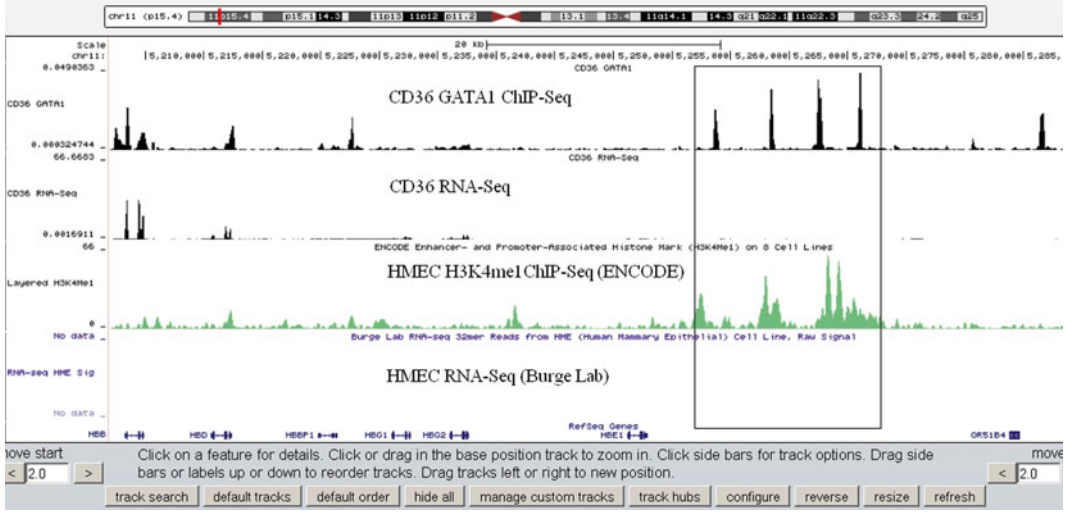


Fig. 1 A screenshot of the uploaded ChIP-Seq and RNA-Seq from the online UCSC genome browser. First track: distribution ChIP-Seq reads of GATA1 at the beta-globin domain for human CD36+ erythrocyte precursor cells. Second track: distribution of RNA-Seq reads from the same cells as in the first track. The *Y*-axis for the two tracks represents the number of tags normalized by total library size and window size. Third track: genomic distribution of H3K4me1 for human mammary epithelial cell line. Enrichment of H3K4me1 in intergenic regions is a marker of enhancer. Forth track: distribution of RNA-Seq reads for the same cells as in the third track. No read was detected in this region. The last two tracks are preexisting in the online UCSC genome browser. Enhancer regions are marked by *rectangle*

online UCSC genome browser is its ability to integrate NGS data sets preexisting in the browser, including those from the ENCODE project. For instance, from the "Encode Regulation …" section, the enhancer regions in the beta-globin domain in erythrocyte precursor cells are marked by H3K4me1 (an enhancer signature) in the distinctive human mammary epithelial cells (third track). However, the enhancers are not likely active, because the nearby globin genes are all silent (forth track).

3.4 Identification of Read Enriched Regions

A common step in analyzing ChIP-Seq data is to identify the genomic regions enriched with mapped reads. The general idea is to test whether the number of tags with a genomic region is significantly more than those generated from a background model. An initial check of the read distribution from the genome browser helps to tell whether the read-enriched regions are broad or narrow. While different methods have been developed to address each situation, a combinational usage of the methods is not uncommon in literatures [16].

Identification of read-enriched regions is also justified for RNA-Seq data under certain circumstances. It is known that reads from the 3′-end of an RNA molecule are more likely sampled than those from the 5′-end, especially for single cell RNA-Seq [17]. In this situation, normalizing the read count within a gene by the

size of read-enriched regions rather than simply by the gene length would improve the quantification of gene expression. The 3′-end-biased sequencing data provide valuable information on the exact ending positions of transcripts, of which the boundaries can be defined by read-enriched islands.

3.5 Gene Expression Quantification

The abundance of mRNA of a gene is quantified by RPKM (the number of *r*eads *p*er *k*ilobases of exon model per *m*illion reads) for RNA-Seq, which normalizes the length of RNA species and sequencing depth [18]. The expression level can be measured at both gene and isoform levels and the choice is project specific. The differentially expressed (DE) genes are identified by examining whether or not the difference in read counts between two conditions is significantly higher than expectation. Different probabilistic distributions are proposed to model read count from RNA-Seq data, including Poisson and negative binomial, with representative tools such as edgeR [19].

3.6 Average Density Profile from ChIP-Seq Data

We previously generated a large number of ChIP-Seq data sets for histone methylations and acetylations in human hematopoietic stem cells, erythrocyte precursors, $CD4^+$ T cells, and B cells [5–8]. Analysis of these data sets revealed that different histone modifications show distinctive preferences in genomic localizations. A plot for the average density of reads for a histone modification surrounding and across genic regions helps to reveal its localization preference. Below is an example for how to obtain the average distribution of H3K4me3 across a genic region from human hematopoietic stem cells:

1. Download the BED file for the H3K4me3 ChIP-Seq data from GEO (GSM317587) [6]. Note that the BED file is based on hg18.
2. Download genome annotation from the online UCSC genome browser following instructions in Subheading 2.1 (choose Human and assembly version hg18).
3. Calculate the average density of H3K4me3 across genes by using an in-house C++ program "AverageDensityAcrossGenes": It divides the promoter region (TSS ± 2 Kbps) into 20 equal size bins, separates gene body region into 10 fractions, and extends to 2 Kbps after TES (10 equal size bins) (*see* **Note 12**). It outputs the density for each bin/fraction in a flat text file.
4. Visualize the average density with any spreadsheet software such as Excel (Fig. 2a).

3.7 Correlating Histone Modifications with Gene Expression Level

After sorting genes based on their expression levels into equal size groups, two strategies are introduced to visualize the correlation between histone modifications and gene expression: (1) Plot the average density profile of a histone modification (*see* **Note 13**), and (2) visualize the read density by using heatmap.

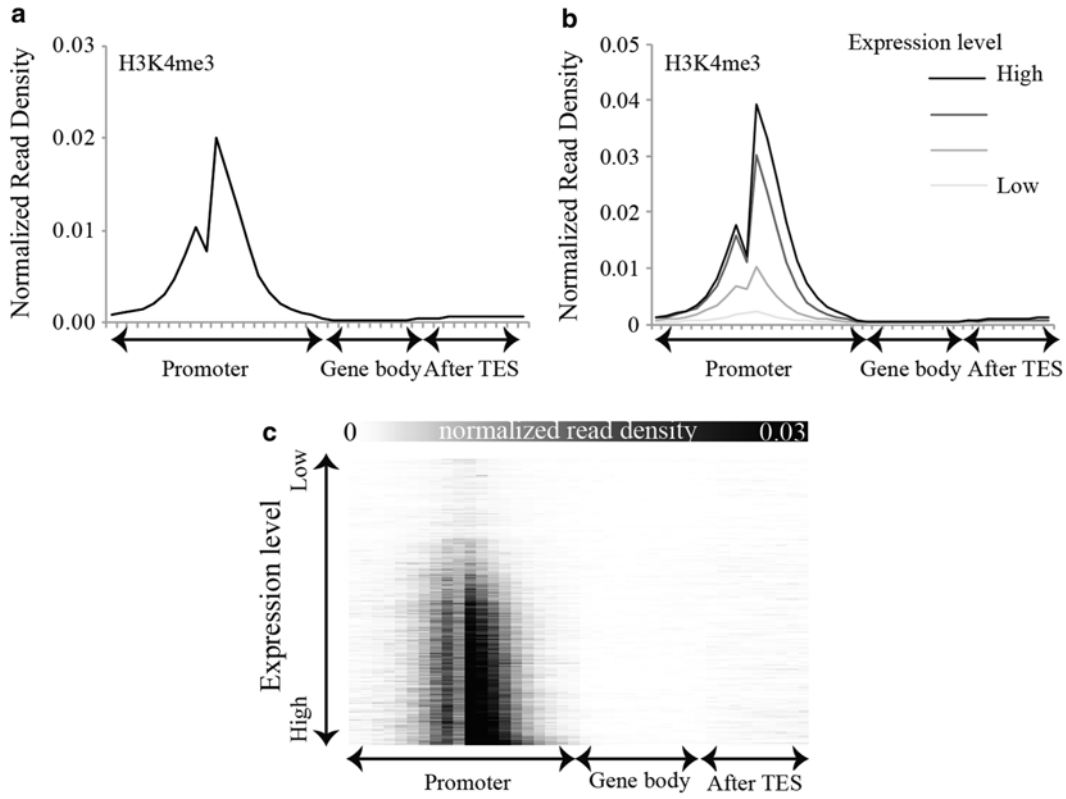


Fig. 2 H3K4me3 read density at promoters positively correlates with gene expression level. (**a**) Average H3K4me3 read density across promoter region, gene body region, and region 2 Kbps downstream to transcription ending site (TES). The promoter region (TSS ± 2 Kbps) is divided into 20 equal size bins. Gene body region, excluding the first 2 Kbps, is separated into ten fractions. Region downstream of TES is divided into 10 equal size bins. (**b**) Similar to panel **a**, except that the average density is plotted independently for four groups of genes, sorted based gene expression levels. (**c**) The average density is visualized as a heatmap. Genes were sorted into 200 equal size groups by gene expression level. Each row corresponds to a group of genes. Each column corresponds to a bin/fraction of genomic region as defined in **a**

Below is a step-by-step guide about how to visualize H3K4me3 densities across groups of genes sorted by gene expression level (both from human hematopoietic stem cells):

1. Download the ChIP-Seq BED file for H3K4me3 (GSM317587) [6] from GEO and genome annotation (hg18) from the online UCSC genome browser.
2. Download the RNA-Seq BED file from GEO (GSM651554) [14].
3. An in-house C++ program "RPKMCalculator" calculates expression at gene level by taking a BED file and a genome annotation file (*see* **Note 14**).
4. Sort gene annotations with the expression file from **step 3** with "SortGeneAnnoByExpr."

5. Set the desired number of groups as the last input parameter of the in-house C++ program: "AverageDensityAcrossGenes." It will output a file containing a matrix of average read density with each row corresponding to a bin or fraction and each column corresponding to a gene group. The matrix can be visualized by any spreadsheet software such as Excel (Fig. 2b). If the number of groups is large, the density matrix can be imported into and visualized as a heatmap by MeV (Fig. 2c).

3.8 Visualization of Histone Modifications and Gene Expression During Hematopoiesis

The dynamics of histone modifications and gene expression during early stages of T-cell development was extensively characterized by Dr. Ellen Rothenberg's laboratory [20]. They generated genome-wide ChIP-Seq data for several histone modifications and RNA-Seq data during the differentiation from "early T-cell precursors" (DN1) to CD4 and CD8 double-positive cells [20]. Using these data sets, we show below an example of a combinational usage of in-house programs, Access/Excel and MeV, to visualize the dynamics of H3K4me2 enrichment at promoters during the early stages of T-cell development. For a concise result, we limited the data analysis to genes that are specifically expressed in DN1 cells. Examples of Access/Excel files are available upon request:

1. Obtain gene expression data. (1) Download RNA-Seq raw sequence data from GEO (GSE31235; "sample1") [20]. Generate BED6 files as described in **steps 1–3** of Subheading 3.3. (2) Download UCSC genome annotation (mm9) as described in Subheading 2.1. (3) Apply the in-house C++ program "RPKMCalculator" to calculate RPKM values for each BED6 file. (4) Apply the in-house C++ program "Cat_expr_file" to concatenate the expression files from **step 3**. It outputs a file containing a matrix of expression values, with each row corresponding to one gene and each column corresponding to one sample.
2. Define DN1 specific genes. (1) Open the expression matrix file from Excel. (2) Create a new column and define the values for the column by using the formula (Fig. 3a).
3. Create a database using Access (*see* **Note 15**). Choose "Blank Database" from the templates that appear after running Access to create a blank database, name, and save it to a local drive.
4. Import the expression data into Access. Click the "External Data" panel and choose "Import/Excel File," which activates the "Import Text Wizard," to import the excel file generated from **step 2**. During the process, note that (1) the first row of the density file contains field names and (2) attribute "ID" should be specified as primary key of the table.
5. Obtain read density of H3K4me2 at promoters. (1) Download ChIP-Seq data from GEO (GSE31235) and process the sra

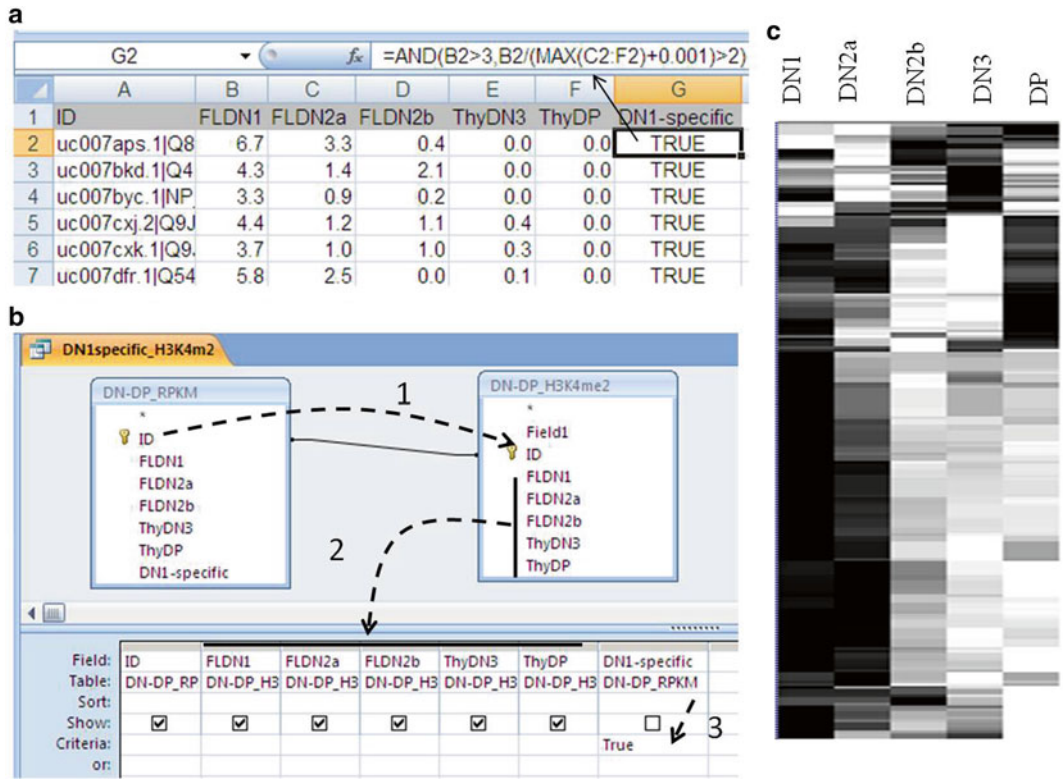


Fig. 3 A combinational use of Excel/Access and Mev to visualize changes of histone modifications. (**a**) A new column "DN1-specific" is created and the value is defined by the formula. As shown in the formula, a gene is defined as DN1-specific if the expression is higher than 3 ("B2 > 3," where "B2" means column "B" row "2") and is at least twofold higher than any other stages ("B2/(max(C2:F2) + 0.001) > 2," where 0.001 is a "pseudo count"). (**b**) Screenshot of the query window. The *dashed arrows* are explained in main text. (**c**) Hierarchical clustering and heatmap visualization of H3K4me2 density at promoters of DN1-specific genes during the differentiation from DN1 to DP cells. H3K4me2 is highly enriched at the promoters of most genes at DN1 stage and decreases during differentiation. However, about 1/3 of genes also show high enrichments of H3K4me2 at promoters at later development stages

files to BED6 files. (2) Apply the in-house C++ program "DensityCalculatorPromoters" to calculate the normalized read density at promoters for all samples (*see* **Note 16**).

6. Import the read density into Access. Similar to **step 4**, click "External data" panel, and chose "Import/Text" to import the text file generated from **step 5** into the database.
7. Intersect the expression and density tables. (1) Click the "create" panel and choose "query design" to activate the "show table" dialog. (2) Add the two tables to the query panel from the dialog. (3) Click the "ID" attribute of one table, hold, drag, and release it to the "ID" attribute of the other (Fig. 3b, dashed arrow 1) to create a join link that implements an intersection operation between the two tables through the specified attributes.

8. Extract read density for DN1 specific genes. (1) Click, hold, drag, and release the attributes associated with read densities to the bottom panel (Fig. 3b, dashed arrow 2). (2) Use "Criteria" in the bottom panel to restrict the query results to DN1-specific gene (Fig. 3b, dashed arrow 3). (3) Execute ("Design/Results/Run") and save the query ("ctrl+s"). (4) Export the results to a flat text file ("External Data/Export/Text File"). During this process, choose to include Field name and "Tab" as delimiter.
9. Visualize read density by using MeV (Fig. 3c). (1) Import the flat text file from **step 6** into MeV ("File/Load Data"). (2) Normalize the read density to highlight changes of H3K4me3 enrichment across samples ("Adjust Data/Gene/Row Adjustments/Normalize Genes/Rows"). (3) Cluster genes based on their read density across samples ("Analysis/Clustering/HCL"; *see* **Note 17**). A "HCL: Hierarchical clustering" dialog will show up before the clustering for users to set parameters. (4) Choose color theme ("Display/Color Scheme"), adjust color scale ("Display/Set Color Scale Limits"), and set size ("Display/Set Element Size") of the heatmap.

4 Notes

1. FastQC is implemented by script language JAVA and therefore has low run-time performance. In practice, one would extract the first 0.1 million reads to supply to FastQC: "head -n 400000 original_fastq_file>0.1_million_fastq_file", where the ">" symbol directs the output to the specified file.
2. Bowtie2 implements an option "--local" to allow reads to be trimmed at both extremes to optimize the alignment score. If the low-quality bases are known from the initial quality inspection, one could use the option "-5" ("-3") from Bowtie2 to specify the number of bases to be clipped at the 5′- (3′-) end.
3. Integrative Genomics Viewer (http://www.broadinstitute.org/igv/) is also frequently used by biologists to visualize ChIP-Seq/RNA-Seq data.
4. When running fastq-dump from SRA Toolkit, the whole path of the sra file is recommended to supply as input for beginners to minimize additional configurations. If several sra files are available for one sample, the fastq files can be concatenated with the "cat" commend: "cat fastq_file_1 fastq_file_2 … fastq_file_n > fastq_file_1_2".
5. The current version of "Sam2Bed6_Bowtie2" deals with single-end alignment.

6. The in-house C++ program "RemoveRedundantRead" removes redundant reads for every mapped position and outputs a sorted BED file.
7. Removing read redundancy is not recommend for RNA-Seq data, since coding regions constitute a small portion of the genome and it is very likely that two reads will hit the same position.
8. To enable comparison among different samples, the forth column of the BEDGraph file generated by "Generate RPBMBasedSummary" is normalized by library size (in millions) and by window size: *r*eads *p*er *b*ase per *m*illion reads (RPBM). To run the program, one needs to specify the number of bases to be shifted. Setting the shifting size as half of the length of a nucleosome DNA plus the linker DNA (approximately 200 bps) would work for most ChIP-Seq of histone modifications. It is recommended to set the shifting size to zero for RNA-Seq data set to ensure the shifted positions be within coding regions. One also needs to specify a window size. While a window size of 200 bps works for most ChIP-Seq, the window should be smaller for RNA-Seq data to account for exons less than 200 bps (e.g., 20 bps).
9. A combinational usage of "echo" and "cat" commands converts a BEDGraph output by "GenerateRPBMBasedSummary" (say file1) to a BEDGraph acceptable by the online UCSC genome browser (say file2): (1) echo track type = bedGraph name = \"track name\" description = \"description of the track\" > file2 and (2) cat file1 >> file2. The first command writes the parameters to a newly created file2, and the second command appends the content of file1 to file2.
10. It is highly recommended to compress the BEDGraph files before uploading to the online UCSC genome browser.
11. In our experiences, the UCSC genome browser only keeps the uploaded tracks for a couple of weeks on the same machine if the session is not saved. Saving the session also allow one to retrieve the tracks from different machines.
12. The last input parameter of the program "AverageDensity AcrossGenes" specifies the number of groups to separate based on the input order of gene annotation file. If the annotation is sorted by for example gene expression level, then by specifying the number of gene groups, the program can be used to correlate the read density with gene expression level.
13. Visualization based on the average density profile may be sensitive to outliers. For instance, if the read density is extremely high for several genes, then the profile would mostly reflect the features of this gene subset.

14. The "RPKMCalculator" program calculates the number of read mapped the annotated transcribed regions of an isoform and uses this number to calculate RPKM; it treats different isoforms from one gene independently.
15. Access manages data in the forms of tables: Each table contains several columns (or attributes), of which one may be marked as key to distinguish different records. It allows one to intersect different tables through the keys.
16. The "DensityCalculatorPromoters" program allows many input BED files, distinguished by different labels from the input. It outputs a matrix of read density for promoters (TSS ± 2 Kbps), with each row corresponding to one promoter and each column corresponding to one sample.
17. When setting parameters for hierarchical clustering in MeV, one needs to uncheck the "Sample Tree" option if the sample orders are known in prior. The "Normalize Genes/Rows" procedure in MeV subtracts the mean (row) and then divides the standard deviation (row) for each value to be normalized.

Acknowledgments

The authors are supported by the Intramural Research Program of the NIH, NHLBI.

References

1. Orkin SH, Zon LI (2008) Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132:631–644
2. Loose M, Swiers G, Patient R (2007) Transcriptional networks regulating hematopoietic cell fate decisions. Curr Opin Hematol 14:307–314
3. Cedar H, Bergman Y (2011) Epigenetics of haematopoietic cell development. Nat Rev Immunol 11:478–488
4. Wei G, Wei L, Zhu J et al (2009) Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. Immunity 30:155–167
5. Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
6. Cui K, Zang C, Roh TY et al (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell 4:80–93
7. Abraham BJ, Cui K, Tang Q et al (2013) Dynamic regulation of epigenomic landscapes during hematopoiesis. BMC Genomics 14:193
8. Wang Z, Zang C, Rosenfeld JA et al (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40:897–903
9. Meyer LR, Zweig AS, Hinrichs AS et al (2013) The UCSC genome browser database: extensions and updates 2013. Nucleic Acids Res 41:D64–D69
10. Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 41:D991–D995
11. Saeed AI, Sharov V, White J et al (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34:374–378
12. Bailey JA, Gu Z, Clark RA et al (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

13. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10:71–73
14. Hu G, Schones DE, Cui K et al (2011) Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. Genome Res 21:1650–1658
15. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
16. Kidder BL, Hu G, Zhao K (2011) ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol 12:918–922
17. Ramskold D, Luo S, Wang YC et al (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30:777–782
18. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628
19. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140
20. Zhang JA, Mortazavi A, Williams BA et al (2012) Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. Cell 149: 467–482

Part IV

Embryo Culture and Derivation of Stem Cells

Chapter 12

In Vitro Maturation and In Vitro Fertilization of Mouse Oocytes and Preimplantation Embryo Culture

Benjamin L. Kidder

Abstract

Epigenetic regulation of gene expression in the germline is important for reproductive success of mammals. Misregulation of genes whose expression is correlated with reproductive success may result in subfertility or infertility. To study epigenetic events that occur during oocyte maturation and preimplantation embryo development, it is important to generate large numbers of ovarian follicles and embryos. Oocyte maturation can be modeled using in vitro maturation (IVM), which is a system of maturing ovarian follicles in a culture dish. In addition, fertilization and early embryogenesis can be modeled using in vitro fertilization (IVF), which involves the fertilization of mature oocytes with capacitated sperm in a culture dish. Here, we describe protocols for in vitro maturation (IVM) and in vitro fertilization (IVF) of mouse oocytes and preimplantation embryo culture. These protocols are suitable for the study of oocyte and embryo biology and the production of embryonic mice.

Key words In vitro maturation, In vitro fertilization, IVM, IVF, Oocytes, Ovarian follicle, Cumulus oocyte complex, COC, Blastocyst, Embryo culture, Mouse

1 Introduction

The survival of mammalian species requires fertilization of an oocyte by a sperm to form the zygote, which is the first developmental landmark in the embryo [1] and is a key determinant of reproductive success. Development of the germline lineage in vertebrate species, which begins around mid-gestation in mice and continues through sexual maturity in adulthood, sets the stage for successful reproduction. Epigenetic regulation of gene expression in the germline, including oocytes in females and sperm in males and supporting lineages such as cumulus, granulose, and theca cells in females and Sertoli and Leydig cells in males, is essential for normal development and reproductive success. Perturbation of regulatory pathways or misregulation or genes whose expression is correlated with reproductive success may result in subfertility or infertility.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_12, © Springer Science+Business Media New York 2014

Experimental models of oocyte maturation [2–4] and fertilization [3, 5] and embryo development [5] have furthered our understanding of reproduction biology. Advancements in artificial reproductive therapies (ART) rely on the development of novel protocols for studying oocyte and embryo biology. To study epigenetic events that occur during oocyte maturation and preimplantation embryo development, including pluripotent cells of the inner cell mass (ICM) and multipotent trophoblast stem (TS) cells of the trophectoderm [6, 7], it is important to generate large numbers of ovarian follicles and embryos, without using a significant number of mice. Development of oocytes can be modeled using in vitro maturation (IVM), which is a system of maturing ovarian follicles in a culture dish. In addition, fertilization and early embryogenesis can be modeled using in vitro fertilization (IVF), which involves the fertilization of mature oocytes with capacitated sperm in a culture dish. There are several advantages with the use of IVM and IVF relative to in vivo fertilization for the study of oocyte and preimplantation embryo biology. Because penetration of sperm is synchronous during IVF, development of fertilized oocytes and embryos should also be synchronous, which is difficult to achieve in vivo. Also, propagation of subfertile species may be enhanced using IVM and IVF techniques. In addition, real-time visualization of oocytes and preimplantation embryos and manipulation of culture conditions can be achieved in a highly controlled environment, which is not as feasible in vivo. Here, we describe protocols for IVM and IVF of mouse oocytes and preimplantation embryo culture. These protocols are suitable for the study of oocyte and embryo biology and the production of embryonic mice.

2 Materials

2.1 Oocyte Retrieval, In Vitro Maturation and In Vitro Fertilization of Oocytes, and Embryo Culture

1. Minimum Essential Medium (MEM) Alpha Medium (Gibco cat # 12571).
2. Fetal bovine serum (FBS).
3. EmbryoMax M2 Medium (1×, Millipore cat # MR-015-D).
4. EmbryoMax Human Tubal Fluid (HTF) (1×, Millipore, cat # MR-070-D).
5. EmbryoMax KSOM Medium (1×, Millipore cat # MR-106-D).
6. Syringe (1 cc) and 26 G needle.
7. Dissecting instruments (stainless steel fine forceps and fine scissors).
8. Stereomicroscope.

9. Sterile plastic Petri dishes (35, 60, and 100 mm).
10. Multidish 4-well culture cell (Nunc cat #176740).
11. Phosphate-buffered saline without calcium or magnesium (PBS, 1×).
12. Mineral oil (pre-gassed) for mouse embryo culture (Sigma).
13. Aspirator tube assembly: aspirator mouth piece, tubing (Sigma A5177).
14. Pulled capillary tube (Drummond microcaps 1-000-0500) or drawn Pasteur pipette.
15. Aspirator tube assembly: aspirator mouth piece, tubing (Sigma A5177).
16. Pipettes.
17. Cell culture incubator with carbon dioxide and nitrogen gas tanks (37 °C and 5 % CO_2 and 5 % O_2).
18. 70 % ethanol for sterilization.
19. C57Bl6 or B6D2F1 female mice (19–23 day old, *see* **Note 1**) and male mice (>6 weeks old). Fertilization rates of sperm and oocytes from F1 hybrid strains such as C57Bl6 or B6D2F1 should be >90 %. Other strains may be used. However, fertilization rates will vary depending on the strain.
20. Penicillin–streptomycin (100×).
21. Lab Armor bead bath (*see* **Note 2**).

2.2 Hormones

1. Pregnant mare serum gonadotropin (PMSG, Calbiochem, Cat. No. 367222). Resuspend lyophilized PMSG in PBS to a final concentration of 50 IU/mL. Aliquot 1 mL into 1.5 mL Eppendorf tubes (50 IU) and freeze at −20 or −80 °C until use. Each female mouse receives a 100 μL injection which is equivalent to 5 IU of PMSG.
2. Human chorionic gonadotropin (hCG, Sigma Cat. # CG-5). Stock solution: Resuspend lyophilized hCG in PBS to a final concentration of 1,000 IU/mL. Aliquot 50 μL into 1.5 mL Eppendorf tubes (50 IU) and freeze at −20 or −80 °C until use. Working solution: Add 950 μL PBS to the tube with 50 μL of hCG to make a final concentration of 50 IU/mL. Each female mouse receives a 100 μL injection equivalent to 5 IU of hCG.
3. Follicle-stimulating hormone (FSH). Maturation of oocytes can be stimulated by addition of exogenous growth factors such as FSH [8] and IL-6 [9], where FSH and IL-6 promote the expansion of cumulus cells in the cumulus oocyte complex (COC).

2.3 In Vitro Maturation, In Vitro Fertilization, and Embryo Culture Media

1. In vitro maturation (IVM) media for the maturation of mouse oocytes harvested from superovulated C57Bl6 or B6D2F1 female mice. Fresh culture dishes should be prepared the day before IVM will be performed. Alpha-MEM, 10 % FBS, 0.2 IU/mL FSH.
2. In vitro fertilization (IVV) media: Fresh culture dishes should be prepared the day before IVF will be performed. Sperm dishes should be prepared as follows: Multidish 4-well culture cells containing 400 μL HTF should be covered with 200 μL mineral oil and incubated at 37 °C with 5 % CO_2. IVF culture dishes should be prepared using 30 mm dishes as follows: Fertilization dishes should contain two 40 μL drop of HTF media for IVF and one 120 μL wash drop of HTF media. Cover the 30 mm culture dish with mineral oil and incubate at 37 °C with 5 % CO_2.
3. Embryo culture media: One 30 mm culture dish should be prepared for every fertilization dish. Add two 40 μL drops of KSOM and one 120 μL wash drop of KSOM to the 30 mm dish. Cover the dish with mineral oil and incubate 37 °C with 5 % CO_2, 5 % O_2, and 90 % N_2.

3 Methods

3.1 Superovulation of Female Mice for IVM

1. Day 1. Female mice (C57Bl6 or B6D2F1) aged 19–23 days should be subjected to a hormonal regiment to induce superovulation. For IVM, female mice should be intraperitoneally injected with 5 IU of PMSG (100 μL of PMSG stock) between 5:00 and 7:00 pm. The timing of the PMSG injection should also depend on the time the oocytes are collected. It is important to include a positive control for maturation of oocytes in vivo. For this purpose, female mice should be intraperitoneally injected 48 h later with 5 IU of hCG (100 μL of hCG working solution).

3.2 Preparation of Culture Dishes for IVM

1. Day 2. Prepare fresh dishes for each condition the day before COCs will be collected. Culture dishes (35 mm) should be prepared containing two 50 μL drops of IVM media (alpha-MEM, 10 % FBS, 0.2 IU/mL FSH) and one 120 μL drop of IVM media without hormone for washing. This dish serves as a positive control. Carefully cover the drops of IVM media by adding 3–4 mL of mineral oil. Incubate the IVM dishes at 37 °C with 5 % CO_2. To test the role of exogenous hormones or agonists in the expansion of COCs, multiple dishes should be prepared for each condition, and hormonal treatments should be empirically optimized. For example, FSH or IL-6 should be titrated to identify the optimal hormonal dose that induces oocyte maturation (e.g. 3 μg/mL, 30 μg/mL, or 300 μg/mL of IL-6).

3.3 In Vitro Maturation (IVM) of Mouse Oocytes

1. Day 3. In the morning (e.g., 9:00 am), warm several 15 mL conical tubes of M2 media on a heat plate at 37 °C.
2. Day 3 afternoon (2:00 and 3:00 pm). Dissect ovaries in M2 media in a 6 cm culture dish. During the dissection procedure it is important to remove the oviduct and clean residual fat from the ovaries. Failure to remove the oviduct and fat will result in excessive cellular debris and droplets of fat, respectively, which may make it more difficult to identify COCs in the downstream steps.
3. Place the cleaned ovaries in a new 6 cm culture dish. To remove the COCs, use two 26 G needles attached to syringes to puncture and disrupt the ovaries.
4. Use a mouth pipette apparatus to transfer the COCs from the 6 cm culture dish to a 35 mm dish containing IVM media. First, transfer the COCs to the wash drop containing a 120 μL drop of IVM media, and then transfer to a 40 μL drop of IVM media. Culture the dishes overnight at 37 °C with 5 % CO_2 and normal (atmospheric) O_2 levels.
5. Day 3 afternoon (5:00 and 7:00 pm). As a control for the IVM, female mice should be intraperitoneally injected with 5 IU of hCG (100 μL of hCG working solution) 48 h after injection of PMSG (5:00 and 7:00 pm).
6. Day 4. Evaluate the expansion of COCs using a dissecting stereomicroscope or an inverted bright field microscope. Control COCs should not undergo expansion (Fig. 1a), while treatment with FSH (0.2 IU/mL; Fig. 1b), IL-6 (3 μg/mL; Fig. 1c), or FSH/IL-6 (0.2 IU/mL, 3 μg/mL; Fig. 1d) should induce the expansion of COCs.
7. Day 5. Compare in vivo ovulated oocytes with IVM oocytes. Prepare in vivo ovulated oocytes by dissecting oviducts from PMSG/hCG-injected female mice in M2 media warmed to 37 °C. In a 6 cm culture dish, dissect the oviduct and transfer to a fresh 35 mm culture dish containing warm M2 media. Use forceps or a 26 G needle and syringe to tear the ampulla of the oviduct and massage out the COCs. Transfer the COCs to a 120 μL drop of IVM media to wash, and then to a 40 μL drop of IVM media. In vivo matured oocytes can be counted and compared to IVM oocytes using microscopy, gene expression, or other experimental techniques.

3.4 Superovulation of Female Mice for IVF

1. Day 1. Female mice (C57Bl6 or B6D2F1) aged 19–23 days should be subjected to a hormonal regiment to induce superovulation. For IVF, female mice should first intraperitoneally injected with 5 IU of PMSG (100 μL of PMSG stock) between 5:00 and 7:00 pm.

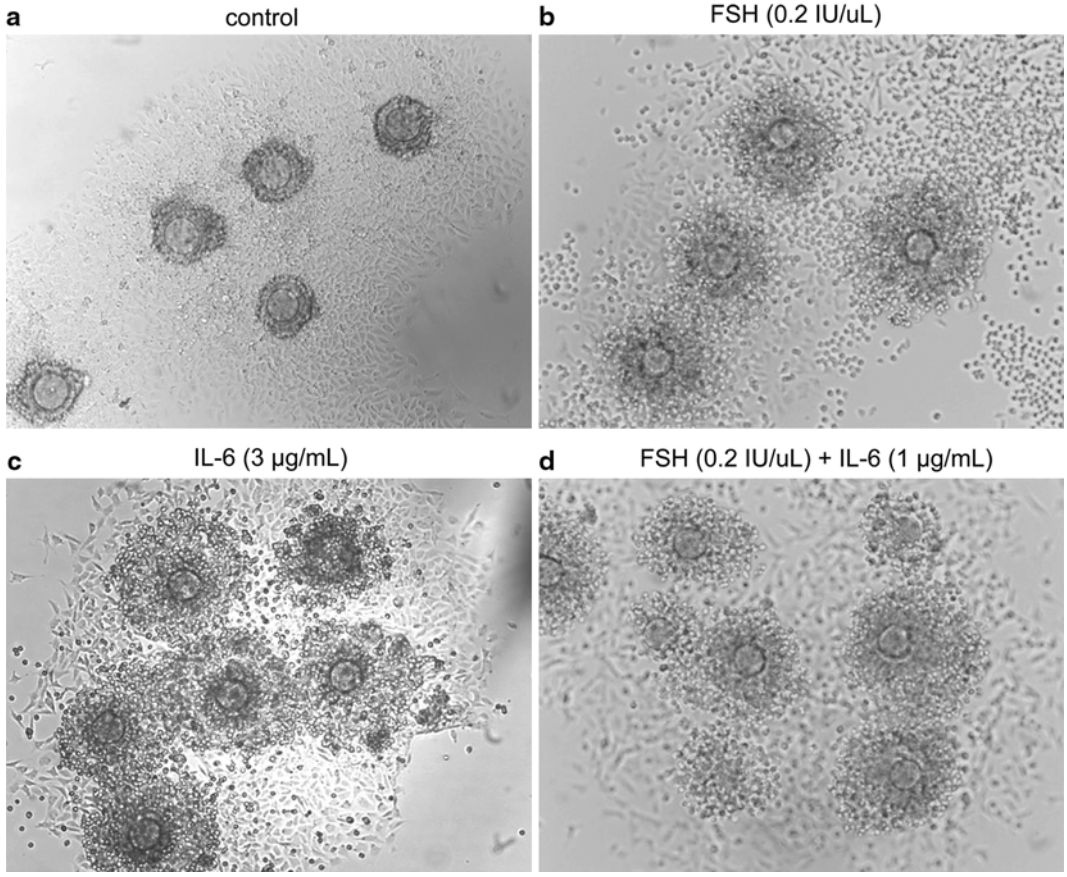


Fig. 1 IVM with FSH or IL-6 hormonal treatment induces COC expansion. Non-expanded COCs were collected from PMSG-primed female mice and cultured in defined IVM media containing HTF and exogenous hormones (FSH, IL-6). While (**a**) control COCs failed to undergo expansion, treatment with (**b**) FSH (0.2 IU/mL), (**c**) IL-6 (3 μg/mL), or (**d**) FSH/IL-6 (0.2 IU/mL, 1 μg/mL) induced the expansion of COCs

2. Day 3. The same female mice should be intraperitoneally injected with 5 IU of hCG (100 μL of hCG working solution) 48 h after injection of PMSG (5:00–7:00 pm).

3.5 Preparation of Culture Dishes for IVF

1. Day 3. Prepare fresh dishes the day before IVF will be performed.
2. Fresh sperm dish: Prepare a multidish 4-well culture cell by adding 400 μL of HTF media and cover with 200 μL of mineral oil. Incubate at 37 °C with 5 % CO_2 5 % O_2 and 90 % N_2 levels. It is important to use pre-gassed mineral oil.
3. Culture dish for IVF (IVF-HTF): Prepare one 35 mm culture dish for each fertilization procedure by adding two 50 μL drops of HTF media and one 120 μL wash drop of HTF (*see* **Note 3**).

4. Embryo culture dish (IVC-KSOM): Prepare several 35 mm culture dishes containing two 50 μL drops of KSOM media and one 120 μL wash drop of KSOM media.
5. Warm five 15 mL conical tubes of M2 overnight on a heat plate at 37 °C.

3.6 In Vitro Fertilization (IVF) of Mouse Oocytes

1. Day 4 (7:30 am). Euthanize a male mouse and dissect out the cauda epididymis in warm M2 media. Transfer the cauda to the 4-well sperm dish containing 400 μL of HTF media and covered with 200 μL mineral oil. Use a 16 gauge needle and a syringe to tear open the cauda and squeeze the sperm out of the vas deferens. Then, use forceps to move down the vas deferens to remove any remaining sperm. Opaque dark clouds of sperm should be visible. Incubate for 5 min at 37 °C.
2. Remove residual tissue from 4-well dish and incubate for 1–1.5 h at 37 °C with 5 % CO_2, 5 % O_2, and 90 % N_2. This step allows the sperm to swim out of the tissue and undergo capacitation, or maturation, which is a biochemical event which involves the destabilization of the acrosomal sperm head membrane allowing increased binding between the sperm and the oocyte. Capacitation is required for mammalian sperm to become competent to fertilize an oocyte [10].
3. Meanwhile, prepare in vivo matured and ovulated oocytes (in vivo COCs) from PMSG/hCG-injected females. Dissect the oviducts from PMSG/hCG-injected female mice in M2 media warmed to 37 °C. In a 6 cm culture dish, dissect the oviduct and transfer to a fresh 35 mm culture dish containing warm M2 media. Use forceps or a 26 gauge needle and syringe to tear the ampulla of the oviduct and massage out the COCs. Transfer the COCs to a 120 μL wash drop of HTF (IVF media), and then transfer to a 50 μL drop of HTF media.
4. Day 4 (9:00 am). Transfer in vitro matured oocytes (IVM COCs) to an HTF wash drop submerged under oil in a 10 μL volume using a 200 μL pipette tip. Next, transfer the COCs to a 50 μL drop of HTF.
5. Remove the sperm from the incubator and evaluate the opacity. The sperm should resemble a homogeneous cloud. Dilute the sperm 1:2 by adding 400 μL HTF media to the 400 μL HTF media in the 4-well dish. Add 10 μL sperm to the COCs in the drop of HTF media, and incubate at 37 °C with 5 % CO_2 for 4–6 h.
6. Day 4 (2:00–3:00 pm): After 4 h of incubation, the oocytes should be fertilized. The COCs should be washed to remove the excess sperm. Forcefully pipette the COCs up and down several times with a 10 μL volume using a 200 μL pipette.

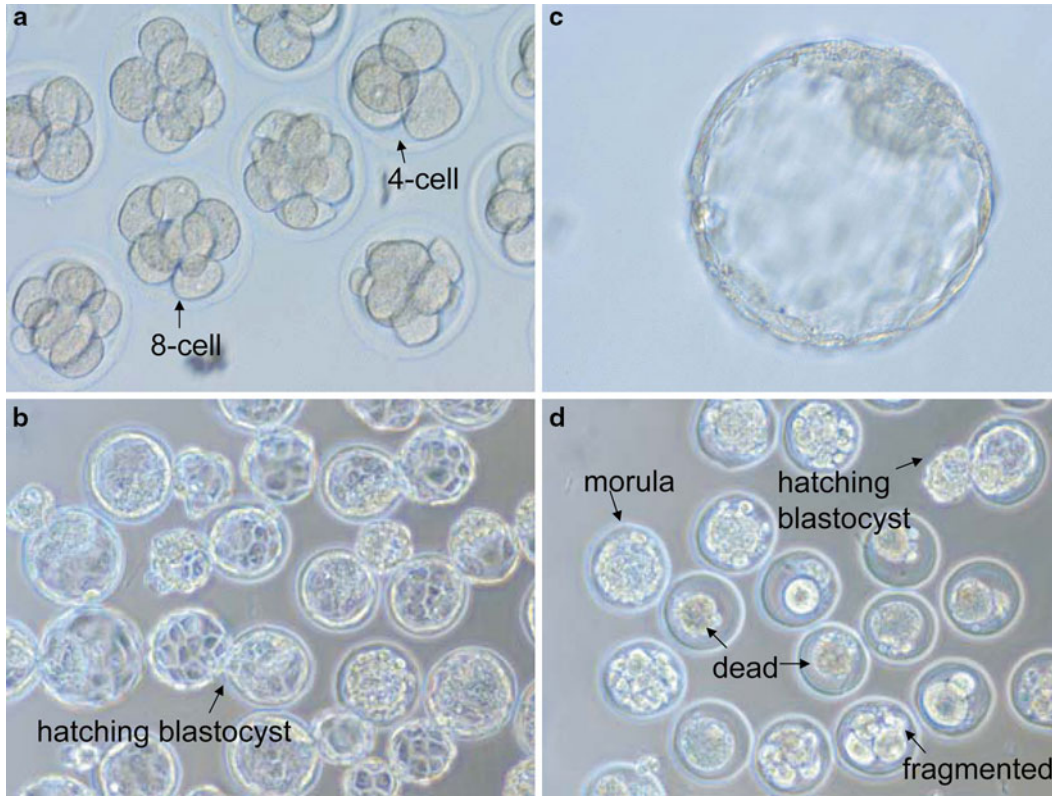


Fig. 2 In vitro embryo culture following IVM and IVF treatment. Bright field microscopy of (**a**) four- and 8-cell stage embryos, (**b**) E4.5 blastocysts undergoing hatching, (**c**) a hatched blastocyst, and (**d**) unhealthy embryos underdeveloped (morula stage), dead embryos with fragmented blastomeres, or embryos arrested at the two-cell stage

7. Transfer fertilized oocytes to a fresh drop of HTF media within the same 35 mm culture dish. Take care to leave behind as much cellular debris as possible.
8. Transfer viable fertilized oocytes to a new 35 mm culture dish containing KSOM media that was prepared previously. Distribute the embryos evenly throughout the culture dish and incubate at 37 °C with 5 % CO_2 overnight. Use a microscope to observe the two pronuclei using a microscope.
9. Day 5: Observe the two-cell cleavage to blastocyst stage (E4.5) using an inverted bright field microscope. If IVM and IVF have been performed successfully, embryos should develop to the four- and eight-cell (Fig. 2a) and blastocyst stages (Fig. 2b, c). Viable blastocysts should undergo hatching following development to the blastocyst stage (Fig. 2b). However, if embryos are to be transferred to pseudopregnant females, they should be transferred prior to the hatching stage. Embryos that exhibit excessive blastomere fragmentation or a

loss of their shiny cellular morphology (Fig. 2d) during in vitro culture represent unhealthy or dead embryos.

10. Transfer embryos to pseudopregnant females or continue to culture embryos in KSOM media.

4 Notes

1. For IVM it is recommended to use female mice aged 19–23 days. However, for IVF, female mice aged 4–6 weeks can be used.
2. Use of a Lab Armor bead bath prevents microbial growth compared to a water bath.
3. It is important to label the bottom of the culture dish.

References

1. Matzuk MM, Lamb DJ (2008) The biology of infertility: research advances and clinical challenges. Nat Med 14(11):1197–1213. doi:10.1038/nm.f.1895, nm.f.1895 [pii]
2. Eppig JJ, O'Brien MJ (1996) Development in vitro of mouse oocytes from primordial follicles. Biol Reprod 54(1):197–207
3. Eppig JJ, Schroeder AC (1989) Capacity of mouse oocytes from preantral follicles to undergo embryogenesis and development to live young after growth, maturation, and fertilization in vitro. Biol Reprod 41(2):268–276
4. O'Brien MJ, Pendola JK, Eppig JJ (2003) A revised protocol for in vitro development of mouse oocytes from primordial follicles dramatically improves their developmental competence. Biol Reprod 68(5):1682–1686. doi:10.1095/biolreprod.102.013029 biolreprod.102.013029 [pii]
5. Edwards RG (1965) Maturation in vitro of mouse, sheep, cow, pig, rhesus monkey and human ovarian oocytes. Nature 208(5008):349–351
6. Tanaka S, Kunath T, Hadjantonakis AK, Nagy A, Rossant J (1998) Promotion of trophoblast stem cell proliferation by FGF4. Science 282(5396):2072–2075
7. Kidder BL, Palmer S (2010) Examination of transcriptional networks reveals an important role for TCFAP2C, SMARCA4, and EOMES in trophoblast stem cell maintenance. Genome Res 20(4):458–472, gr.101469.109 [pii] 10.1101/gr.101469.109
8. Cortvrindt R, Smitz J, Van Steirteghem AC (1996) In-vitro maturation, fertilization and embryo development of immature oocytes from early preantral follicles from prepuberal mice in a simplified culture system. Hum Reprod 11(12):2656–2666
9. Liu Z, de Matos DG, Fan HY, Shimada M, Palmer S, Richards JS (2009) Interleukin-6: an autocrine regulator of the mouse cumulus cell-oocyte complex expansion process. Endocrinology 150(7):3360–3368. doi:10.1210/en.2008-1532, en.2008-1532 [pii]
10. Visconti PE, Galantino-Homer H, Moore GD, Bailey JL, Ning X, Fornes M, Kopf GS (1998) The molecular basis of sperm capacitation. J Androl 19(2):242–248

Chapter 13

Derivation and Manipulation of Trophoblast Stem Cells from Mouse Blastocysts

Benjamin L. Kidder

Abstract

The trophoblast is the first lineage to undergo differentiation during mammalian development. In the preimplantation blastocyst embryo, two cell types are present including the inner cell mass (ICM) and the trophectoderm (TE). ICM cells exhibit pluripotent potential, or the capacity to give rise to all cells represented in the adult organism, while TE cells are multipotent and are therefore only capable of differentiating into trophoblast lineages represented in the placenta. The TE is essential for implantation of the embryo into the uterine tissue, formation of trophoblast lineages represented in the placenta, and exchange of nutrients and waste between the embryo and the mother. Trophoblast stem (TS) cells, which can be derived from the TE of preimplantation embryos in the presence of external signals such as FGF4, can self-renew indefinitely, and because they are capable of differentiating into epithelial lineages of the trophoblast, TS cells are a useful in vitro model to study the biology of the trophoblast including epigenetic regulation of gene expression. In this chapter we describe protocols for derivation of TS cells from mouse blastocysts, culture conditions that promote self-renewal and differentiation, and methods to transduce TS cells with lentiviral particles encoding shRNAs. These protocols are sufficient for efficient derivation of TS cells and robust RNAi knockdown of target genes in TS cells.

Key words Trophoblast stem cells, Trophectoderm, FGF4, Derivation, Blastocyst, Differentiation, RNA interference, shRNA, Lentivirus

1 Introduction

Blastocyst formation in mammals yields the generation of two distinct cellular layers including the inner cell mass (ICM) and the trophectoderm (TE). Three days after fertilization of the oocyte in mice (4–5 in humans), the TE forms a spherical layer of epithelial cells around the ICM of the blastocyst. In contrast to the ICM, which generates all somatic and germline lineages, the TE gives rise to epithelial trophoblast lineages of the placenta including trophoblast giant cells, spongiotrophoblasts, glycogen trophoblast cells, and syncytiotrophoblasts [1]. The ICM and TE both contain stem cell compartments that are capable of self-renewing in the presence of appropriate external signals. The ICM and TE can be

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_13, © Springer Science+Business Media New York 2014

modeled in vitro through the derivation of embryonic stem (ES) cells [2] and trophoblast stem (TS) cells [3], respectively. ES cells can be maintained in vitro indefinitely in the presence of LIF, BMP4, and WNT3a, while TS cells can be propagated indefinitely in the presence of FGF4, INHBA, NODAL, and TGFB1 [3]. TS cells can be differentiated by removal of the self-renewal factors, which is accompanied by decreased proliferation and trophoblast differentiation. In contrast, ES cells can be differentiated into all cellular phenotypes of the adult organism that are programmed into the mammalian genome.

TS cells are a useful model to study the development, gene expression, and the epigenomic landscape of placental lineages. Previous findings have demonstrated that TS cells and ES cells share common and unique expression programs. ES cell self-renewal and pluripotency require the core transcription factors Oct4 [4], Sox2 [5], and Nanog [6, 7], while TS cell multipotency requires transcription factors such as Cdx2 [8], Eomes [9], Esrrb [10], and Tead4 [11]. Recently, genome-wide transcriptome analysis and epigenomic analyses (e.g., ChIP) have provided global views of molecular networks that regulate transcription in TS cells [12, 13]. Previously, we found that ES cells and TS cells co-express a number of transcriptional regulators including Sox2, Stat3, Tbx3, Esrrb, Klf5, Lin28, Rest, Rex1, Sall4, Smarca4, and Utf1 [12]. Using an RNAi knockdown approach, we also found that Tcfap2c (Ap-2g), Smarca4 (Brg1), and Eomes are important for TS cell self-renewal [12]. These findings provide additional insight into epigenetic mechanisms of TS cell self-renewal. Here, we describe protocols for deriving, maintaining, and differentiating TS cells from mouse blastocysts and protocols to transduce TS cells with lentiviral particles encoding shRNAs.

2 Materials

2.1 Cell Culture Reagents

1. Dulbecco's Modified Eagle's Medium (DMEM), high glucose.
2. Penicillin–streptomycin (100X) (Invitrogen).
3. L-Glutamine (200 mM).
4. Fetal bovine serum (FBS, ES cell qualified).
5. Nonessential amino acids (NEAA, 100×) (Invitrogen).
6. Phosphate-buffered saline without calcium or magnesium (PBS, 1×).
7. 0.25 % Trypsin-EDTA (Invitrogen).
8. 0.05 % Trypsin-EDTA (Invitrogen).
9. Roswell Park Memorial Institute (RPMI) 1640 medium.
10. Sodium pyruvate (100 mM) (Invitrogen).
11. 2-Mercaptoethanol (100×, cell culture grade, Invitrogen).

12. FGF4 (25 μg, R&D Systems). Resuspend lyophilized FGF4 in PBS with 0.1 % BSA. Aliquot tubes and freeze at −80 °C until use (*see* **Note 1**). To make a solution of 0.1 % BSA in PBS, dissolve BSA in PBS, filter using a 0.45 μM syringe, aliquot, and store at −80 °C until use.
13. Heparin (10,000 units, Sigma H3149). Prepare a 1000× stock tube of heparin by resuspending in PBS to 1 mg/mL (1,000×). Store at −80 °C until use.
14. 24-, 12-, and 6-well culture dishes.
15. 0.1 % gelatin solution in water.
16. M2 medium (EmbryoMax 1×, Millipore).
17. Polybrene.
18. Puromycin.
19. All-*trans*-retinoic acid (Sigma R2625).
20. 5, 15, and 50 mL polystyrene conical tubes.
21. Lab Armor bead bath.
22. 70 % ethanol.

2.2 Equipment

1. Pulled capillary tube or drawn Pasteur pipette.
2. Aspirator tube assembly: aspirator mouth piece and tubing (Sigma A5177).
3. Flushing needle (1 cc syringe and 30–32 G needle).
4. Dissecting instruments (fine forceps, fine scissors).
5. Sterile Petri dishes (10 cm, 35 mm).
6. Stereomicroscope.
7. Clinical centrifuge.
8. Cell culture incubator.

2.3 Lentivirus Production

1. HEK 293 T cells.
2. Opti-MEM medium (Invitrogen).
3. Lipofectamine 2000 (Invitrogen).
4. Lentiviral plasmids: envelope plasmid (pLP/VSVG), packaging vector (e.g. psPAX2), and shRNA lentiviral expression vector (e.g., pGreenPuro, System Biosciences).
5. 10 cm culture dish.
6. 0.45 μM cell strainer (BD Biosciences).

2.4 Culture Media

1. Mouse embryonic fibroblast (MEF) media for culture of mitotically inactivated MEFs (iMEFs) derived from E13.5-E14.5 mouse embryos. Store in liquid nitrogen until use. DMEM high glucose, 10 % FBS, penicillin–streptomycin (1×), and L-glutamine (1×) at 37 °C with 5 % CO_2. Mitotically inactivated

MEFs (iMEFs) can be harvested from mouse embryos (E13.5–14.5) [14] or purchased from a commercial vendor.

2. TS cell derivation media: RPMI 1640, 20 % FBS, penicillin–streptomycin (1×), L-glutamine (1×), sodium pyruvate (1 mM), 2-mercaptoethanol (1×), 25 ng/mL FGF4, and 1 μg/mL heparin [3]. Store at 4 °C until use.
3. Feeder-conditioned TS cell media: Conditioned media (CM) is prepared by culturing iMEFs in TS cell media without FGF4 and heparin and harvesting the conditioned media 2–3 days later. CM can be stored at −20 °C until use or at 4 °C for a short period. Avoid multiple freeze–thaw cycles.

3 Methods

3.1 Preparation of iMEF Feeder Layer

1. For culture of iMEFs on 24-, 12-, and 6-well gelatin-coated plates, add 2 mL of gelatin solution to culture plates, and incubate at 37 °C for 10 min.
2. Pre-warm MEF media at 37 °C in a Lab Armor bead bath or water bath.
3. Thaw a vial of frozen MEFs in a Lab Armor bead bath or water bath, add cells to 3 mL of pre-warmed MEF media in a 15 mL conical tube, and spin at 1,500 rpm (500 × *g*) for 3–5 min.
4. Meanwhile, remove gelatin from culture plates. After spinning has completed, resuspend iMEFs in MEF medium, and plate the iMEFs in the culture dishes. The total volume of MEF media for a 24-well, 12-well, and 6-well plates should be 0.5 mL, 1 mL, and 2 mL, respectively. Before placing the plate in the incubator, move the plate vertically then horizontally to evenly distribute MEFs.

3.2 Isolation of Blastocysts

1. Day 0. Set up mouse matings. Place a female mouse in estrus into a cage housing a male in the afternoon. Check for the presence of a vaginal plug the following morning (E0.5). Place the female in a new cage.
2. Day 3. Collection of blastocysts. Aspirate the MEF media from the 24-well culture plates containing iMEFs, and add TS cell media containing 25 ng/mL FGF4 and heparin. Place the plate back in the incubator to equilibrate the media. Meanwhile, prepare reagents for harvesting blastocysts from the female mouse. Pre-warm M2 medium at 37 °C:
   (a) Sterilize dissection area by spraying with 70 % ethanol (*see* **Note 2**).
   (b) Sacrifice female mouse by cervical dislocation or CO_2 asphyxiation (*see* **Note 3**).

(c) Spray abdomen with 70 % ethanol to sterilize the fur, then use fingers to open abdomen, or use scissors to cut through skin and peritoneum.

(d) Dissect uterus by cutting below the oviduct and above the cervix.

(e) Wash uterus briefly in PBS in a 10 cm Petri dish.

(f) Place the uterus in a 35 mm dish containing M2 medium.

(g) Use a 1 cc syringe with a 32 G needle to flush the E3.5 blastocysts from the oviduct side of the uterus horn through the cervix. Use forceps to hold the uterus near the oviduct side while flushing the uterus. Flush each horn twice with M2 medium. When finished flushing, place uterus in another dish.

(h) Use mouth pipette setup and stereomicroscope to transfer blastocysts from 35 mm dish to 24-well plate containing iMEFs and TS cell media. Place one blastocyst per well and incubate at 37 °C with 5 % CO_2.

3.3 Derivation of TS Cells

1. The following day, use an inverted microscope to check the attachment of the blastocysts to the layer of feeder cells. Blastocysts should attach to the iMEFs within 2–3 days.
2. For wells that contain attached blastocysts after 48 h, aspirate the medium and add 500 μL of fresh TS cell media with FGF4 and heparin. For wells that contain unattached blastocysts, add fresh medium without aspirating the old medium.
3. Continue to culture blastocysts until the outgrowth has reached a size that is adequate for disaggregation (Fig. 1). Culture of blastocysts in TS cell media should promote the expansion of the trophoblast layer, while proliferation of the ICM clump should be minimal.
4. Disaggregate the blastocyst outgrowth before it becomes too large in size (*see* **Note 4**). If the outgrowth is cultured too long, it may result in differentiation of the TS cells into giant cells or expansion of endoderm progenitors. Wash the well with 500 μL PBS, and add 100 μL pre-warmed 0.05 trypsin-EDTA. Incubate for 2–3 min at room temperature. Briefly pipette and add to a 15 mL conical tube containing 3 mL of pre-warmed MEF media. Centrifuge for 3 min at 1,500 rpm (500 × *g*).
5. Resuspend the pellet in 500 μL TS cell media containing FGF4 and heparin, and plate on a fresh well of a 24-well plate containing iMEFs. Continue to incubate at 37 °C with 5 % CO_2.
6. TS cell colonies should become visible within 7–10 days after disaggregation and replating. Note the compact colony morphology of early passage TS cells resembling epithelial sheets (Fig. 2). Continue to change the TS cell medium every 2 days.

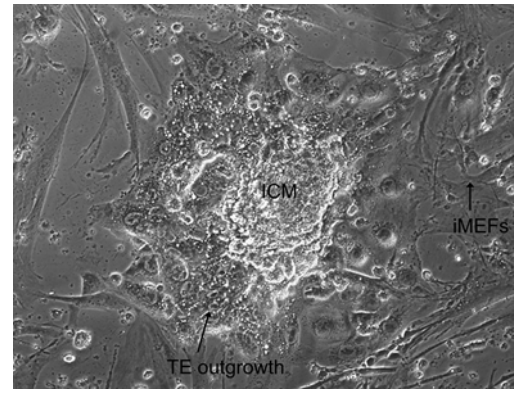


Fig. 1 Blastocyst outgrowth. Bright-field microscopy of a blastocyst outgrowth on a feeder layer. Note the presence of the inner cell mass (ICM) and trophectoderm (TE) outgrowth on a layer of mitotically inactivated MEFs (iMEFs). Trophoblast stem (TS) cells can be derived from multipotent stem cells of the TE. The outgrowth is an appropriate size for disaggregation and replating on a culture dish of equal size containing iMEFs

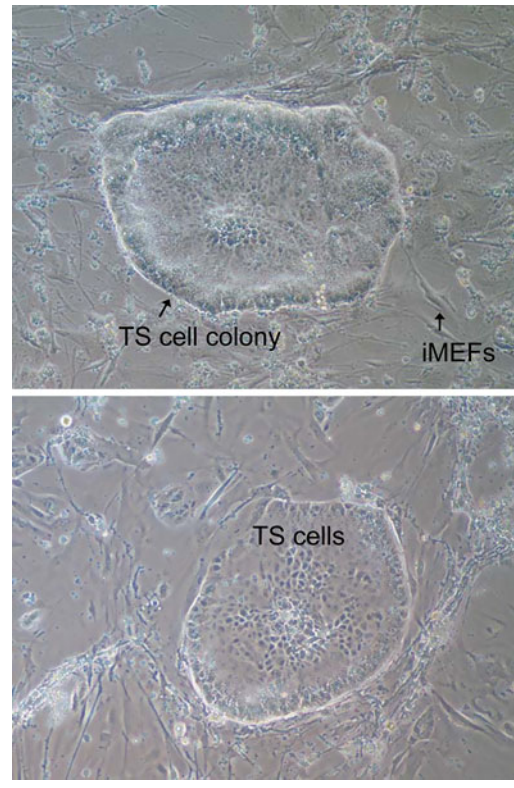


Fig. 2 TS cells growth on iMEFs. Trophoblast stem (TS) cell colonies growing on a layer of iMEFs. Note the presence of a tight cell boundary and a colony morphology resembling epithelial sheets which are typical characteristics of TS cells

7. Once the colonies become 50 % confluent, passage the TS cells onto a 12-well dish. Wash the 24-well with 500 μL PBS, then add 100 μL of pre-warmed 0.05 % trypsin, and incubate for 2–3 at room temperature. Pipette up and down several times to dissociate the TS cells but attempt to leave small clumps of cells which maintain greater self-renewal relative to individual cells.
8. When the colonies become 50 % confluent, passage the TS cells onto a 6-well dish. Repeat the steps for passaging TS cells as described in **step 7**. It is imperative to maintain small clumps of TS cells while passaging during the early stages of derivation to avoid differentiation of TS cells.
9. It is best to passage established TS cell colonies once they reach 70 % confluency on 6-well plates or 10 cm culture dishes at a high split ratio (1:10). A high split ratio seems to maintain the integrity of TS cells relative to a low split ratio. However, because each TS cell line is unique, splitting ratios must be empirically optimized. TS cell lines that have a lower proliferative rate can be passaged 1:5 or 1:6.

3.4 Culture of TS Cells in Feeder-Free Conditions

1. Prepare TS cell conditioned media. TS cells can be cultured on tissue culture plates or dishes without gelatin.
2. Culture TS cells in 70 % TS cell conditioned media and 30 % TS cell media (e.g., for a 10 cm culture dish, add 7 mL of TS cell conditioned media and 3 mL of TS cell media and 25 ng/mL FGF4 and heparin to 1×). The morphology of TS cells cultured in feeder-free conditions should be similar to culture on iMEFs (Fig. 3). TS cell colonies should maintain tight cell–cell contact at the colony boundary.
3. TS cells grown in feeder-free conditions should be passaged in a similar manner as TS cells grown on iMEFs. It is important to maintain small clumps of cells during the trypsinization step because TS cells grown in feeder-free conditions are more susceptible to differentiation. Maintaining the appropriate confluency and split ratios is also an important factor to empirically determine when establishing new TS cell lines.

3.5 Differentiation of TS Cells

1. TS cells can be differentiated by culturing in TS cell media without FGF4 or heparin in feeder-free conditions. Passage TS cells as described above and replate in a 6-well culture plate in TS cell media without iMEFs or growth factors.
2. After 2 days, aspirate the old TS cell media and add 2 mL of fresh TS cell media without FGF4 and heparin.
3. After 4 days, differentiation of TS cells should be visible. Use an inverted bright-field microscope to visualize the morphology

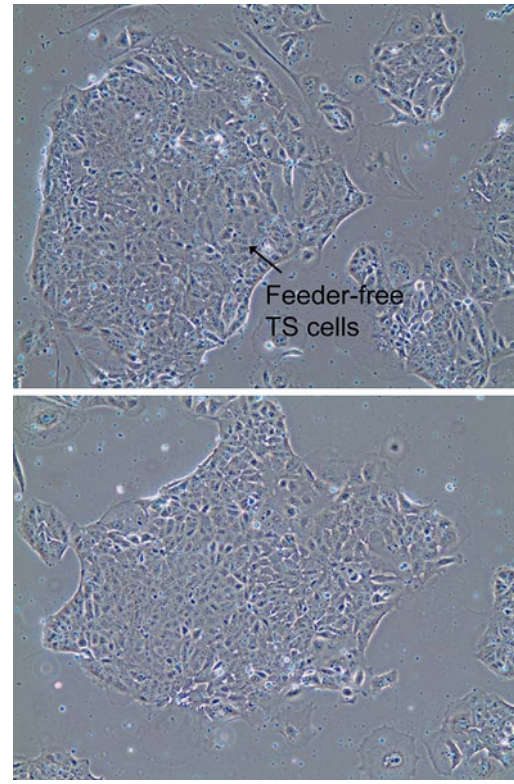


Fig. 3 TS cells grown in feeder-free conditions. Bright-field microscopy of trophoblast stem cells grown in feeder-free conditions maintain a tight cell boundary and typical colony morphology

of TS cells. TS cells undergoing differentiation should exhibit an increased cytoplasm to nucleus ratio. Giant cells should be highly visible as they are morphologically distinct from TS cells. Other cell types represented in the trophoblast lineage should be present but may not be as easily distinguished in the heterogeneous differentiation conditions. Polyploidy, which is a characteristic of giant cells and syncytiotrophoblasts, can also be visualized within 6 days of differentiation. TS cells cultured without FGF4 for 9–12 days should contain a mixture of differentiated cells including giant cells (Fig. 4).

4. TS cells can also be differentiated using small molecules such as retinoic acid (RA) [15]. To promote differentiation, add RA to a final concentration of 2 μM in TS cell media without FGF4 or heparin. Perform differentiation experiments as described above. Change medium every 2 days.

3.6 Production of Lentiviral Particles

1. Thaw a vial of frozen HEK 293 T cells in a bead or water bath, add several milliliters of MEF media to a 15 mL conical tube, and spin at 1,500 rpm (500 × *g*) for 3 min.

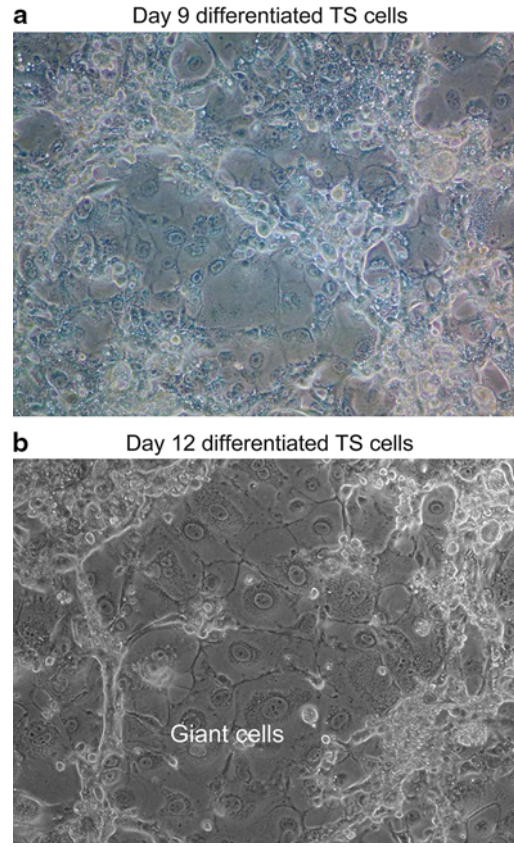


Fig. 4 Differentiation of TS cells. Bright-field microscopy of trophoblast stem cells differentiated for (**a**) 9 and (**b**) 12 days following removal of FGF4 and iMEFs. Note the presence of cells of the trophoblast lineage including giant cells, which exhibit an increased cytoplasmic to nuclear ratio relative to undifferentiated TS cells

2. Plate 293 T cells on a 10 cm culture dish in MEF media. Passage 293 T cells 1:10 or 1:15 when they reach 70 % confluency.
3. Once the 293 T cells reach an exponential growth phase, passage the cells 24 h before production of lentiviral particles at a density that will achieve ~20–25 % confluency the following day.
4. On the day of transfection, warm up Opti-MEM media to room temperature.
5. Next, add 1.5 mL of Opti-MEM media and 36 μL Lipofectamine 2000 to a 15 mL conical tube. Mix by gently swirling the tube several times. Incubate for 5 min at room temperature.
6. To another tube, add 1.5 mL of Opti-MEM media and 4 μg of envelope plasmid (pLP/VSVG), 10 μg packaging vector (e.g., psPAX2), and 10 μg of the shRNA lentiviral expression

vector (e.g., pGreenPuro, System Biosciences). Tap the tube gently to mix.

7. Add 1.5 mL of plasmid mixture to the tube containing 1.5 mL of Lipofectamine 2000. Mix by gently inverting the tube several times. Incubate for 20 min at room temperature.
8. Meanwhile, add 10 mL of MEF media (without pen–strep) to the 293 T cells.
9. Add 3 mL of the transfection mix slowly to the 293 T cells and gently swirl to mix.
10. Incubate at 37 °C for at least 4–6 h.
11. The next morning, remove the media and add 10 mL of TS cell media without FGF4 or heparin and incubate at 37 °C.
12. Harvest the supernatant after 24–48 h by filtering the virus-containing media through a 0.45 μM cell strainer and spinning at 1,500 rpm (500 × *g*) for 5 min.

3.7 Lentiviral Infection of TS Cells

1. At least 24 h before transduction, plate TS cells in a 6-well plate without feeders at a concentration that will achieve 50 % confluency the following day. TS cell cultures that are too sparse may result in elevated cell death following transduction, and cultures that are too dense may not be efficiently transduced.
2. To transduce the TS cells, add media containing viral particles and polybrene (4–8 μg/mL) to the 6-well plate containing TS cells. The next day, remove the media and add fresh TS cell media containing FGF4 and heparin.
3. TS cells can be stably selected in the presence of 1–2 μg/mL puromycin.
4. After 2 days, check the viability and transduction rate of infected TS cells using inverted bright-field microscopy (Fig. 5a), epifluorescence analysis (Fig. 5b), and FACS analysis.
5. Change the media 2 days after infection by aspirating the TS cell media containing puromycin and add fresh TS cell media containing FGF4 and heparin. Allow the TS cells to grow for another 2 days or until they reach 70 % confluency.
6. Harvest stably selected TS cells for shRNA knockdown experiments by washing 10 cm dish with 2 mL of PBS (1×), then add 1 mL of pre-warmed 0.05 % trypsin solution, and incubate at room temperature or 37 °C for 2–3 min or until colonies begin to disassociate. Spin TS cells at 1,500 rpm (500 × *g*) for 3 min and resuspend pellet in TS cell media with FGF4 and heparin.

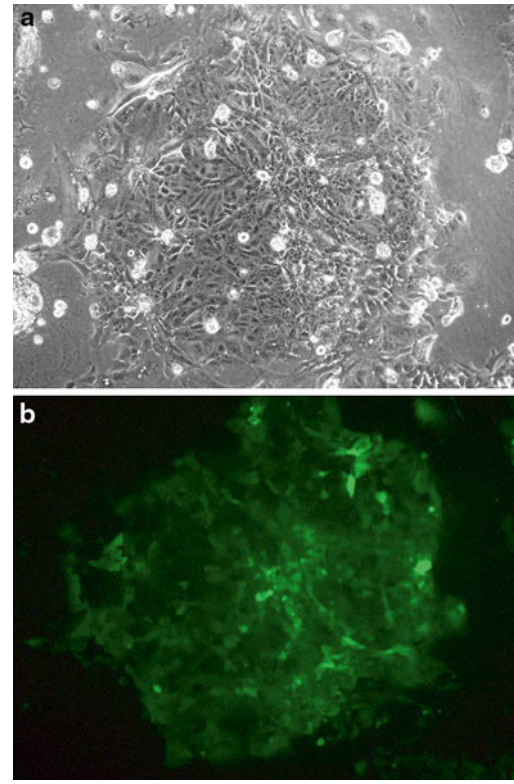


Fig. 5 Fluorescence microscopy of lentiviral particle-infected TS cells. (**a**) Bright-field and (**b**) immunofluorescence microscopy of TS cells infected with lentiviral particles encoding control shRNAs and GFP. Note the high expression of GFP following stable selection of infected TS cells in the presence of 1 μg/mL puromycin for 3 days

7. Individual TS cell clones can be picked and expanded for characterization, or bulk TS cell cultures can be expanded for downstream applications.
8. Perform downstream experiments (e.g., expression analysis, global ChIP).

4 Notes

1. It is important to avoid multiple freeze–thaw cycles.
2. Dissection instruments can be sterilized by autoclaving and sprayed with 70 % ethanol prior to use.
3. Mice should be euthanized according to the Institutional Animal Care and Use Committee (IACUC) protocol used for the study.
4. Blastocysts should be disaggregated when an appropriate outgrowth perimeter has been achieved.

References

1. Cross JC, Baczyk D, Dobric N, Hemberger M, Hughes M, Simmons DG, Yamamoto H, Kingdom JC (2003) Genes, development and evolution of the placenta. Placenta 24(2–3): 123–130
2. Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci U S A 78(12): 7634–7638
3. Tanaka S, Kunath T, Hadjantonakis AK, Nagy A, Rossant J (1998) Promotion of trophoblast stem cell proliferation by FGF4. Science 282(5396):2072–2075
4. Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Scholer H, Smith A (1998) Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell 95(3): 379–391
5. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R (2003) Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev 17(1):126–140
6. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell 113(5):631–642
7. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113(5):643–655
8. Strumpf D, Mao CA, Yamanaka Y, Ralston A, Chawengsaksophak K, Beck F, Rossant J (2005) Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. Development 132(9):2093–2102. doi:10.1242/dev.01801, dev.01801 [pii]
9. Russ AP, Wattler S, Colledge WH, Aparicio SA, Carlton MB, Pearce JJ, Barton SC, Surani MA, Ryan K, Nehls MC, Wilson V, Evans MJ (2000) Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature 404(6773):95–99
10. Luo J, Sladek R, Bader JA, Matthyssen A, Rossant J, Giguere V (1997) Placental abnormalities in mouse embryos lacking the orphan nuclear receptor ERR-beta. Nature 388(6644):778–782. doi:10.1038/42022
11. Nishioka N, Yamamoto S, Kiyonari H, Sato H, Sawada A, Ota M, Nakao K, Sasaki H (2008) Tead4 is required for specification of trophectoderm in pre-implantation mouse embryos. Mech Dev 125(3–4):270–283
12. Kidder BL, Palmer S (2010) Examination of transcriptional networks reveals an important role for TCFAP2C, SMARCA4, and EOMES in trophoblast stem cell maintenance. Genome Res 20(4):458–472. doi:10.1101/gr.101469.109, gr.101469.109 [pii]
13. Kidder BL, Palmer S (2012) HDAC1 regulates pluripotency and lineage specific transcriptional networks in embryonic and trophoblast stem cells. Nucleic Acids Res 40(7):2925–2939. doi:10.1093/nar/gkr1151, gkr1151 [pii]
14. Kidder BL, Oseth L, Miller S, Hirsch B, Verfaillie C, Coucouvanis E (2008) Embryonic stem cells contribute to mouse chimeras in the absence of detectable cell fusion. Cloning Stem Cells 10(2):231–248
15. Yan J, Tanaka S, Oda M, Makino T, Ohgane J, Shiota K (2001) Retinoic acid promotes differentiation of trophoblast stem cells to a giant cell fate. Dev Biol 235(2):422–432. doi:10.1006/dbio.2001.0300S0012-1606(01)90300-8 [pii]

Part V

Transcriptional Programs that Promote Self-Renewal, Reprogramming, and Transdifferentiation

Chapter 14

Conversion of Epiblast Stem Cells to Embryonic Stem Cells Using Growth Factors and Small Molecule Inhibitors

Jyoti Rao and Boris Greber

Abstract

Stem cell in vitro culture is a useful model system to study mechanisms underlying transitions between defined cell states. Epiblast stem cells, in addition to being capable of somatic differentiation, can be converted to a more primitive embryonic stem cell-like state, by overexpression of specific transcription factors. Here, we describe a reliable method to accomplish—and potentially further study—the transgene-independent reversion from epiblast stem cells to ES cells using administration of specific growth factors and small molecule inhibitors.

Key words Mouse embryonic stem cells, Epiblast stem cells, Self-renewal, EpiSC reversion

1 Introduction

Pluripotency is the ability of a cell to potentially generate all somatic cell types of a given organism. In the mouse, there are two types of pluripotent stem cells that can be isolated and expanded in an undifferentiated state in vitro. These are the preimplantation embryo-derived "embryonic stem cells" (ESCs) [1, 2] and the postimplantation embryo-derived "epiblast stem cells" (EpiSCs) [1, 2]. In line with their early embryonic origin, both ESCs and EpiSCs share several characteristics including the expression of the pluripotency controlling transcription factors OCT4, NANOG, and SOX2 [1]. ESCs and EpiSCs also show similarities in their global expression profiles [1, 3]. On the other hand, there are several other transcription factors specifically expressed in ESCs that assist the abovementioned trio in maintaining self-renewal of this developmentally more primitive cell state. Significantly, overexpression of any of these "accessory" ESC-specific transcription factors in EpiSCs is sufficient to convert these into ESCs [4–8].

Apart from the differential expression of these transcription factors, ESCs and EpiSCs also display different growth factor requirements to enable self-renewal. While EpiSCs rely on FGF

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_14, © Springer Science+Business Media New York 2014

and Nodal/Activin/TGFβ signaling [3, 5], self-renewal of mouse ESCs is promoted by activation of the LIF/STAT3 cascade, together with small molecule-based inhibition of FGF/ERK and GSK3β [11, 12]. Given that the activation or inactivation of signaling pathways may directly impact the expression of key genes in the two types of pluripotent stem cells, it was tempting to speculate that a switch in culture conditions may be sufficient to revert EpiSCs into ESCs. Indeed, transgene-independent EpiSC reversion has been demonstrated by several laboratories, whereas reversion efficiencies tend to vary depending on the specific culture conditions applied and the EpiSC line used [3, 9–12].

The reversion from EpiSCs to ESCs with the protocol described here [modified from ref. 5] proceeds rapidly, i.e., within a few days, and at comparatively high efficiency. This is enabled by several steps. First, it is helpful if the *basic* culture conditions, namely, substrate and basal medium, are compatible with both EpiSC and ESC growth. This is true for Knockout™ serum replacement-based medium and use of inactivated mouse embryonic fibroblast (MEF) feeder layers as substrate. For example, if EpiSCs are cultured on gelatin-coated dishes, which is compatible with ESC renewal but not EpiSC growth, the vast majority of EpiSCs would differentiate instead of reverting into ESCs. Secondly, it is important that the reversion process is initiated from single cells rather than from large flat EpiSC colonies, because ESC colonies form best from single cells. On the other hand, EpiSCs tend to apoptose when dissociated into single cells. Therefore, in the below protocol, EpiSC reversion is initiated from single cells but upon adding an EpiSC survival factor to the medium [13]. Thirdly, it is helpful if switching from EpiSC culture conditions to particularly *stringent* ESC conditions. These are the above mentioned two-inhibitor + LIF conditions [14]. Moreover, we found that inclusion of an additional inhibitor against SMAD2/3 signaling as well as pretreating EpiSCs with the reversion medium prior to passaging more efficiently selects against the EpiSC state and significantly enhances reversion efficiencies, respectively. Taken together, the protocol potentially enables sufficiently high EpiSC reversion efficiencies to further investigate the transition between the two pluripotent stem cell states in vitro, even as bulk cultures.

2 Materials

2.1 Stocks of Cells, Growth Factors, and Inhibitors

1. Cells:
   - Vials of frozen mitotically inactivated MEFs. Store in liquid nitrogen or at −150 °C (*see* **Note 1**).
   - Vials of frozen EpiSCs, e.g., of line E3 [3]. Store in liquid nitrogen or at −150 °C (*see* **Note 2**).

2. Recombinant growth factors:
   - Basic fibroblast growth factor (FGF2, mouse or human): Carefully dissolve in 0.1 % cell culture grade bovine serum albumin/PBS to yield a final concentration of 10 μg/mL. Freeze aliquots at −20 °C. Once thawed, FGF2 aliquots are stored at 4 °C and used within a week. Avoid repeated freeze-thaw cycles.
   - Leukemia inhibitory factor (LIF, e.g., Millipore, mouse, or human, 10 μg/mL). Store at 4 °C.
3. Inhibitors:
   - ROCK inhibitor Y27632, MEK inhibitor PD032590, ALK4/5/7 inhibitor SB431542, and GSK3β inhibitor CHIR99021 (*see* **Note 3**): Dissolve in DMSO at 10 mM. Store as aliquots in amber tubes at −20 °C. Thawed aliquots can be kept at 4 °C for days or be refrozen.

2.2 Preparation of Feeder Layers

1. Coating solution: 0.2 % gelatin solution in cell culture grade phosphate buffered saline (PBS) and filter-sterilize. Store at room temperature.
2. MEF culture medium: Knockout™ DMEM (catalogue number 10829018, Life Technologies—or conventional DMEM) supplemented with 10 % fetal bovine serum (FBS, e.g., Biowest) and 1 % L-glutamine/penicillin/streptomycin mix (e.g., catalogue number P11013, PAA). Store at 4 °C.
3. Tissue culture 6-well plates.

2.3 Cell Culture

1. EpiSC culture medium: Knockout™ DMEM supplemented with 20 % Knockout™ serum replacement (catalogue number 10828028, Life Technologies) and L-glutamine/penicillin/streptomycin. Can be stored at 4 °C for days. Freshly add FGF2 to a final concentration of 5 ng/mL before use.
2. Reversion medium: Knockout™ DMEM supplemented with 20 % Knockout™ serum replacement and L-glutamine/penicillin/streptomycin. Can be stored at 4 °C for days. Before use, add 0.5 μM PD0325901, 3 μM CHIR99021, and 20 ng/mL LIF from stocks. 10 μM SB431542 is additionally included in the medium where indicated in the protocol.

2.4 Splitting Cells

1. Cell culture grade PBS without bivalent cations.
2. Collagenase Type IV: Dissolve at 2 mg/mL in KO-DMEM, snap-freeze aliquots, and store at −20 °C. Pre-warm to 37 °C before use (*see* **Note 4**).
3. Accutase™ (e.g., catalogue number L11-007, PAA). Store frozen as aliquots. Pre-warm to 37 °C before use.
4. Sharp plastic scrapers (e.g., catalogue number 99010, TPP).

5. Hemocytometer.
6. Strongly recommended but not absolutely required: Stereo microscope with good transmission light source placed in laminar flow hood.

3 Methods

3.1 Preparing Feeder Layers

1. Add 2 mL of gelatin solution per well of a 6-well plate and let stand for about 2 h at RT. Pre-warm MEF medium to 37 °C (*see* **Note 5**).
2. Thaw vial of frozen mitotically inactivated MEFs in 37 °C water bath, add cells to several milliliters of pre-warmed MEF medium in a 15 mL conical tube, and centrifuge at 200×*g* for 2 min.
3. Aspirate off gelatin solution from 6-well plate, gently resuspend cell pellet in appropriate volume of MEF medium, and plate at 250,000 cells per well (*see* **Note 6**). Upon placing the plate into the cell culture incubator (37 °C, 5 % CO_2, saturated humidity), agitate plate in several quick figure eight motions—this ensures a uniform distribution of cells in the wells.

3.2 Culturing EpiSCs

1. Next day, confirm that the feeder cells are evenly distributed (cover the entire surface of the well(s) and look overall "healthy") (Fig. 1a). Prepare and warm up several milliliters of EpiSC medium supplemented with 10 μM ROCK inhibitor (*see* **Note 7**).
2. Aspirate off the MEF medium and replace with 2 mL of fresh EpiSC medium containing ROCK inhibitor. Return the culture plate back to the incubator.

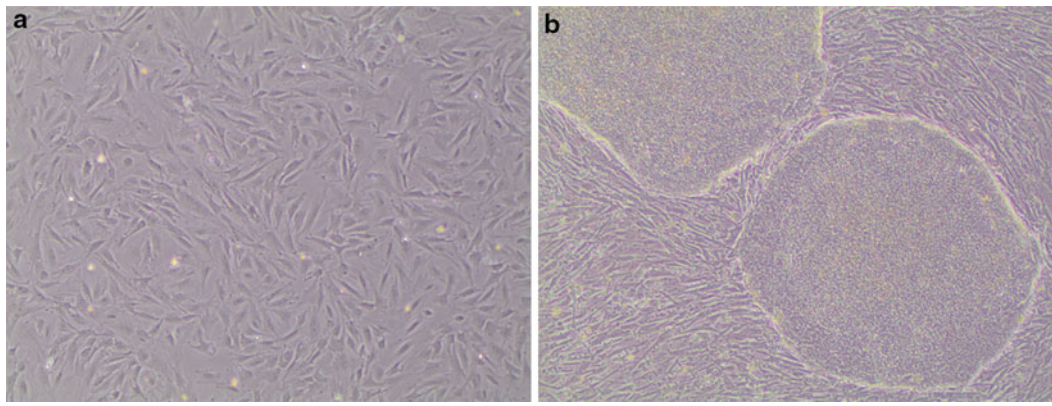


Fig. 1 Culturing EpiSCs. (**a**) Appropriate morphology and density of MEFs 1 day after seeding 250,000 frozen cells per well of a 6-well plate. (**b**) Typical morphology of EpiSC colonies (line E3). EpiSCs grow as flat large colonies with sharp boundaries. Note that the morphology of the feeder layer is different as a result of FGF2 contained in the EpiSC medium

3. Thaw frozen vial of EpiSCs in a 37 °C water bath. Quickly transfer thawed EpiSCs to a 15 mL tube containing ~4 mL of EpiSC medium. Spin briefly at 200 × *g* for several seconds. In this way, the small aggregates of EpiSCs will loosely settle at the bottom of the tube while the individual cells will remain in the supernatant (*see* **Note 8**).
4. Soak off the supernatant and, using a 1 mL pipette, gently resuspend in ~300 μL of EpiSC medium. Dropwise add to prepared well(s) with feeders (*see* **Note 9**). Upon placing the plate back into the incubator, swirl plate twice in clockwise orientation and then once in an anticlockwise manner, to distribute the clumps evenly (*see* **Note 10**).
5. Next day, replace the used medium with 2 mL of fresh EpiSC medium without ROCK inhibitor (*see* **Note 11**). Replace with increasing volumes of fresh medium every day (*see* **Note 12**). Within the next days, EpiSCs will grow to form large flat colonies (Fig. 1b).

   Once the EpiSC colonies start reaching sub-confluence, they will need to be passaged. For maintaining the cells as EpiSCs, follow the next steps. For initiating EpiSC reversion to a mESC-like state, proceed to Subheading 3.3.
6. To passage the EpiSCs, prepare fresh feeder layers as described above (*see* **Note 13**). Next day, replace MEF medium by pre-warmed EpiSC medium with 10 μM ROCK inhibitor and pre-warm an aliquot of collagenase solution (1 mL needed per well of a 6-well plate) (*see* **Note 14**).
7. Wash the cells once with several milliliters of PBS, add 1 mL of collagenase solution, and incubate for 5–10 min at 37 °C.
8. Once the feeder layer loosens up significantly as a result of the collagenase treatment, remove the collagenase solution and wash with several milliliters of PBS. Add 1 mL of EpiSC medium to the well and—ideally under a stereo microscope and using a sharp plastic spatula—scrape off the still-intact EpiSC colonies along with the loosened feeder layer. Using a 1 mL pipette, carefully pipette up and down several times to break up the colonies into smaller clumps. Ideally, you should end up with few single cells but many EpiSC aggregates in the range of ~100 to several hundred cells (*see* **Note 15**).
9. Transfer the suspension to a 15 mL conical tube. To eliminate single EpiSCs and old MEFs, centrifuge for only a few seconds at 200 × *g*. Remove the supernatant and gently resuspend the clumps in ~300 μL of EpiSC medium. Be careful not to break up the clumps any further. In a dropwise manner, add EpiSC clumps to new feeder-containing wells with equilibrated EpiSC medium. Feed daily with EpiSC medium as before (*see* **Note 16**).

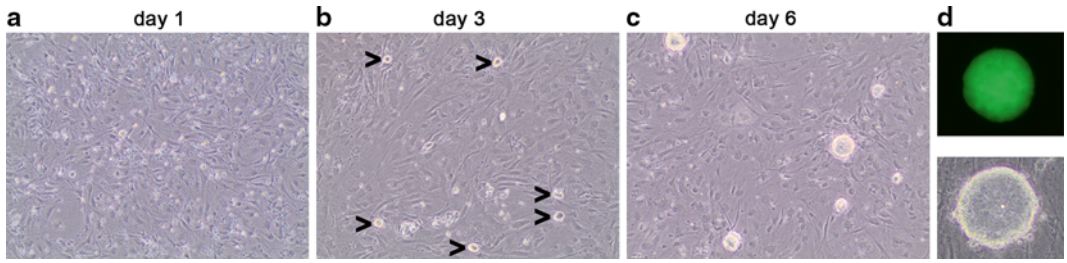


Fig. 2 Media-induced reversion of EpiSCs. (**a**, **b**) EpiSCs usually revert to an ESC-like state within 3 days after passaging as single cells. ESC-like cells grow as compact and dome-shaped colonies (see *arrows*). Reverted colonies grow larger on subsequent days (**c**) and could be picked/bulk passaged at this stage for establishing pure cultures of ESC-like cells. (**d**) Oct4-GFP fluorescence and phase contrast morphology of a typical reverted colony (EpiSC line used: E3)

3.3 Reversion of EpiSCs into ESCs

1. Estimate the day when cells may be ready for splitting. This may often be the 5th day after splitting or thawing of the EpiSCs. Two days before that day, i.e., on day 3, replace the EpiSC medium with freshly prepared reversion medium instead of feeding with EpiSC medium.
2. On day 4, feed EpiSCs with reversion medium and prepare a fresh well with feeders as above (*see* **Note 17**).
3. On day 5, prepare and warm up ~4 mL of reversion medium supplemented with 10 μM ROCK inhibitor. Use 2 mL to replace MEF medium on fresh feeder-containing well. Place plate back into incubator.
4. Wash the sub-confluent EpiSCs with PBS and add 1 mL of pre-warmed Accutase™ containing 10 μM ROCK inhibitor. Incubate for 5–10 min at 37 °C, then inactivate enzyme by adding 1 mL of reversion medium (*see* **Note 18**). Using a 1 mL pipette, pipette up and down several times to completely dissociate the cells. Transfer the cell suspension into a 15 mL tube and centrifuge at 200 × *g* for 2 min.
5. Resuspend the pellet in 1 mL reversion medium by gently pipetting up and down, determine cell titer using a hemocytometer and plate ~20,000 cells into the fresh well containing feeders, reversion medium, and ROCK inhibitor (*see* **Note 19**). Upon placing the plate back into the incubator, agitate as when plating MEFs (see above), to distribute cells uniformly within the well.
6. Next day, feed with fresh reversion medium (without ROCK inhibitor). The day after, when small ESC-like colonies already start to emerge, start feeding with reversion medium plus 10 μM SB431542 over the next few days (*see* **Note 20**).
7. Over the next days, allow reverted colonies to expand (Fig. 2, *see* **Note 21**). Daily feed the cells with slightly increasing volumes of reversion medium containing 10 μM SB431542. Figure 3 shows typical day-6 morphologies and illustrates the

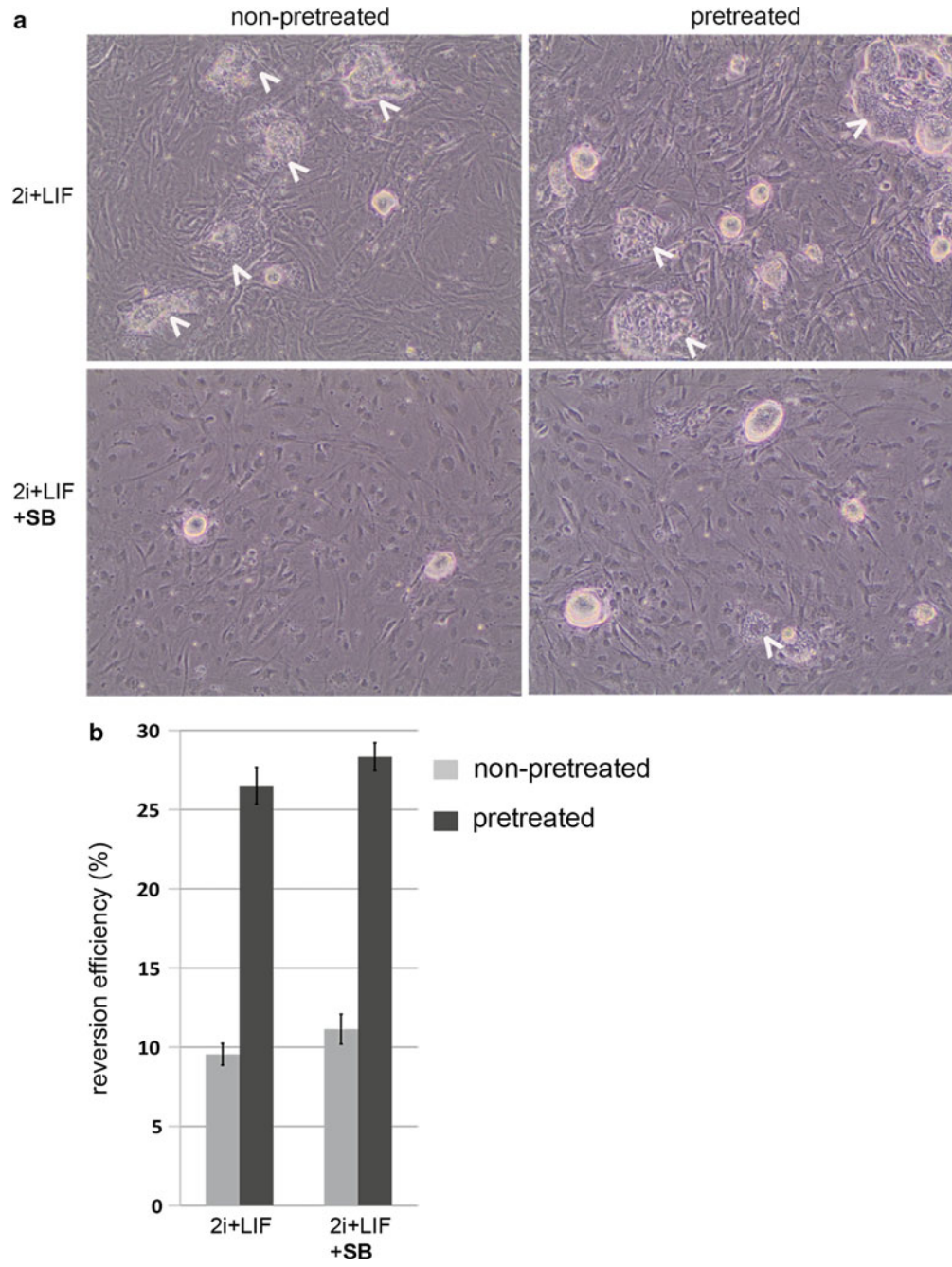


Fig. 3 Effects of 2i + LIF pretreatment and SB431542 addition in EpiSC reversion experiments. (**a**) EpiSC pretreated with 2i + LIF for 2 days before single-cell passaging reverted more efficiently into ESC-like cells compared to non-pretreated cultures (note higher density of dome-shaped colonies in *right panels* compared to *left panels*). In comparison, addition of SB during EpiSC reversion did not appear to increase the numbers of reverted colonies, but it blocked the expansion of remaining EpiSC-like colonies (compare *top* and *bottom panels—white arrows* indicate non-ESC-like background colonies). Hence, pretreatment of EpiSCs with 2i + LIF enhances reversion efficiency, while SB addition improves purity of reverted cultures. (**b**) Quantification of results in part (**a**): Cultures shown in (**a**) were stained for alkaline phosphatase activity to determine reversion efficiencies: Reversion efficiency was defined as the percentage of dome-shaped, strongly alkaline phosphatase-positive colonies compared to the total number of colonies. Note that SB addition did not change the numbers of these background colonies, but it effectively prevented their further expansion

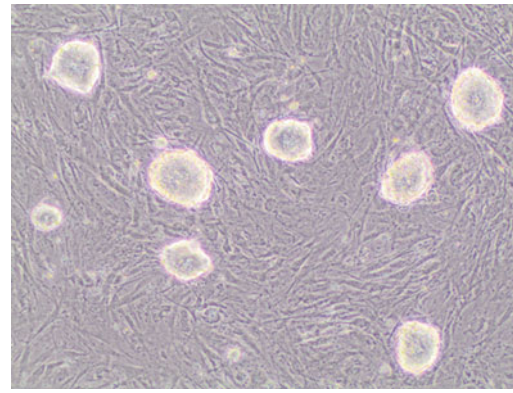


Fig. 4 Stable culture of reverted cells: ESC-like colonies were picked from cultures shown in Fig. 2c, dissociated into single cells and replated in 2i + LIF

effects of SB addition and of pretreating cells with reversion medium before splitting. Under these optimized conditions, after about 1 week, most background cells will be differentiated and the cultures will be free of remaining EpiSC-like colonies.

3.4 Passaging Reverted EpiSCs

Depending on the numbers of mESC-like colonies, by day ~7, the cultures can either be bulk-passaged or mESC-like colonies need to be picked for establishing pure cultures of mESCs. The following procedure is for picking mESC-like colonies:

1. Prepare fresh well with feeders 1 day in advance or coat a culture well with gelatin (*see* **Note 22**).
2. Pre-warm reversion medium and add to new well. Place plate into incubator.
3. As reverted cells can easily be identified by their typical compact dome-shaped morphology (Figs. 2 and 3), isolating these colonies is convenient. Under a stereo or conventional cell culture microscope, using a 200 μL pipette tip, carefully detach the reverted colonies without much damaging the feeder layer (*see* **Note 23**).
4. Collect all the detached and floating colonies into, e.g., 1 well of a 96-well plate containing 100 μL of reversion medium (*see* **Note 24**). Allow the colonies to settle down, carefully soak off the medium, wash the colonies with 200 μL PBS, and finally resuspend in 100 μL of pre-warmed Accutase™.
5. Incubate for ~10 min at 37 °C and then dissociate the colonies into single cells by pipetting up and down. Add to fresh gelatin-coated or feeder-containing well with 2i + LIF medium. Feed with fresh medium every day and allow cells to grow into stable ESC colonies (Fig. 4). The cells can now be handled like bona fide mESCs and are molecularly and functionally indistinguishable from these [3].

4 Notes

1. A description about isolating MEFs goes beyond the scope of this chapter. Less experienced users may obtain frozen vials of inactivated MEFs from commercial suppliers. We usually freeze one million inactivated MEFs per vial in 10 % DMSO/MEF medium.
2. To date, a number of established EpiSC lines have successfully been reverted to a mESC-like cell state by switching culture conditions. But there are differences in growth behavior and reversion efficiencies. Line E3 is an easy-to-maintain line and shows robust reversion rates. It is freely available upon request.
3. With these small molecules, we have not seen significant differences in activity between different manufacturers.
4. In comparison to other applications, the activity of collagenase IV in the EpiSC splitting procedure is not a particularly critical factor. Activity testing of batches from specialized manufacturers has never been necessary in our hands.
5. We usually use 6-well plates for most purposes, but the protocol can certainly be adapted to any other plate format like, for instance, single 6 cm dishes; gelatin coating serves to enhance the longer-term attachment of MEF cells to some degree.
6. For example, if there are one million inactivated MEF cells per Cryovial, this would be sufficient to prepare feeders in 4 wells of a 6-well plate. So, in this example, resuspend cells in 10 mL of MEF medium and dispense 2.5 mL each into 4 wells of a gelatin-coated 6-well plate. For thawing or maintaining EpiSC, you may actually need less numbers of wells. If using a different plate format, scale up or down accordingly. The aim is to hit the right feeder density: Next day, there should be no gaps between the feeder cells, but they should still have enough space to spread out nicely. If you are not sure about the survival rates of your MEFs after thawing, you may also want to plate out a dilution series of different titers into several gelatin-coated wells. Next day, choose the well with the most appropriate feeder density for plating out the EpiSCs.
7. In principle, you will only need one well of EpiSCs for initiating a reversion experiment. But you may want to have a second well running for maintaining the EpiSCs. So, although this also depends on the titer at which the EpiSCs have been frozen and on the EpiSC survival rates after thawing, you could, e.g., plate out the contents of one frozen EpiSC vial into two wells of a 6-well plate. For that, you would need about 2 mL of EpiSC medium per well, i.e., 4 mL. In addition, you need several mL of extra medium in the thawing process. So, in this example, make up ~8 mL of EpiSC medium containing ROCK inhibitor.

8. EpiSCs are usually frozen as small aggregates of cells.
9. As said, the optimal amount of EpiSC suspension from frozen vials to be plated out per well is dependent on several factors such as overall frozen cell titer, clump size, and cell line-specific survival. There are no strict rules. If you are uncertain about how much of the EpiSC suspension to plate out per well, you may also want to try plating different amounts across different wells of a 6-well plate. Several days later, you will be able to judge in which sample you hit a good density and may then proceed with this one. We usually plate EpiSC aggregates out under a stereo microscope placed into a laminar flow hood. This allows you to directly monitor and control the numbers of cell aggregates that you plate.
10. Note that this way of agitating the plate is different from the way you need to agitate it when plating out single cells. You can also practice a bit and check the resulting distribution of EpiSC clumps within the well under a stereo microscope. In general, achieving an even distribution of clumps—and thereby EpiSC colonies forming subsequently—is quite important for successfully maintaining EpiSC cultures in an undifferentiated state.
11. Upon feeding with EpiSC medium, you may notice that the MEFs undergo a change in morphology: They will become more elongated and stringy. This is normal and results from the FGF2 contained in the medium.
12. EpiSCs grow fast and tend to significantly acidify the medium. Start with 2 mL of medium per well of a 6-well plate on the first day and increase the volume on subsequent days. Again, there are no strict rules. You need to frequently observe the cultures. Do not change medium without first observing the cells. Pay attention to overall confluency, occurrence of spontaneous differentiation, and color of the medium. Do not allow medium to turn bright yellow overnight. Toward the end of a passage, you may need to feed with as much as 6 mL of medium per well.
13. The number of new feeder-containing wells depends on your needs. Some EpiSC lines can be split at ratios of up to 1:10, but mostly, there is no need to culture EpiSCs in such many wells. One possible way of handling the cells would, e.g., be to only replate a fraction of the cells from one well into two new ones: One of these could later be used for initiating an experiment, while the other one could be used for further maintaining the cell line in culture.
14. Addition of ROCK inhibitor not only promotes cell survival after thawing but is also useful when splitting aggregates of cells.
15. While scraping, apply sharp and quick strokes to avoid uncontrolled disruption of EpiSC colonies. When pipetting up and down to break up the colonies, you should aim for ending up with few single cells but instead with many EpiSC aggregates

in the range of ~100 to several hundred cells. You can monitor this by carrying out the procedure under a stereo microscope. Alternatively, check results of the pipetting under a conventional cell culture microscope at low magnification.

16. As before, no strict rules regarding splitting ratios can be given. It is best to plate out a few droplets of EpiSC suspension into one well, monitor the resulting clump density under a (stereo) microscope and then decide whether to add more cell aggregates to the well or not. When splitting EpiSCs, the aim is to achieve a seeding density of several hundred aggregates per well of a 6-well plate. Be aware that when including ROCK inhibitor in the medium, the vast majority of aggregates will indeed attach on the feeder layer to form new colonies. When starting to work with a new cell line, it may also be useful to try out plating different EpiSC densities into several feeder-containing wells, to compare results over the following days and gain experience this way.
17. In comparison with our previous procedure [3], we found that pretreating EpiSCs with reversion medium before splitting enhances reversion efficiencies (Fig. 3). However, the time-point to initiate the pretreatment has not been fully optimized. When starting the pretreatment on day 3, as suggested here, the EpiSCs will still look like normal EpiSC colonies on day 5.
18. To promote EpiSC reversion into ESCs, the EpiSCs are dissociated to single cells, in contrast to what is done when routinely splitting EpiSCs. Accutase™ enables mild cell dissociation but might be substituted by comparable products. Addition of ROCK inhibitor is useful even during Accutase™ treatment, to prevent blebbing of the cells and subsequent cell death.
19. ROCK inhibitor does not promote survival of reverted cells but only of EpiSC. However, since at this point the vast majority of cells are still EpiSC-like, addition of ROCK inhibitor ensures that sufficient numbers of EpiSC survive the procedure to attach to the feeder layer.
20. Under 2i + LIF conditions, depending on the cell line, few or many cells may first remain in the EpiSC state, despite the fact that 2i + LIF medium does not support self-renewal of EpiSCs. Therefore, the addition of SB431542 serves to better select against the epiblast state because EpiSCs are strictly dependent on Nodal/Activin/TGFβ signaling. As a result, several days after passaging and SB treatment, the cultures will essentially be free of remaining EpiSCs (Fig. 3). The time-point of SB431542 addition has not been fully optimized.
21. If the EpiSCs carry a reporter transgene such as Oct4-GFP, green fluorescence can be used as an additional hallmark for identifying reverted colonies (Fig. 2d).
22. Reverted cells can be maintained on feeders or under feeder-free conditions on gelatin.

23. The colonies are attached to the feeder layer with a relatively small surface, and hence, they can be detached from it quite easily. Not damaging the feeder layer will also help reducing contamination from non-reverted cells.
24. Other plate formats or vessels can also be used at this step. The point here is simply to collect the colonies in a rather small volume of medium.

Acknowledgment

This work was supported by the Bundesinstitut für Risikobewertung (BfR) grant FK-3-1329-471 and the Chemical Genomics Centre of the Max Planck Society.

References

1. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, Gardner RL, McKay RD (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 448(7150):196–199
2. Brons IG, Smithers LE, Trotter MW, Rugg-Gunn P, Sun B, de Sousa C, Lopes SM, Howlett SK, Clarkson A, Ahrlund-Richter L, Pedersen RA, Vallier L (2007) Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature 448(7150):191–195
3. Greber B, Wu G, Bernemann C, Joo JY, Han DW, Ko K, Tapia N, Sabour D, Sterneckert J, Tesar P, Scholer HR (2010) Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. Cell Stem Cell 6(3):215–226
4. Guo G, Yang J, Nichols J, Hall JS, Eyres I, Mansfield W, Smith A (2009) Klf4 reverts developmentally programmed restriction of ground state pluripotency. Development 136(7):1063–1069
5. Yang J, van Oosten AL, Theunissen TW, Guo G, Silva JC, Smith A (2010) Stat3 activation is limiting for reprogramming to ground state pluripotency. Cell Stem Cell 7(3):319–328
6. Guo G, Smith A (2010) A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency. Development 137(19):3185–3192
7. Gillich A, Bao S, Grabole N, Hayashi K, Trotter MW, Pasque V, Magnusdottir E, Surani MA (2012) Epiblast stem cell-based system reveals reprogramming synergy of germline factors. Cell Stem Cell 10(4):425–439
8. Hall J, Guo G, Wray J, Eyres I, Nichols J, Grotewold L, Morfopoulou S, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A (2009) Oct4 and LIF/Stat3 additively induce Kruppel factors to sustain embryonic stem cell self-renewal. Cell Stem Cell 5(6):597–609
9. Bao S, Tang F, Li X, Hayashi K, Gillich A, Lao K, Surani MA (2009) Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells. Nature 461(7268): 1292–1295
10. Zhou H, Li W, Zhu S, Joo JY, Do JT, Xiong W, Kim JB, Zhang K, Scholer HR, Ding S (2010) Conversion of mouse epiblast stem cells to an earlier pluripotency state by small molecules. J Biol Chem 285(39):29676–29680
11. Bernemann C, Greber B, Ko K, Sterneckert J, Han DW, Arauzo-Bravo MJ, Scholer HR (2011) Distinct developmental ground states of epiblast stem cell lines determine different pluripotency features. Stem Cells 29(10): 1496–1503
12. Hanna J, Markoulaki S, Mitalipova M, Cheng AW, Cassady JP, Staerk J, Carey BW, Lengner CJ, Foreman R, Love J, Gao Q, Kim J, Jaenisch R (2009) Metastable pluripotent states in NOD-mouse-derived ESCs. Cell Stem Cell 4(6):513–524
13. Watanabe K, Ueno M, Kamiya D, Nishiyama A, Matsumura M, Wataya T, Takahashi JB, Nishikawa S, Muguruma K, Sasai Y (2007) A ROCK inhibitor permits survival of dissociated human embryonic stem cells. Nat Biotechnol 25:681–686
14. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A (2008) The ground state of embryonic stem cell self-renewal. Nature 453(7194):519–523

Chapter 15

Generation of Induced Pluripotent Stem Cells Using Chemical Inhibition and Three Transcription Factors

Benjamin L. Kidder

Abstract

Generation of induced pluripotent stem (iPS) cells from differentiated cells has traditionally been performed by overexpressing four transcription factors: Oct4, Sox2, Klf4, and c-Myc. However, inclusion of c-Myc in the reprogramming cocktail can lead to expansion of transformed cells that are not fully reprogrammed, and studies have demonstrated that c-Myc reactivation increases tumorigenicity in chimeras and progeny mice. Moreover, chemical inhibition of Wnt signaling has been shown to enhance reprogramming efficiency. Here, we describe a modified protocol for generating iPS cells from murine fibroblasts using chemical inhibition and overexpression of three transcription factors. Using this protocol, we observed robust conversion to iPS cells while maintaining minimal contamination of partially reprogrammed transformed colonies.

Key words Embryonic stem cells, Induced pluripotent stem cells, Reprogramming, Oct4, Sox2, c-Myc, Klf4, Inhibitors, Small molecules, Transformation

1 Introduction

Embryonic stem (ES) cells have the ability to self-renew indefinitely and differentiate into all cells represented in the three germ layers. As such, ES cells are an ideal model system to study mammalian development, genetics, and epigenetics and are a potentially unlimited source of cells for therapeutic application, disease modeling, drug screening, and personalized medicine. Induced pluripotent stem (iPS) cells, which are typically generated from differentiated cells by overexpression of four transcription factors including Oct4, Sox2, Klf4, and c-Myc, are considered to be functionally equivalent to ES cells, through their ability to differentiate in vitro into multiple lineages and in vivo through teratoma formation and by contributing to cells represented in the three germ layers in mouse chimeras [1–3]. Activation of Wnt signaling via exogenous proteins or chemical inhibition of glycogen synthase kinase-3 (GSK3) signaling has been shown to enhance reprogramming in the

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_15, © Springer Science+Business Media New York 2014

presence of the self-renewal factor LIF [4, 5]. Moreover, while c-Myc has been shown to bind promoters of self-renewal genes in ES cells [6, 7], inclusion of c-Myc in the reprogramming mix may lead to induction of transformed cells that are not fully reprogrammed, and studies have demonstrated that c-Myc reactivation increases tumorigenicity in chimeras and progeny mice [8]. In this chapter, we describe a modified protocol for generating iPS cells using chemical inhibition and overexpression of three transcription factors. This protocol allows for robust derivation of iPS cells under conditions that minimize expansion of contaminating transformed cells

2 Materials

2.1 Cell Culture

1. Dulbecco's Modified Eagle's Medium (DMEM), high glucose.
2. Penicillin–streptomycin (100×) (Invitrogen).
3. L-Glutamine (200 mM).
4. Fetal bovine serum (FBS, ES cell qualified).
5. Nonessential amino acids (NEAA, 100×) (Invitrogen).
6. Phosphate buffered saline without calcium or magnesium (PBS, 1×).
7. 0.25 % trypsin–EDTA (Invitrogen).
8. 2-Mercaptoethanol (100×, cell culture grade).
9. 24-, 12-, and 6-well culture dishes.
10. 0.1 % gelatin solution in water.
11. Polybrene.
12. 5, 15, and 50 mL polystyrene conical tubes.
13. Lab Armor bead bath.
14. Mitotically inactivated mouse embryonic fibroblasts (iMEFs) derived from E13.5 to E14.5 mouse embryos. Store in liquid nitrogen until use (*see* **Note 1**).
15. Mouse embryonic stem cells (ES cells) with a normal karyotype and at low passage. Store in liquid nitrogen until their use (*see* **Note 2**).

    Growth factor.

16. Mouse leukemia inhibitory factor (LIF, e.g., Millipore, 10^6 or 10^7 units). Store at 4 °C.

    Small molecule inhibitors.

17. GSK3β inhibitor CHIR99021 (GSK3i, Stemgent or Selleck Chemicals, *see* **Note 3**). Resuspend CHIR99021 in DMSO to a

final concentration of 10 mM. Aliquot and store at −20 °C. Store thawed aliquots at 4 °C for several days.

Cell culture.

18. Mouse ES cell culture media: DMEM high glucose, 15 % ES cell-qualified FBS, LIF (10 ng/mL), 1× penicillin–streptomycin (pen–strep), 1× glutamine, 1× 2-mercaptoethanol, and nonessential amino acids (NEAA) at 37 °C with 5 % CO_2.
19. MEF culture media: DMEM high glucose, 10 % FBS, 1× pen–strep, and 1× glutamine at 37 °C with 5 % CO_2.
20. 0.1 % gelatin solution in water.
21. Lab armor bead bath.

Passage of cells.

22. PBS without calcium and magnesium stored at room temperature.
23. 0.25 % trypsin–EDTA (1×, phenol red) warmed to 37 °C.

2.2 Retrovirus Production

1. HEK 293T cells.
2. Opti-MEM medium (Invitrogen).
3. Lipofectamine 2000 (Invitrogen).
4. Retroviral plasmids: pMXs-Oct4, pMXs-Sox2, pMXs-Klf4, MLV VSV-G plasmid (Addgene), and MLV gag-pol (Addgene).
5. 10 cm culture dishes.
6. 0.45 μM cell strainer (BD Biosciences).

2.3 Reagents for Isolation of Total RNA

1. Qiagen miRNeasy or RNeasy Kit.

3 Methods

3.1 Preparation of Cell Culture Media

1. Mouse embryonic fibroblasts (MEFs) are cultured in media containing DMEM high glucose, 10 % FBS, penicillin–streptomycin (1×), and L-glutamine (1×) at 37 °C with 5 % CO_2. Mitotically inactivated MEFs (iMEFs) can be harvested from mouse embryos (E13.5–14.5) [9] or purchased from a commercial vendor. Reprogramming can be performed with or without a MEF feeder layer.
2. Generation of iPS cells is performed in ES cell media containing DMEM high glucose, 15 % ES cell-qualified FBS, LIF (10 ng/mL), penicillin–streptomycin (1×), glutamine (1×), 2-mercaptoethanol (1×), and nonessential amino acids (1×) at 37 °C with 5 % CO_2.

3.2 Preparation of Feeder Layer

1. Twenty-four hours before infection of MEFs with retroviral particles iMEFs are cultured on 24-, 12-, and 6-well gelatin-coated plates. Add 2 mL of gelatin solution to culture plates and incubate at 37 °C for 10 min.
2. Pre-warm MEF media at 37 °C in a Lab Armor bead bath or water bath.
3. Thaw a vial of frozen MEFs in a Lab Armor bead bath or water bath, add cells to 3 mL of pre-warmed MEF media in a 15 mL conical tube, and spin at 1,500 rpm for 3–5 min.
4. Meanwhile, remove gelatin from culture plates. After spinning has completed, resuspend iMEFs in MEF medium, and plate the iMEFs in the culture dishes. The total volume of MEF media for a 24-, 12-, and 6-well plates should be 0.5, 1, and 2 mL, respectively. Before placing the plate in the incubator, move the plate vertically then horizontally to evenly distribute MEFs.

3.3 Production of Retroviral Particles

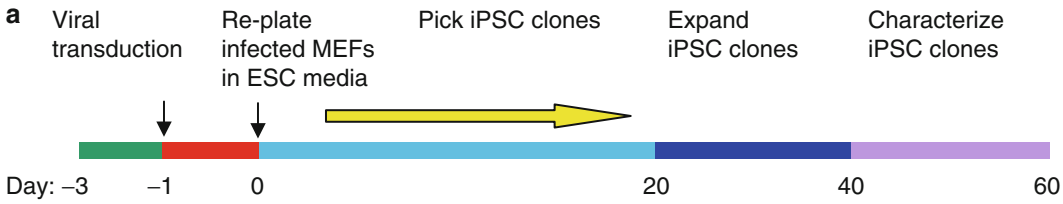
1. Thaw a vial of frozen HEK 293T cells in a bead or water bath, add several milliliters of MEF media to a 15 mL conical tube, and spin at 1,500 rpm for 3 min.
2. Plate 293T cells on a 10 cm culture dish in MEF media. Passage 293T cells 1:10 or 1:15 when they reach 70 % confluency.
3. Once the 293T cells reach an exponential growth phase, passage the cells 24 h before production of retroviral particles at a density that will achieve ~20–25 % confluency the following day.
4. The day of transfection, warm up Opti-MEM media to room temperature.
5. Next, add 1.5 mL of Opti-MEM media and 36 μL Lipofectamine 2000 to a 15 mL conical tube. Mix by gently swirling the tube several times. Incubate for 5 min at room temperature.
6. To another tube, add 1.5 mL of Opti-MEM media and 4 μg of envelope plasmid (pVSV-G) and 10 μg of MLV gag-pol packaging vector and 10 μg of the reprogramming vectors (pMXs-Oct4, pMXs-Sox2, and pMXs-Klf4). Tap the tube gently to mix.
7. Add 1.5 mL of plasmid mixture to the tube containing 1.5 mL of Lipofectamine 2000. Mix by gently inverting the tube several times. Incubate for 20 min at room temperature.
8. Meanwhile, add 10 mL of MEF media (without pen–strep) to the 293T cells.
9. Add 3 mL of the transfection mix slowly to the 293T cells and gently swirl to mix.
10. Incubate at 37 °C for at least 4–6 h.
11. The next morning, remove the media and add 10 mL of MEM media and incubate at 37 °C.

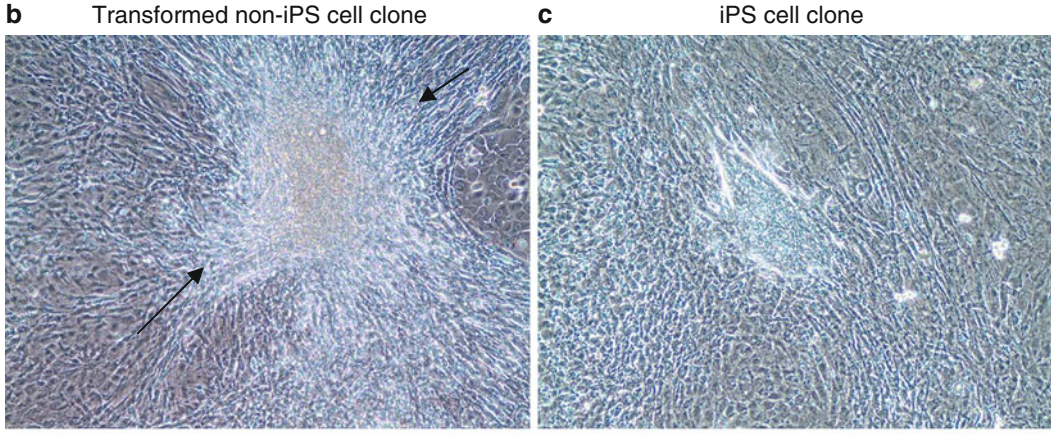
12. Harvest the supernatant after 24–48 h by filtering the virus-containing media through a 0.45 μM cell strainer and spinning at 2,500 rpm for 5 min.
13. Transduce MEFs by adding media containing viral particles and polybrene (4–8 μg/mL). The next day, remove the media and add fresh media.
14. Harvest infected MEFs for reprogramming experiments by washing 10 cm dish with 2 mL of PBS (1×), then add 1 mL of pre-warmed 0.25 % trypsin solution, and incubate at room temperature or 37 °C for 2–3 min or until colonies begin to disassociate. Spin MEFs at 1,500 rpm for 3 min and resuspend pellet in ES cell media.

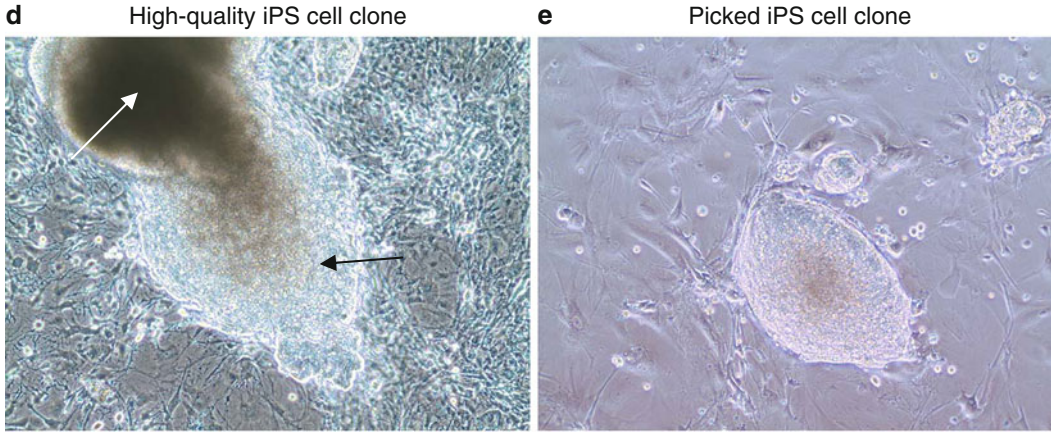
3.4 Reprogramming of Fibroblasts to iPS Cells

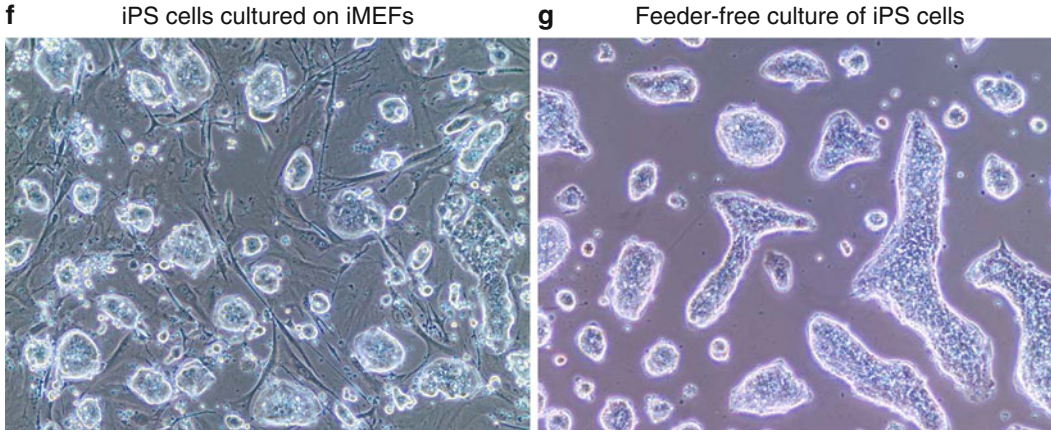
The scheme for reprogramming is shown in Fig. 1a.

1. At least 24 h before transduction, plate MEFs in a 6-well plate at a concentration that will achieve ~60–70 % confluency the following day. MEF cultures that are too sparse may result in elevated cell death following transduction, and cultures that are too dense may not be efficiently transduced.
2. To transduce the MEFs, add polybrene to the Oct4, Sox2, and Klf4 viral supernatant to a concentration of 8 μg/mL. Add 2–4 mL of viral mixture to the MEFs and culture overnight at 37 °C.
3. Twenty-four hours after transduction (Day 0), passage the MEFs 1:2 onto an equally sized gelatin-coated culture dish (e.g., 6-well plate) in 2 mL of ES cell media containing DMEM high glucose, 15 % ES cell-qualified FBS, LIF (10 ng/mL), 1.5–3 μM GSK3i (CHIR99021), penicillin–streptomycin (1×), glutamine (1×), 2-mercaptoethanol (1×), and nonessential amino acids (1×) at 37 °C. Alternatively, transduced MEFs can be plated on iMEFs using the same conditions. Culture of transduced MEFs on iMEFs may enhance reprogramming efficiency due to secretion of self-renewal factors from the feeder layer. Freeze the remaining transduced MEFs in liquid nitrogen.
4. Day 1: Check the viability of infected MEFs using an inverted bright field microscope. Infected MEFs should be morphologically similar to uninfected MEFs.
5. Day 2: Aspirate the old ES cell media and add fresh ES cell media containing 1.5–3 μM GSK3i.
6. Day 4: Aspirate the old ES cell media and add fresh ES cell media containing 1.5–3 μM GSK3i. Media should be changed every 48 h and every 24 h if the cells become dense enough that the medium changes to a yellow color within 24 h after changing the media. Cells should be highly proliferative at


a
Viral transduction
Re-plate infected MEFs in ESC media
Pick iPSC clones
Expand iPSC clones
Characterize iPSC clones
Day: –3
–1
0
20
40
60


b
Transformed non-iPS cell clone
c
iPS cell clone


d
High-quality iPS cell clone
e
Picked iPS cell clone


f
iPS cells cultured on iMEFs
g
Feeder-free culture of iPS cells

this point. Check the reprogramming status using an inverted bright field microscope. Cells undergoing reprogramming will become compact and exhibit an increase in the nuclear to cytoplasm ratio which is a characteristic of ES cells and iPS cells.

7. Day 6: Several colonies resembling reprogrammed cells should become more visible at this stage. Change the media again by aspirating the old ES cell media and adding fresh media. Compact 2D colonies of cells should be visible at this point, along with intermittent transformed colonies (Fig. 1b). Because we excluded c-Myc from the reprogramming mix, the presence of transformed colonies should be minimal. However, transformed colonies that remain flat without a smooth shiny 3D surface after extended culture should be avoided when picking colonies in subsequent steps.
8. Day 7: Prepare an iMEF feeder layer to passage the iPS cell colonies. Add 2 mL of gelatin solution to a 24-well plate and incubate at 37 °C for 10 min. Thaw a vial of frozen MEFs and add cells to 3 mL of pre-warmed MEF media in a 15 mL conical tube and spin at 1,500 rpm for 3–5 min. Aspirate the gelatin from the 24-well plate, resuspend iMEFs in 12 mL of MEF media, and aliquot 0.5 mL to each well.
9. Day 8: Change the ES cell media and monitor the progress of reprogramming. The emergence of 3D colonies that resemble reprogrammed iPS cells should become more apparent during each successive day of reprogramming.
10. Days 10–18: Continue to change the ES cell media every 24–48 h. As iPS cell colonies become visible (Fig. 1c–d), pick individual colonies using a 10 or 200 μL pipette tip or a finely drawn capillary tube. Prior to picking colonies, prepare a 24-well plate with iMEFs. Then aspirate off old MEF media and add ES cell media containing CHIR99021 (1.5–3 μM). Next, pick individual iPS cell colonies and wash with PBS (1×), briefly trypsinize (0.05 % trypsin), and replate in a well

Fig. 1 Derivation of iPS cells with chemical inhibition and without c-Myc. (**a**) Experimental design for derivation of iPS cells from mouse embryonic fibroblasts. (**b**) Bright field microscopy of a representative "transformed" non-iPS cell colony. Transformed colonies can be morphologically excluded from bona fide iPS cell colonies by several criteria including an ill-defined cell border (*arrows*) that seems to resemble "stretched" fibroblasts. This phenomenon seems to correspond with an elastic physical property that is observed when picking colonies. While bona fide iPS cell colonies easily detach from the surrounding cells when picked, transformed cells may demonstrate elastic physical restraint and thus can be avoided when picking colonies. (**c**) High-quality iPS cell colonies should have a well-defined cell border and resemble a 3D shiny colony. (**d**) Representative high-quality iPS cell colony (*black arrow*) that has overgrown and partially differentiated into cells resembling endoderm (*white arrow*). (**e**) Picked iPS cell clone cultured on iMEFs for several days. (**f**) Expanded iPS cell line cultured on iMEFs. (**g**) Feeder-free and serum-free culture of iPS cells on a 6-well plate for two passages. One more passage should be sufficient to remove all contaminating fibroblasts

of a 24-well plate. These iPS cell colonies are at passage 0 at this stage.

11. Twenty-four hours later, check the status of the picked colonies.
12. Two days after picking and reseeding iPS cell colonies in 24-well plates (Fig. 1e), aspirate the old ES cell media and add 0.5 mL fresh ES cell media containing CHIR99021 (1.5–3 μM).
13. After cells become ~70 % confluent, or individual colonies grow excessively large, passage the iPS cell onto a 12-well plate.
14. Continue to pick iPS cell colonies as they emerge, and maintain picked and expanded iPS cell clones by passaging cells onto successively larger culture plates. iPS cell clones can be frozen as they are passaged from a 12-well plate onto a 6-well plate (Fig. 1f) for expression analysis, alkaline phosphatase staining, or other screening measures.

3.5 Feeder-Free and Serum-Free Culture of iPS Cells

iPS cells can be transitioned to feeder-free conditions to minimize contamination of fibroblasts in downstream applications such as expression analysis and epigenomics studies (e.g., ChIP-Seq).

1. Thaw a vial of iPS cells in 3 mL of MEF media (DMEM, high glucose, 10 % FBS, pen–strep, L-glutamine). Centrifuge for 3 min and resuspend in 2 mL of ES cell media. Plate on a gelatin-coated 6-well plate and incubate at 37 °C with 5 % CO_2. Examine the cells on a daily basis and change the media as necessary when the media changes to a yellowish color. Three successive passages of iPS cells at a ratio of 1:6 are sufficient to eliminate feeder cells by diluting the number of non-proliferative iMEFs.
2. To culture iPS cells in feeder-free and serum-free conditions, prepare gelatinized 6-well plates in advance. Aspirate the ES cell media, wash with PBS (1×), and add 1 mL of 0.25 % trypsin and incubate at room temperature for 2–3 min or until cells become detached. Pipette several times with a 1 mL tip and add to 3 mL of MEF media. Centrifuge and resuspend pellet in 2 mL of a 4:1 ratio of ES cell media to ESGRO complete media containing LIF, BMP4, and CHIR99021. These conditions have been adapted from a previously published protocol [10] and commercially optimized. Alternatively, iPS cells can be cultured under the original conditions including N2B27 media, LIF, and BMP4 [10].
3. Examine the cells daily and passage the cells as they become ~70 % confluent. The goal is to transition the iPS cells from serum-containing media to feeder-free conditions over several passages. For the second passage, prepare a gelatinized 6-well plate in advance. Aspirate the ES cell/ESGRO complete media, wash with PBS (1×), and add 1 mL of 0.25 % trypsin and incubate at room temperature for 2–3 min. Pipette, centrifuge,

and resuspend in 2 mL of a 3:2 ratio of ES cell media to ESGRO complete media. After the second passage, few feeder cells should be visible (Fig. 1g). Repeat the above conditions for the third passage, except resuspend the pellet in 2 mL of a 2:3 ratio of ES cell media to ESGRO complete media. For the fourth passage, resuspend the pellet in 2 mL of a 1:4 ratio of ES cell media to ESGRO complete media. For the fifth passage, iPS cells should be fully transitioned to feeder-free and serum-free ESGRO complete ES cell media.

4 Notes

1. MEFs can also be obtained from various commercial vendors or prepared from E13.5 to E14.5 MEFs. Protocols for harvesting MEFs can be readily found in the literature.
2. Established and characterized ES cell lines can be obtained from academic labs or commercial sources. Alternatively, ES cells can be derived by plating E3.5 mouse blastocysts on iMEFs in ES cell derivation media (DMEM high glucose, 20 % ESC-qualified FBS, LIF, pen–strep, nonessential amino acids, 1× 2-mercaptoethanol, L-glutamine) and incubating at 37 °C with 5 % CO_2. After the blastocyst adheres to the layer of MEFs and the trophectoderm layer spreads out, the inner cell mass (ICM) outgrowth can be picked, dissociated in trypsin briefly, and replated in ES cell derivation media.
3. The GSK3β inhibitor, CHIR99021, can be obtained from several commercial vendors. However, because the activity of small molecule inhibitors may vary between vendors, it is important to test the activity of small molecule inhibitors acquired from various sources.

Acknowledgment

The authors would like to thank Allegra Geller for helpful discussions.

References

1. Okita K, Ichisaka T, Yamanaka S (2007) Generation of germline-competent induced pluripotent stem cells. Nature 448(7151):313–317
2. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature 448(7151):318–324
3. Kidder BL, Hu G, Yu ZX, Liu C, Zhao K (2013) Extended self-renewal and accelerated reprogramming in the absence of Kdm5b. Mol Cell Biol 33:4793–4810. doi:10.1128/MCB.00692-13, MCB.00692-13 [pii]
4. Silva J, Barrandon O, Nichols J, Kawaguchi J, Theunissen TW, Smith A (2008) Promotion of reprogramming to ground state pluripotency

by signal inhibition. PLoS Biol 6(10):e253. doi:10.1371/journal.pbio.0060253, 08-PLBI-RA-2877 [pii]

5. Marson A, Foreman R, Chevalier B, Bilodeau S, Kahn M, Young RA, Jaenisch R (2008) Wnt signaling promotes reprogramming of somatic cells to pluripotency. Cell Stem Cell 3(2):132–135. doi:10.1016/j.stem.2008.06.019, S1934-5909 (08)00334-2 [pii]
6. Kidder BL, Yang J, Palmer S (2008) Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells. PLoS ONE 3(12):e3932. doi:10.1371/journal.pone.0003932
7. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133(6): 1106–1117
8. Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S (2008) Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat Biotechnol 26(1):101–106. doi:10.1038/nbt1374, nbt1374 [pii]
9. Kidder BL, Oseth L, Miller S, Hirsch B, Verfaillie C, Coucouvanis E (2008) Embryonic stem cells contribute to mouse chimeras in the absence of detectable cell fusion. Cloning Stem Cells 10(2):231–248
10. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A (2008) The ground state of embryonic stem cell self-renewal. Nature 453(7194): 519–523. doi:10.1038/nature06968, nature 06968 [pii]

Chapter 16

Transdifferentiation of Mouse Fibroblasts and Hepatocytes to Functional Neurons

Samuele Marro and Nan Yang

Abstract

Nuclear reprogramming by defined transcription factors became of broad interest in 2006 with the work of Takahashi and Yamanaka (Cell 126:663–676, 2006), but the first example of cell fate reshaping via ectopic expression of transcription factor was provided back in 1987 when Davis and colleagues induced features of a muscle cell in fibroblast using the muscle transcription factor MyoD (Davis et al., Cell 51:987–1000, 1987). In 2010 our laboratory described how forced expression of the three neuronal transcription factors Ascl1, Brn2, and Myt1l rapidly converts mouse fibroblasts into neuronal cells that exhibit biochemical and electrophysiological properties of neurons. We named these cells induced neuronal cells (iN cells) (Vierbuchen et al., Nature 463:1035–1041, 2010; Vierbuchen and Wernig, Nat Biotechnol 29:892–907, 2011). Interestingly, iN cells can also be derived from defined endodermal cells such as primary hepatocytes, suggesting the existence of a more general reprogramming paradigm (Marro et al., Cell Stem Cell 9:374–382, 2011). In this chapter we describe the detailed methods used to attain the direct conversion.

Key words Nuclear reprogramming, Fibroblast, Hepatocyte, Induced neuronal cells

1 Introduction

In normal development, cells from the inner cell mass give rise to all the cell lineages composing an adult organism. During this process cells progressively acquire functional identities by establishing epigenetic landmarks and are generally stable in their state throughout the life of the organism. However, as early as the mid-1960s, cell transplantation experiments suggested that the state of specialized cells was plastic and could be altered in response to the extracellular environment [6, 7]. Change of cell identity by nuclear reprogramming has been achieved by three distinct experimental approaches: somatic cell nuclear transfer, cell fusion, and transcription-factor transduction. While the first two approaches require a host cell adding technical and ethical issues to the method, the third approach has become particularly attractive for its simplicity.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_16, © Springer Science+Business Media New York 2014

We have recently developed approaches to directly convert fibroblasts into induced neuronal (iN) cells, which indicate that direct lineage conversions are possible between very distantly related cell types [3]. In mouse cells this is realized by transduction of the fibroblastic cells with lentivirus encoding the three transcription factors Ascl1, Brn2, and Myt1l (namely, BAM factors). The same reprogramming cocktail can also convert definitive endodermal cells—hepatocytes—into iNs (Hep-iN cells) [5]. Gene profiling data indicated that albeit the Hep-iN cells required more time to induce a neuronal transcriptional program than fibroblast-iN cells, BAM factors can induce silencing of both fibroblast and hepatocyte specific transcriptional program and obtaining neuronal features.

2 Materials

The following specific items are required for this protocol. General tissue culture equipment is not listed.

2.1 Mouse Hepatocyte Derivation Components

1. Perfusion Buffer 1 (PB1): To make 500 mL use 499 mL Krebs-Ringer Bicarbonate Buffer (KRB) (Sigma) and 1 mL 0.5 M EDTA (Lonza). Sterile filter with 0.22 μm bottle top filter into sterile media bottle. Store at 4 °C for up to 1 month.
2. Perfusion Buffer 2 (PB2): To make 500 mL use 499 mL KRB and 0.5 mL 150 mM $CaCl_2$ (Lonza). Sterile filter with 0.22 μm bottle top filter into sterile media bottle and store at 4 °C. Right before use add 20 mg Collagenase 1 (Sigma) to 40 mL of PB2.
3. Acetic acid solution: To make 500 mL: 2.25 mL acetic acid (Sigma, A6283) in distilled water.
4. Polybrene: To make 5 mL of 1,000× stock solution dissolve 40 mg of hexadimethrine bromide (Sigma) in distilled water, filter and store at 4 °C.

2.2 Mouse Glial Cell Isolation Components

1. Dissociation solution: 5 mL HBSS, 80 μL Papain solution (Worthington), 5 μL of 0.5 M EDTA, and 5 μL of 1 mM $CaCl_2$.
2. DNase solution: 10 mg/mL in sterile water, adjust pH to 7.3, filter with 0.22 μM syringe filter, aliquot, and store at −20 °C.

2.3 Cell Culture Media

1. MEF media: To make 500 mL: Add 50 mL calf serum (Thermo Scientific), 5 mL 100× Pen/Strep (Invitrogen), 5 mL 100× sodium pyruvate (Invitrogen), 5 mL 100× nonessential amino acids (NEAA) (Invitrogen), 4 μL of 2-mercaptoethanol (Sigma) into 435 mL DMEM (Invitrogen).

2. N3 media: To make 500 mL: 490 mL DMEM/F12 (Invitrogen), 5 mL 100× Pen/Strep, 12.5 mg insulin (Sigma), 25 mg apo-transferrin (Sigma), 3.2 μg progesterone (Sigma), 8 mg putrescine (Sigma), and 259.5 μg sodium selenite (Sigma) (*see* **Note 1**).
3. Hepatocyte growth media: To make 500 mL: 145 mL DMEM, 355 mL low glucose DMEM (Invitrogen), 1 g BSA (Sigma), 1 g galactose (Sigma), 50 mg ornithine (Sigma), 15 mg proline (Sigma), 305 mg nicotinamide (Sigma), 12 μg $ZnCl_2$ (Sigma), 375 mg $ZnSO4{:}7H_2O$ (Sigma), 100 μg $CuSO4{:}5H_2O$ (Sigma), 12 μg MnSO4, 5 mM glutamine, 2.5 mg insulin, 2.5 mg apo-transferrin, 2.5 μg sodium selenite, 0.1 μM dexamethasone (Sigma), 5 mL 100× Pen/Strep, and 40 ng/mL hepatocyte growth factor (HGF) (PeproTech). Right before use add 20 mg/mL epidermal growth factor (EGF) (R&D systems).
4. Hepatocyte plating media: Add 10 % calf serum to hepatocyte growth media.
5. 2× freezing media: To make 100 mL: 60 mL MEF media, 20 mL calf serum, 20 mL DMSO. Dilute to 1× with MEF media.
6. Doxycycline: To make 5 mL of 1,000× stock solution: Dissolve 10 mg of doxycycline (Sigma) in distilled water, filter, and store protecting from light at 4 °C.
7. 0.25 % trypsin: Dilute 2.5 % trypsin stock with HBSS–EDTA for 0.25 % stock solution. Store aliquots at −20 °C.

2.4 Lentivirus Production Components

1. DNA plasmids: Tet-O-FUW-Ascl1 (Addgene, plasmid 27150), Tet-O-FUW-Brn2 (Addgene, plasmid 27151), Tet-O-FUW-Myt1l (Addgene, plasmid 27152), Tet-O-FUW-EGFP (Addgene, plasmid 30130), FUW-M2rtTA (Addgene, plasmid 20342), pRSV-rev (Addgene, plasmid 12253), pMDLg/pRRE (Addgene, plasmid 12251), and pMD2.G (Addgene, plasmid 12259).
2. HEK293T/17 cells: (ATCC, CRL-11268).
3. 2× BBS: To make 500 mL: Add 8.18 g NaCl, 5.32 g BES (Sigma), 0.10 Na_2HPO_4, into distilled water, set pH to 6.95 with 1 M NaOH. Validate before use (*see* **Notes 2** and **3**).
4. 2.5 M $CaCl_2$: To make 100 mL: Add 27.74 g $CaCl_2$ in distilled H_2O. Filter with 0.22 μM filter and store at −20 °C.
5. Polyornithine: To make 100× stock, dilute 100 mg in 67 mL water. Filter with 0.22 μM filter and store at −20 °C in 5 mL aliquots. Dilute 100× stock in sterile water and filter into sterile bottle to make the 1× working solution. Store at 4 °C up to 3 months.

2.5 Immunofluorescence Components

1. Antibodies: Albumin (Bethyl), Tuj1 (Covance), MAP2 (Sigma), synapsin (Synaptic Systems).
2. Blocking buffer: To make 20 mL: Dissolve 0.8 g bovine serum albumin and 0.2 mL calf serum in D-PBS.

2.6 Mice Strains

Tau::GFP, Alb::Cre, loxSTOPlox tdTomato (Jackson's lab) (*see* **Notes 4** and **5**).

3 Methods

3.1 Mouse Fibroblast Derivation

1. Remove E13.5-E14.5 embryos from uterine horns and place in a 10 cm tissue culture dish filled with HBSS. Wash twice with 20 mL HBSS and place in a fresh 10 cm tissue culture dish.
2. Under a dissection microscope remove the limbs of the embryo with surgical scissors and pincers. Be careful to not include any tissue above the shoulder and hip joints. Place limbs in a few drops of HBSS in a 15 cm tissue culture dish.
3. Place extremities from 3 to 4 mice on a 15 cm tissue culture dish. Using curved scissors, thoroughly mince until homogenous (approximately 2–3 min).
4. Add 1 mL of 0.25 % trypsin solution and incubate for 15 min at 37 °C. Briefly triturate dissociated tissue using a 10 mL pipette. Add another 20 mL MEF media and place cells in a 37 °C incubator. Typically 2–4 days later, the cells will become confluent.
5. When confluent each dish is frozen in 3 cryovials using freezing media and stored in liquid nitrogen. The best reprogramming efficiency is achieved using MEF in between passage 3 and 4.

3.2 Mouse Hepatocytes Derivation

The following protocol is a modification of Seglen's (Preparation of isolated rat liver cells, P.O. Seglen, Methods Cell Biol, 1976) two-step collagenase perfusion method to isolate and purify rat hepatocytes from non-parenchymal cells for the mouse. Hepatocytes are not the only liver population purified with this protocol; presence of fibroblast-like cells has been reported in liver neonatal primary cultures [8]. Addition of hormones and growth factors to the culture medium reduces the number of non-epithelial cells in these cultures, but nevertheless, immunofluorescence with anti-albumin antibodies should be performed to test the purity of the cultures. Additionally, liver preparations can be done from reporter mice such as Alb::Cre crossed with a loxSTOPlox tdTomato Cre reporter as shown in the Fig. 1.

1. Prepare one 10 cm tissue culture dish by coating the surface with 0.3 mg/mL collagen (BD) in acetic acid solution for 1 h in 37 °C incubator.

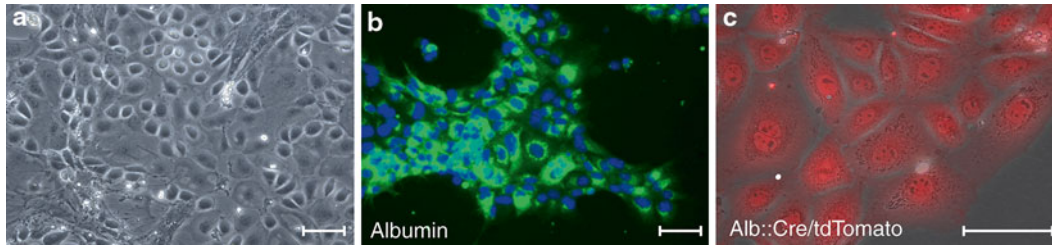


Fig. 1 (**a**) Phase contrast image of liver cultures derived from Alb::Cre loxSTOPlox tdTomato mice 5 days after isolation. (**b**) Albumin expression is detected by immunostaining using an antibody against albumin (*green*). (**c**) Hepatocyte lineage tracing using tdTomato (*red*) live imaging. Scale bars: 50 μm

2. Euthanize one postnatal day 2 pup by decapitation. Make incision through the skin on the ventral midline and remove the liver using curved tweezers. Place the liver in one 10 cm plate containing 5 mL PB1 and wash for 3 min at room temperature.
3. Carefully remove PB1 (do not let the tissue dry) and repeat the wash twice. Mince with scissors very gently for about 1 min or until small pieces are obtained (≤0.5 mm).
4. Clump all of the minced parts in the center of the dish and add 3 mL pre-warmed PB2. Place in 37 °C incubator for 20 min.
5. Add 2 mL of PB2 and gently pipet for several times and incubate for another 20 min.
6. Add 8 mL of ice cold KRB and produce a single cell suspension by pipetting 5–10 times and filter the suspension through a 70 μm cell strainer into a 50 mL conical tube. Collect remaining cells with an extra 10 mL KRB and repeat suspension and filtration. After collecting all cells, fill the tube with cold KRB up to 30 mL and centrifuge at 140 × *g* for 5 min at 4 °C.
7. Remove the supernatant and add 30 mL ice cold KRB to wash the pellet and centrifuge again. Repeat wash two times.
8. Wash collagen plates with abundant KRB twice. Resuspend cells in 6 mL of pre-warmed hepatocyte plating media and plate the cells. After 4 h, change media to hepatocyte growth media. The tissue culture dish should contain approximately 4×10^6 cells at 90 % confluence within 4–5 days. Hepatocytes at passage 2 can be split up to three times by incubating at 37 °C in 0.25 % trypsin solution. Collagen-coated dishes and seeding density of 30,000 cells/cm^2 should be used.

3.3 Glia Cell Isolation

1. Anesthetize postnatal day 3–5 pups on ice. Remove heads from pups with surgical scissors and place in a 10 cm tissue culture dish (2–3 brains for each 10 cm tissue culture dish). To prepare a significant amount of cells, at least three pups are required.

2. Remove mouse heads one at a time and place in a 15 cm dish to remove brain from skull. Then put the brain in a 3 cm dish filled with cold HBSS, dissect the cortices from the brain, and separate the two hemispheres. Remove the meninges using fine tweezers and place the hemisphere into a 15 mL Falcon tube filled with cold HBSS and put on ice (*see* **Note 6**).
3. After collecting all the hemispheres, remove the HBSS and add 5 ml dissociation solution (*see* **Note 7**). Put the tube in the 37 °C incubator for 15 min and shake the tube every 5 min.
4. Remove the dissociation solution with caution and wash the tissue twice with MEF media.
5. Add 1 mL MEF media and use a pipette to triturate the tissue. Add another 4 mL MEF media and transfer through a 40 μm cell strainer into the 50 mL Falcon tube with 5 mL MEF media. Plate onto a 10 cm plate.
6. Change the media on the next day. Massive cell death will be observed. Passage the cells in MEF media when confluent. Typically 4–6 days later, the cells will become confluent and can be split into three 10 cm dishes using MEF media (*see* **Note 8**).

3.4 Lentivirus Production

The reprogramming factors are delivered by a doxycycline-inducible lentiviral expression system. This expression system employs the recombinant tetracycline trans-activator (rtTA) expressed via an FUW lentiviral vector [9]. Thus, for proper expression of the viral transgenes, the recipient cells need to be infected with both the Tet-O-FUW lentiviral particles containing the reprogramming factors and FUW-M2rtTA and cultured in media containing doxycycline. Lentiviruses are produced by transfecting HEK/293T cells with the calcium-phosphate precipitation method [10].

1. Coat 10 cm tissue culture dishes with polyornithine by adding 5 mL of 1× polyornithine solution and incubating for 1 h in 37 °C incubator. Remove polyornithine solution and wash twice with PBS.
2. Seed 5×10^6 HEK/293T cells per tissue culture dish in MEF media (*see* **Note 9**).
3. 24 h after plating, remove the media and add 9 mL fresh MEF media.
4. Prepare five transfection mixtures (one per lentivirus: Ascl1, Brn2, Myt1l, EGFP, M2rtTA) by mixing in a total volume of 500 μL distilled water, 10 μg of lentiviral plasmid, 2.5 μg pRSV-rev, 2.5 μg pMD2.G, and 5 μg pMDLg/pRRE.
5. Add 40–120 μL of 2.5 M $CaCl_2$ while vortexing.
6. Add 500 μL of 2× BBS to the transfection mixture drop by drop while vortexing and incubate for 10 min at room temperature.

7. Lightly mix solution by pipetting up and down and add 1 mL dropwise to each plate of HEK/293T cells.
8. 16 h later replace media with 5 mL of MEF media (*see* **Note 10**). Check EGFP fluorescence of Tet-O-EGFP plate to ensure that the transfection worked properly.
9. Harvest viral supernatant 24 h later (40–44 h after transfection), filter through 0.45 μm cellulose acetate filter, and centrifuge at 50,000 × *g* for 1.5 h at 4 °C to concentrate the viral particles.
10. Reconstitute the pellet in 50 μL DMEM (to obtain a 100× concentrated virus stock) and store at 4 °C (*see* **Note 11**).

3.5 Transduction

Titration of each reprogramming factor will require 6 wells of a 12-well dish. Each reprogramming factor must be titrated together with FUW-RtTA virus and the infection efficiency should be measured 48 h after the addition of doxycycline using immunofluorescence with appropriate antibodies for each factor. An approximate starting point for MEFs would be to perform a dilution series across the 6 wells using a range of 0.5–10 μL/virus in 1 mL of polybrene-containing media/well. Fix and stain cells 48 h after the addition of doxycycline using standard procedures. Determine infection efficiency for each reprogramming factor by estimating the fraction of infected cells and use DAPI staining to mark all nuclei. Optimal reprogramming efficiencies will be achieved when approximately 60–80 % of cells express Ascl1, Brn2, and 30–40 % express Myt1l. For hepatocyte transduction, a range of 5–50 μL/virus in 1 mL of media/well should be tested (*see* **Note 12**).

1. Seed 300,000 hepatocytes or 250,000 MEF suspended in a volume of 1.5 mL of media per well of a 6-well plate and incubate 24 h at 37 °C incubator.
2. Change media to 1 mL of fresh media containing polybrene (add 1 μL of the 1,000× stock per well) and appropriate amount of Ascl1, Brn2, Myt1L, and FUW-RtTA virus and incubate 16–18 h in 37 °C incubator. EGFP virus can be included to help visualize the iN cell morphology.
3. Change media to doxycycline-containing media by diluting doxycycline solution to final concentration of 1× in MEF or hepatocyte media to activate the transcription.
4. 24 h later switch both hepatocyte and MEF cultures to doxycycline-containing basic neuronal media (N3 media) and change half of the media every 2–3 days.

3.6 iN Cell Maturation

1. Detach glia using 0.25 % trypsin, spin down for 3 min at 200 × *g*, remove supernatant, and resuspend in N3 media. Count cells using a hemocytometer.

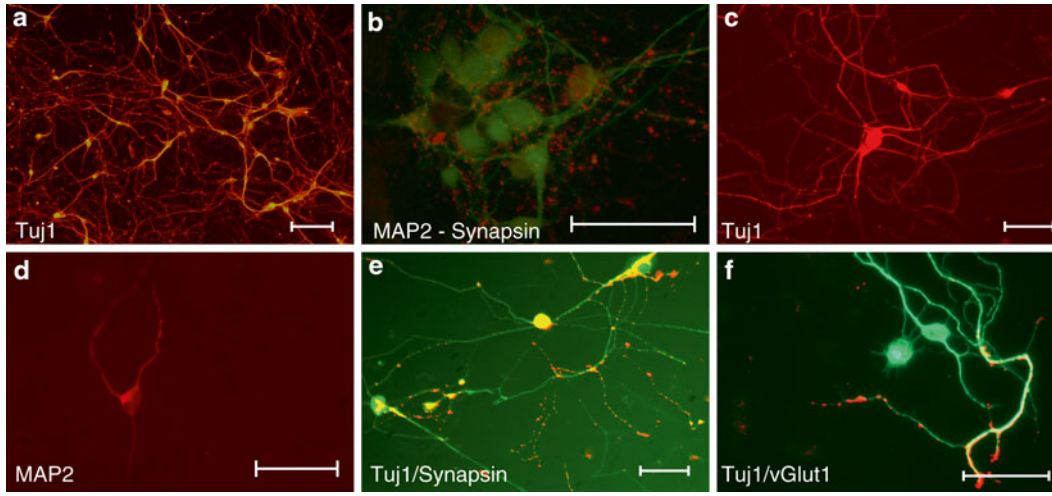


Fig. 2 (**a**) MEF-iN 13 days after infection are detected by staining with Tuj1 antibody (*red*). (**b**) 22 days after infection MEF-iN cells express MAP2 (*green*) and synapsin (*red*). (**c**) Hep-iN 13 days after infection are positive for Tu1j (*red*). (**d**) Hep-iN 22 days after infection express MAP2 (*red*) as well as synapsin (*red*) (**e**) and vGlut1 (*red*) (**f**). Scale bars: 200 μm (**c**) and 50 μm (**b**–**f**)

2. Plate 60–150,000 glia on iN cultures 3–7 days after addition of doxycycline to the infected cells. The appropriate number of glia to add will depend on the density of cells (*see* **Note 13**).
3. Change 50 % of the media every 3 days with extra attention toward gentle pipetting to avoid the dethatching of iN cells from the plastic surface. iN cells with immature neuronal morphology should be visible within 7 days after doxycycline induction.

3.7 Basic Characterization of iN Cells Using Immunofluorescence

iN cells express pan neuronal markers such as Tuj1 and MAP2 as early as 2 weeks after infection, but extra 1–2 weeks are needed to show mature characteristics such as firing of action potentials, synaptic activity, and expression of mature markers such as synapsin and vGlut1 (Fig. 2).

The overall quality of the conversion can be easily assessed by immune fluorescence analysis to confirm expression of neuronal markers. More sophisticated characterization requires electrophysiology techniques, analysis of transcriptional profile by microarray, as well quantitative PCR [5]. Here, we describe a simple procedure for immunofluorescence.

1. Wash the plate twice with D-PBS. Add 4 % PFA/D-PBS and incubate for 10 min at room temperature to fix.
2. Wash the plate three times with D-PBS.
3. Add 0.2 % Triton X-100/D-PBS and incubate for 5 min at room temperature.
4. Wash the plate three times with D-PBS.

5. Add blocking buffer and incubate for 30 min at room temperature.
6. Remove the blocking buffer and add primary antibody diluted in blocking buffer to appropriate concentration and incubate for 30 min at room temperature.
7. Remove the primary antibody solution and wash the plate three times with D-PBS.
8. Add secondary antibody diluted in blocking buffer to appropriate concentration and incubate for 30 min at room temperature.
9. Remove the secondary antibody solution and wash the plate three times with D-PBS. For the first wash using D-PBS containing DAPI to stain the nuclei.

4 Notes

1. Components for N3 media and hepatocyte growth media should be stocked at −80 °C.
2. The pH value of 2× BBS is critical.
3. For each batch of 2× BBS solution, use Tet-O-EGFP DNA with lentiviral packaging plasmids as test DNA to optimize transfection. Make identical transfection reactions and test 2.5 M $CaCl_2$ amounts from 40 to 120 μL in increments of 10 μL. 16 h later, check EGFP levels of each plate using fluorescence microscope. In order to efficiently make lentivirus, transfection efficiency should be at least 70 %. Changes in the amount of total DNA, the tissue culture plate size, the amount of media on the cells, cell density, or the timing of the media change before the transfection can dramatically alter results.
4. All procedures with live animals should adhere to relevant institutional guidelines for animal welfare.
5. Timed pregnant mice can be ordered from commercial suppliers (Charles River, Taconic, etc.) or timed mating can be set up in-house.
6. When making glial cell culture, it is critical to remove the meninges thoroughly; otherwise, fibroblasts can outgrow glia.
7. It is critical that the dissection solution is pre-warmed until the solution is clear. DNase in the dissociation solution is to reduce the stickiness of the mix.
8. Glial cells can be cryopreserved at passage 1.
9. High-passage HEK/293T cells can result in low viral titer regardless of the transfection efficiency.
10. At this point transfection efficiency can be measured my imaging EGFP expression using a fluorescence microscope.

In order to efficiently make lentivirus, transfection efficiency should be at least 70 %.

11. The concentrated lentivirus preparation is stable at 4 °C for up to 10 days.
12. High infection rates (70–85 %) are critical to generate iN cells efficiently, and efficient transduction of hepatocytes requires a MOI (multiplicity of infection) approximately ten times higher than the MOI required to transduce fibroblasts. EGFP virus should be included in every experiment as a proxy for measuring the lentiviral titer. The EGFP virus is also useful for visualizing the axons and dendrites extending from the iN cells, which can be difficult to visualize under phase contrast among fibroblasts.
13. Glial cells are required to promote the functional maturation of iN cells.

References

1. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–676
2. Davis RL, Weintraub H, Lassar AB (1987) Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell 51: 987–1000
3. Vierbuchen T et al (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463:1035–1041
4. Vierbuchen T, Wernig M (2011) Direct lineage conversions: unnatural but useful? Nat Biotechnol 29:892–907
5. Marro S et al (2011) Direct lineage conversion of terminally differentiated hepatocytes to functional neurons. Cell Stem Cell 9:374–382
6. Gehring W (1967) Clonal analysis of determination dynamics in cultures of imaginal disks in Drosophila melanogaster. Dev Biol 16:438–456
7. Hadorn E (1966) Constancy, variation and type of determination and differentiation in cells from male genitalia rudiments of Drosophila melanogaster in permanent culture in vivo. Dev Biol 13:424–509
8. Pagan R et al (1995) Epithelial-mesenchymal transition in cultured neonatal hepatocytes. Hepatology 21:820–831
9. Gossen M et al (1995) Transcriptional activation by tetracyclines in mammalian cells. Science 268:1766–1769
10. Tiscornia G, Singer O, Verma IM (2006) Production and purification of lentiviral vectors. Nat Protoc 1:241–245

Chapter 17

Direct Lineage Conversion of Pancreatic Exocrine to Endocrine Beta Cells In Vivo with Defined Factors

Claudia Cavelti-Weder, Weida Li, Gordon C. Weir, and Qiao Zhou

Abstract

Pancreatic exocrine cells can be directly converted to insulin$^+$ beta cells by adenoviral-mediated expression of three transcription factors Pdx1, Mafa, and Ngn3 in the adult mouse pancreas (Zhou et al., Nature 455(7213):627–632, 2008). This direct reprogramming approach offers a strategy to replenish beta-cell mass and may be further developed as a potential future treatment for diabetes. Here, we provide a detailed protocol for inducing exocrine to beta-cell reprogramming in mice. We also describe key analyses we routinely use to assess the phenotype and function of reprogrammed cells.

Key words Pancreatic exocrine to beta-cell reprogramming, Transcription factors, Polycistronic gene expression, Adenovirus production and injection, Diabetes induction in mice, Blood glucose and insulin monitoring

1 Introduction

Diabetes is a growing health problem that affects hundreds of millions of people worldwide [1]. Type I diabetes results from autoimmune destruction of insulin-secreting beta cells of the pancreas, whereas type 2 diabetes is associated with insulin resistance by peripheral tissues [2]. A hallmark of both type I diabetes and advanced type 2 diabetes is a decrease in beta-cell mass [3, 4]. Therefore, there is great interest in developing strategies to regenerate beta cells and replenish beta-cell mass. To date, several approaches have been described, including promoting replication of existing beta cells [5–7], stimulating neogenesis of new beta cells from precursor cells within the pancreas [8–12], and direct reprogramming of non-beta cells to beta cells [13–17].

Cellular reprogramming is a process in which one cell type is converted to another cell type [18]. A reprogramming approach has been developed to directly convert pancreatic exocrine cells to

Claudia Cavelti-Weder and Weida Li have equally contributed to this work.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_17, © Springer Science+Business Media New York 2014

insulin+ beta cells by intrapancreatic injection of adenoviral vectors expressing three transcription factors Pdx1, Ngn3, and Mafa [13]. In the original work, injection of three separate viruses each carrying a single factor was used to induce reprogramming [13]. We present here an improved method where all three factors are simultaneously expressed from a polycistronic construct mediated by 2A peptides. The 2A peptides mediate translational "skipping," which allows generation of multiple proteins from a single transcript [19]. With the new polycistronic viral construct, robust reprogramming results can be routinely achieved.

We describe the construction of the polycistronic construct, adenoviral production and titration, and the surgical procedure for intrapancreatic viral injection. In addition, we outline how to confirm successful reprogramming by immunohistochemistry and a rigorous functional test. Briefly, the functional test is carried out as the following: Rag−/− mice are treated with streptozotocin to eliminate the majority of endogenous beta cells within the islets. Due to extreme hyperglycemia, the health condition of these animals declines rapidly. Isolated mouse islets are transplanted under the right kidney capsule to normalize blood glucose levels. Next, purified adenovirus (2×10^9 pfu) carrying reprogramming factors or control Cherry is injected directly into the pancreas. The islet graft is removed by nephrectomy 3 weeks after viral induction. The blood glucose level is now under the direct control of induced insulin+ cells residing in the pancreas. Detailed physiological properties of the induced beta cells are evaluated by glucose tolerance tests and fasting tests.

The protocol described here offers guidance to achieve successful and consistent beta-cell reprogramming. This simple model serves as a useful tool to understand the remarkable phenomenon of cellular reprogramming. This model could also serve as a springboard to further develop the reprogramming approach towards a possible future treatment strategy for diabetes.

2 Materials

2.1 Making Polycistronic Construct

1. Gateway pENTR vector (Invitrogen, San Diego, CA).
2. Polycistronic sequence (oligos synthesized by IDT, Coralville, IA)

   TCGACACTAGTGCCACGAACTTCTCTCTGTTAAA
   GCAAGCAGGAGATGTTGAAGAAAACCCCGGGCC
   TGGATCCGAGGGCAGAGGAAGTCTTCTAACA
   TGCGGTGACGTGGAGGAGAATCCCGGCCCT
   ATCGATCAGTGTACTAATTATGCTCTCTTGAAA
   TTGGCTG GAGATGTTGAGAGCAACCCAGGTCCCGC.
3. pAd/CMV/V5-DEST™ Gateway Vector Kit (Invitrogen).

2.2 Viral Production and Purification

1. Viral Power Adenoviral Expression System (Invitrogen).
2. Vivapure AdenoPACK 100 (Sartorius Stedim, Bohemia, NY).
3. Opti-MEM I medium (Gibco, Grand Island, NY).
4. Lipofectamine 2000 (Invitrogen).
5. 293a cell complete culture medium: DMEM with 10 % FBS, 2 mM L-glutamine, and 1 % penicillin/streptomycin (Sigma).

2.3 Surgical Procedure

1. Animal preparation:
   - Animals to be used: immunocompromised animals such as Rag1−/− mice.
   - 70 % EtOH.
   - Anesthesia for survival surgery (*see* **Note 1**).
   - 18 G and 27 G needles (BD, Franklin Lakes, NJ), 1 ml syringes for injection.
   - Shaver.
   - Alcohol preps (Kendall, Mansfield, MA) and Betadine solution (Santa Cruz, Dallas, TX).
2. Virus injection:
   - Syringes (for virus injection 3/10 cc insulin syringes (BD), for anesthesia 1 ml syringe) and needles: 27 G+18 G (BD).
   - Warm pads.
   - Dissecting microscope Leica stereo, zoom 7 (Leica, Germany).
   - Sterile drapes (IMCO, Daytona Beach, FL).
   - Surgery tools (Stapler/staples/small scissors and forceps), autoclaved.
   - Suture (5-0 Chromic gut) (Butler Schein, Dublin, OH).
   - Bead sterilizer (Fine Science Tool, Foster City, CA).
   - Betadine solution (Santa Cruz).
3. After procedure:
   - Heating lamp.
   - Banamine (Merck, Whitehouse Station, NJ).

2.4 Standard Methods of Analysis

2.4.1 Phenotype Analysis: Immunohistochemistry (IHC)

1. Tissue collection, fixation, and sectioning:
   - Anesthesia for terminal procedures: ketamine (Putney, Portland, ME, 100 mg/ml), xylazine (Lloyd, Shenandoah, IA, 100 mg/ml), saline 0.9 % (Hospira, Lake Forest, IL) (*see* **Note 2**).
   - Surgical tools.
   - Dissecting microscope.

- PBS (Lonza, Walkersville, MD).
- Fluorescence microscope.
- 4 % paraformaldehyde (Sigma).
- 30 % sucrose solution (Sigma).
- O.C.T. Compound (Tissue-Tek, Torrance, CA).

2. Immunohistochemistry:
   - PBS (*see* **Note 3**).
   - Triton X-100 10 % (Sigma).
   - Normal donkey serum (NDS) (Jackson ImmunoResearch, West Grove, PA).
   - The following primary antibodies are used to assess beta-cell phenotype: guinea pig anti-insulin (1:200; Invitrogen); rabbit anti-GLUT2 (1:200; Santa Cruz); rabbit anti-Pax6 (1:300; Millipore, Billerica, MA); mouse anti-Nkx6.1 (1:50; Hybridoma Bank, Iowa City, IO) followed by biotinylated anti-mouse (1:200; Jackson, West Grove, PA).
   - To show absence of endocrine non-beta-cell markers, the following antibodies are used: rabbit anti-glucagon (1:3,000; Millipore), rabbit anti-pancreatic polypeptide (1:200; Millipore), and rabbit anti-somatostatin (1:200; Millipore).
   - Secondary antibodies: Alexa Fluor 488 anti-guinea pig (1:200)/anti-rabbit (1:200); Alexa Fluor 647 anti-guinea pig (1:200)/anti-rabbit (1:200; all from Jackson); SA-HRP antibody (1:100; Perkin Elmer, Waltham, MA); and tyramide solution (1:50; Perkin Elmer).
   - Vectashield mounting media with DAPI (Vector Labs, Burlingame, CT).

2.4.2 Functional Analysis: In Vivo Testing

1. Administration of streptozotocin (STZ):
   - Calculator.
   - Citric acid buffer (*see* **Note 4**).
   - Streptozotocin (STZ) (Sigma).
2. Islet isolation: *See* ref. 20 for detailed protocol.
3. Making islet aliquots for transplantation:
   - RPMI 1640 with 10 % newborn calf serum and 1 % penicillin–streptomycin (all from Cellgro, Herndon, VA).
   - Eppendorf tubes, sterile, 1.7 ml (Denville, South Plainfield, NJ).
   - Bucket of ice.
4. Islet preparation before transplantation:
   - 15 ml Falcon tube for each Hamilton syringe.
   - Hamilton syringe (Fisher).

- 200 μl tip, previously cut and sterilized (*see* **Note 5**).
- Eppendorf tubes containing the islets on ice.
- Transplant tubing (*see* **Note 6**).
- 2 cm long tube (VWR, Tubing Tygon 1/16 × 1/8 ft 1/32).
- 1 ml syringe, tip cut off (*see* **Note 7**).
- Centrifuge (*see* **Note 8**).
- 1 ml tips and pipette.
- Measuring tape.

5. Islet transplantation:
   - Anesthesia solution for survival surgery (*see* **Note 1**).
   - Shaver.
   - Gauze.
   - Warming pads and Delta Phase operating board (Braintree Scientific, Braintree, MA).
   - 70 % alcohol.
   - Transplant tubing containing the islet pellet.
   - Surgical tools (especially bulldog clamp and fine forceps).
   - 23 G needle.
   - 0.9 % Saline.
   - Cautery.
   - Suture (Look Suture, Angiotech, Reading, PA).
   - Staples (BD, Sparks, MD).
   - Betadine (Purdue Frederick CO, Stamford, CT).
6. Virus injection: *See* under Subheading 2.3.
7. Nephrectomy:
   - Anesthesia solution for survival surgery (*see* **Note 1**).
   - Gauze.
   - Warming pads and operating board.
   - 70 % alcohol.
   - Surgical tools (especially bulldog clamp).
   - Suture (Look Suture, Angiotech, Reading, PA).
   - Staples (BD, Sparks, MD).
   - Betadine (Purdue Frederick CO, Stamford, CT).
8. Glucose monitoring:
   - Glucometer (One Touch Ultra, LifeScan Inc., Milpitas, CA).
9. IPGTT:
   - Timer.
   - 10 % glucose solution (*see* **Note 9**).

- 1 ml syringes with 18 G and 27 G needles.
- Glucometer (One Touch Ultra).
- Capillary tubes (Fisher).
- Box of dry ice and box of regular ice.
- Labeled small Eppendorf tubes (2 per animal; labeled).

10. Fasting test:
    - Glucometer (One Touch Ultra, LifeScan Inc., Milpitas, CA).
    - Capillary tubes (Fisher).
    - Box of dry ice and box of regular ice.
    - Labeled small Eppendorf tubes (2 per animal; labeled).

3 Methods

3.1 Making Polycistronic Construct (See Note 10)

1. Synthesize the two complementary strands of DNA oligonucleotides as shown in Fig. 1a. This linker contains restriction sites and three 2A sequences. Anneal the two oligos with a standard procedure and clone the double-strand linker into pENTR 2B vectors by Sal I and NotI. This construct is named pENTR-linker 2A.
2. Then, sequentially clone the mouse cDNA of three reprogramming factors and mCherry into the pENTR-linker 2A: Ngn3 by Sal and SpeI, Pdx1 by BamHI, Mafa by ClaI, and mCherry by NotI (Fig. 1a, b). This construct is named pENTR-linker 2A-M3C.
3. At same time, mCherry is cloned into the pENTR-linker 2A by SalI and SpeI as a control.

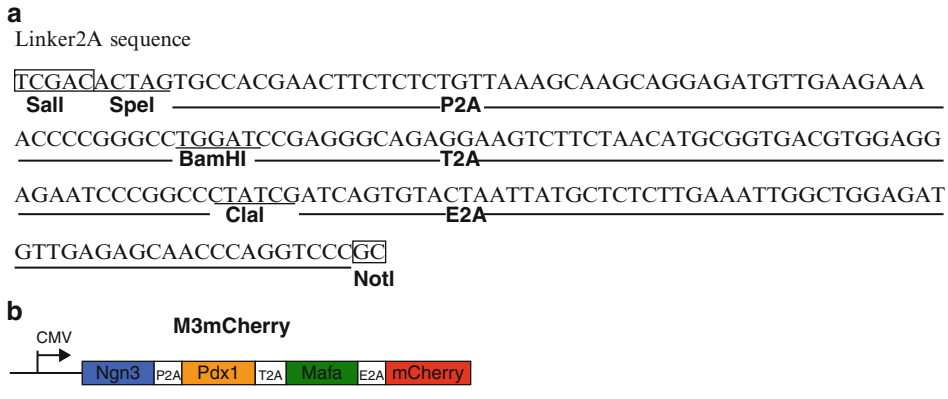


Fig. 1 Polycistronic M3Cherry viral construct. (**a**) Sequence for linker 2A. (**b**) Three reprogramming factors and mCherry are linked by 2A peptides in the polycistronic system. The three different 2A peptides used were P2A (porcine teschovirus-1 2A), T2A (Thosea asigna virus 2A), and E2A (equine rhinitis A virus 2A). CMV, cytomegalovirus promoter

All molecular constructs described in this protocol are freely available. Please send requests to qiao_zhou@harvard.edu.

3.2 Viral Production and Purification

1. Clone the M3C fragment on the pENTR-linker 2A-M3C further into a pAd-V5 DEST vector through Gateway cloning following manufacturer's instructions. This construct is named pAd-M3C.
2. Digest about 5 μg of pAd-M3C plasmid by Pac I to expose the left and right ITRs for the adenovirus packaging and replication. Purify the digested pAd-M3C further by phenol/chloroform extraction and finally elute by sterile water to a concentration of 1.0 μg/μl.
3. Plate 5×10^5 293a cells (from Invitrogen) with 2 ml growth medium with serum into each well of a 6-well cell culture plate the day before the transfection.
4. On the day of transfection, replace the cell culture medium with 1.5 ml Opti-MEM I medium containing serum without antibiotics. Then incubate about 1–5 μg purified pAd-M3C (from **step 2**) with 250 μl Opti-MEM I medium without serum for 5 min at room temperature. At the same time, add 3 μl Lipofectamine 2000 into another 250 μl Opti-MEM I medium without serum and incubate for 5 min at room temperature.
5. After incubation, mix the medium containing pAd-M3C or pAd-Cherry and the medium containing Lipofectamine 2000 and incubate at room temperature for 20 min. Then add the mixture to the 293a in 6-well plate.
6. On the second day after transfection, replace the medium with complete culture medium.
7. 48 h after the transfection, monitor the mCherry expression under a fluorescent microscope. Tryptinize the cells and transfer them into a 10 cm tissue culture plate with 10 ml 293a complete culture medium.
8. Harvest the cells by a cell scraper once 80 % of the 293a cells express mCherry (*see* **Note 11**).
9. Transfer the cell lysate into a 15 ml tube, freeze and thaw for three cycles on dry ice to release the adenovirus from the cells.
10. Centrifuge at $329 \times g$ for 15 min to remove the cell pellet.
11. Take the supernatant as crude viral lysate and store at −80 °C (*see* **Note 12**).
12. Expand cultured 293a cells to ten 15 cm plates at 80 % confluency.
13. Add 200 μl of the crude viral lysate into each plate to infect the cells.

14. Harvest the infected cells until 90 % of the cells have rounded up (*see* **Note 11**).
15. Collect the concentrated adenovirus with the Vivapure AdenoPACK 100 kit.
16. Store the purified virus in storage buffer (20 mM Tris/HCl, 25 mM NaCl, 2.5 % glycerol (w/v), pH = 8.0) at −80 °C (*see* **Note 13**).
17. Plate 1×10^5 293a cells per well in a 24-well plate. You need 3 wells. Culture the cells at 37 °C incubator for 1 day.
18. Thaw the purified virus on ice. Make serial dilutions of the virus to $1{:}10^6$.
19. Add 1 μl of the final dilution to infect cells in each well on the 24-well plate.
20. 24–36 h post infection, count the mCherry-positive cells under a fluorescent microscope. Virus titration (pfu/ml) = mCherry-positive cell number $\times 10^9$. Get the average from the three infection replicates. Expect titers of $2–10 \times 10^{10}$ pfu/ml.

3.3 Surgical Procedure

1. Before procedure: Wipe surfaces with 70 % alcohol prep for cleaning purposes. Weigh all animals as anesthesia will be weight adapted. Animals are anesthetized for survival surgery (*see* **Note 1**). Once asleep, shave animals' left side and clean the skin three times alternately with alcohol preps and Betadine solution.
2. Virus handling: Keep adenovirus at −80 °C for long-term storage and prevent freezing and thawing for more than three times (*see* **Note 13**). When filling the syringe with virus, be careful not to create bubbles. Dilute the virus with storage buffer (20 mM Tris/HCl, 25 mM NaCl, 2.5 % glycerol (w/v), pH = 8.0) to the final injection titer of about 2×10^{10} pfu/ml. Use 100 μl of the diluted virus for each animal.
3. Virus injection: The animal lies on its right side. Palpate the left costal arch and make a small incision about 0.5 cm distally of the costal arch with sharp scissors. Separate the skin from the subcutaneous tissue with scissors. Move the incision you made so that the red-colored spleen can be seen shining through the peritoneum. Lift up the peritoneum and cut little incision. Enlarge the incision above the spleen. From now on use a microscope for surgery. Pop out the spleen by slight pressure. With forceps in your left hand, get hold of the tail of the pancreas where it is attached to the spleen. If necessary clear the pancreas from the mesentery, which is slightly glossier than the pancreas. Inject 100 μl of purified virus in 1–2 loci with a 3/10 ml insulin syringe. A bubble builds where virus is injected successfully. Put the pancreas and spleen back into the abdomen. Suture the peritoneum and staple the skin.

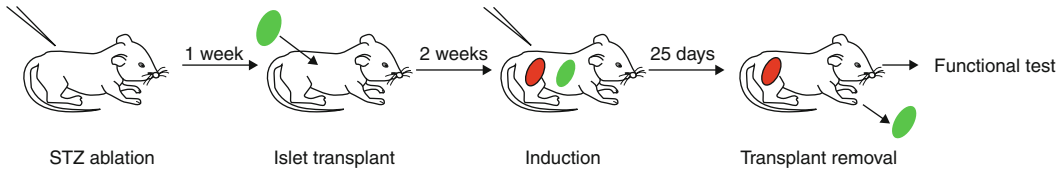


Fig. 2 Schematic diagram of function test of induced beta cells. STZ, streptozotocin. *Green circle*, islet transplantation. *Red circle*, infected pancreas

4. After procedure: Put Betadine solution on the wound. Inject analgesia (Banamine, Merck) intramuscularly into the opposite rear leg of anesthesia injection. Place the animals under a heating lamp until anesthesia recovery.

3.4 Standard Methods of Analysis

Two methods for analyzing induced cells will be discussed: (1) phenotype analysis by immunohistochemistry (IHC) and (2) functional in vivo analysis by glucose monitoring, oral glucose tolerance, and fasting tests. In order to test induced cells' function in vivo independently of endogenous beta cells (Fig. 2), we established a protocol where animals are made diabetic first, subsequently transplanted with islet cell transplants, and then reprogrammed under normoglycemic conditions. Later, transplanted islet cell grafts are removed by nephrectomy, so that glycemic control solely depends on reprogrammed cells.

3.4.1 Phenotype Analysis: Immunohistochemistry (IHC)

1. Tissue collection, fixation, and sectioning: Animals are anesthetized for terminal procedure (*see* **Note 2**). Under a dissecting microscope, an incision is made at the site of the suture from the virus injection (cut out staples). The infected dorsal pancreas is loosened from the surrounding tissue, separated from the pancreatic body with sharp scissors, and immediately placed in PBS in a Petri dish. Under a fluorescence microscope, the infected area of the pancreas is further dissected out by visualizing the Cherry fluorescence (*see* **Note 14**). Then, the dissected pancreas is fixed in 4 % paraformaldehyde for 2 h at 4 °C. Samples are subsequently washed with PBS, incubated in 30 % sucrose solution overnight (6–12 h) and embedded with O.C.T. Compound.
2. Immunohistochemistry: Frozen sections will be cut 12–14 μm thick and stored in a freezer of −80 °C until further processing (*see* **Note 15**). Thaw sections briefly at room temperature and wash with PBS 10–15 min at RT. Incubate in PBS with 0.2 % Triton X and 10 % normal donkey serum (NDS) for 30 min at RT. The slides are then incubated with the primary antibody/antibodies in 5 % NDS blocking buffer and 0.2 % Triton X overnight at 4 °C. The next day, wash with PBS with 0.2 % Triton X for 2 × 5 min at RT. Add the secondary antibody/antibodies in 5 % NDS blocking buffer and 0.2 % Triton X for

1 h at RT. Wash with PBS/0.2 % Triton X for 2×5 min. Incubate with DAPI in PBS for 5 min at RT. Wash with PBS for 2×5 min. Mount slides with Vectashield and store at 4 °C until analyzing by confocal microscopy.

3.4.2 Functional Analysis: In Vivo Testing

1. Administration of STZ: Weigh animals and calculate weight-adapted STZ dose (160–180 mg/kg). Make citrate buffer (*see* **Note 4**). Prepare a 2 % solution of streptozotocin (STZ) dissolved in citrate buffer (*see* **Note 16**). Inject STZ solution as a single dose intraperitoneally into Rag−/− mice after an 8-h daytime fast (*see* **Note 17**). Monitor glucose in the days following STZ injection by blood from snipped tails with a glucometer (*see* **Note 18**).
2. Islet isolation: Islets from adult 6-week-old male C57BL/6J mice are isolated by collagenase digestion with rodent Liberase RI and purified by gradient separation using Histopaque. For a detailed protocol, *see* ref. 20. Estimating the islet number: To estimate the number of collected islets, take three separate drops of 100 μl out of a well-suspended islet suspension of defined volume (e.g. 30 ml). Count the number of islets in each 100 μl drop and average the number of islets per 100 μl. Extrapolate the total number of islets in the islet suspension (*see* **Note 19**).
3. Making islet aliquots for transplantation: Spin down the islet suspension, aspirate media, and resuspend in volume needed (*see* **Note 20**). Make 1 ml aliquots of islet suspension in sterile Eppendorf tubes on ice (*see* **Note 20**). Also prepare a 50 ml Falcon tube with approximately 25 ml RPMI media for transplantation.
4. Islet preparation before transplantation: Have one 15 ml Falcon tube ready for each Hamilton syringe you intend to use. From the 50 ml Falcon tube with media, add 0.5 ml of media into each 15 ml Falcon tube. Fill each Hamilton syringe with 0.1–0.3 ml of media from the 50 ml Falcon tube. Attach a previously cut and sterilized 200 μl tip and fill it with media from the Hamilton syringe (*see* **Note 5**). Make sure there are no air bubbles in the Hamilton syringe. Spin down the Eppendorf tubes containing the islets (*see* **Note 8**) and put them back on ice. Remove the media of one of the Eppendorf tubes and gently stir islet pellet with the tip of the Hamilton syringe to detach islets from the bottom of the tube. Slowly draw all islets into the tip of the Hamilton syringe. After drawing the tissue up, draw up some air into the Hamilton syringe and place it tip down in one of the 15 ml tubes with 0.5 ml media in it. Let the syringe sit for 1–5 min for the islets to settle on the air bubble. Attach the transplant tubing (*see* **Note 6**) to

the Hamilton syringe. Twist the syringe clockwise until all the tissue is in the tubing. Fold the tubing where the tissue ends at the bottom of the tube and put a 2 cm long tube over the bend to hold it secure. Remove the tip and tubing from the Hamilton syringe, place it in a cut off 1 ml syringe in a 15 ml conical centrifuge tube (*see* **Note 7**) and centrifuge (*see* **Note 8**). If some media was lost during the centrifugation, refill the tip with media and reattach to the Hamilton syringe. Make sure to turn the syringe back (counterclockwise) a quarter of a turn to relieve the pressure on the islets. Remove the 2 cm tube from the transplant tubing and unfold the tubing. Determine the length of the islet pellet as a measure for islet mass. Hand the Hamilton syringe to the person performing the islet transplantation.

5. Islet transplantation (*see* **Note 21**): Animals are anesthetized for survival surgery (*see* **Note 1**) and shaved on their right side. Cut the transplant tubing containing the islet pellet just below the islets with sharp scissors (about a 60° cut). Once the animal is asleep, put it on its left side on gauze on the operating board and clean the skin with 70 % alcohol. Make an incision just above the kidney (*see* **Note 22**). Gently pop the kidney out by pressing under the kidney. Clamp the skin with a bulldog clamp to hold the kidney out of the animal (make sure not to clamp the blood vessels going to the kidney). Make a small incision in the kidney capsule on the lower pole of the kidney using a 23 G needle (*see* **Note 23**). Pick up the capsule with fine forceps and gently slide the transplantation tubing containing the islet pellet under the capsule. Slowly turn the syringe plunger clockwise to send the islets out of the transplantation tubing. When most of the islets are out, slowly begin to pull the tubing out while still sending the islets forward. Once the tubing is out, cauterize the incision hole. Close up the animal by sewing the muscle and stapling the skin. Drip Betadine on the wound and keep the animal warm until it wakes up from anesthesia.
6. Virus injection: Procedure details *see* Subheading 3.3 (*see* **Note 24**).
7. Nephrectomy (*see* **Note 25**): Animals are anesthetized for survival surgery (*see* **Note 1**). Clean with 70 % alcohol and cut out the staples at the site of islet transplantation. Gently pop the kidney out by pressing under the kidney. Clamp the hilus of the kidney with a bulldog clamp where vessels and the ureter enter the kidney. Ligate the hilus just below the clamp with very firm knots. Cut through hilus just above the clamp and slowly open clamp to check for bleeding (*see* **Note 26**). Close up the animal as indicated above.

8. Glucose monitoring: The easiest way to assess glycemia is by measuring blood glucose from snipped tails with a glucometer in regular intervals.
9. Intraperitoneal glucose tolerance test (IPGTT; *see* **Note 27**): Measure body weight of animals. Make a 10 % glucose solution (*see* **Note 9**) and inject at a dose of 2 g/kg 10 % glucose solution intraperitoneally. Measure blood glucose with a glucometer at designated time points (0, 15, 30, 60, 90, 120 min). For plasma, take capillary tube of blood at designated time points (0, 15, 30 min; *see* **Note 28**).
10. Fasting test: Change cages so that animals have access to water only (*see* **Note 17**). Measure baseline blood glucose at time point 0 and collect capillary tube of blood for insulin measurement (*see* **Note 28**). Take blood glucose every 2 h and capillary tubes every 4 h. Freeze samples at −80 °C until further processing by ELISA.

4 Notes

1. Anesthesia for survival surgery: Mix 1.0 ml ketamine (Putney, 100 mg/ml), 0.2 ml xylazine (Lloyd, 100 mg/ml), and 5.4 ml saline 0.9 %. Inject a weight-adapted dose intraperitoneally (ketamine 90–100 mg/kg, xylazine 10 mg/kg, e.g., for a 20 g mouse inject 0.1 ml of anesthesia solution). When using diabetic animals, they might not absorb well from the peritoneum and the injection should be given intramuscularly in the rear leg muscle. Place animals in a warm and calm environment to let them fall asleep.
2. Anesthesia for terminal procedure: Mix 10 ml of ketamine (100 mg/ml), 0.6 ml of xylazine (100 mg/ml), and 2.5 ml of saline 0.9 %. The xylazine and saline can be added directly to the ketamine bottle. Per 1 g body weight of animal, inject 0.00205 ml of anesthesia solution intraperitoneally. Doses might vary for different mouse strains.
3. PBS preparation for immunohistochemistry: 10× PBS is prepared by adding 85 g NaCl (Sigma), 10.7 g Na_2HPO_4 (dibasic, Sigma), and 3.9 g $Na_2H_2PO_4$ (monobasic, Sigma) to 1 l distilled H_2O. To dilute to 1× PBS, add 900 ml distilled H_2O to 100 ml 10× PBS. Titrate pH to 7.4.
4. Citrate buffer: To get 100 ml 0.9 % saline, dissolve 0.9 g NaCl (Sigma) in 100 ml water. Dissolve 0.21 g citric acid (Sigma) in 100 ml 0.9 % saline. Titrate pH to 4.5. Filter through 0.22 μM filter (Millipore) into sterile bottle/tubes. Aliquots can be stored at −20 °C.
5. Cut 200 μl tips (take 0.5 cm of the base off) and sterilize. Tips are cut so that they will fit the Hamilton syringe.

6. Prepare transplant tubing: Siliconize PE50 tubing (VWR, Radnor, PA) with Sigmacote (Sigma). Cut 10–15 cm lengths of the PE50 tubing. Cut 1 cm lengths of the connector tubing (Cole Parmer, Vernon Hills, IL) and attach to the PE50 tubing on one side. Have the transplant tubing gas sterilized.
7. Take a 1 ml syringe and cut the tip off. These syringes are put into the 15 ml conical tubes of the centrifuge to hold the transplant tubing with attached tip during centrifugation.
8. Centrifuge at $190 \times g$ for 1:15 min. We use an IEC clinical centrifuge.
9. Glucose 10 % solution: Add 4.4 ml glucose 45 % (Sigma) to 15.6 ml saline 0.9 % (Hospira).
10. All the virus production and handling should be performed under BL2 conditions.
11. It is very important to determine the proper time for harvesting the 293a cells infected by crude lysate: harvesting the cells too early or too late will yield a poor viral titer (usually below 1×10^{10} pfu/ml). The best infection status for harvesting is the following: (a) 100 % of the 293a cells express mCherry, and (b) 90 % of the cells have rounded up or float in the medium.

    If the infected cells are lysed very slowly (more than 7 days) after adding the crude lysate, it might be due to a too small amount of the crude lysate. In that case, prepare the cells again, and repeat the infection with more crude lysate (e.g., five times higher than before).

    If the results are still poor, the following steps should be carried out:

    - Re-titer the virus to exclude the possibility of a mistake in calculation.
    - Check whether the 293a cells used have a high passage number. Too many passages may yield a poor viral titer. 293a cells with less than 15 passages should be used for the virus production.
    - Large-scale viral production for animal injection should always begin with crude lysate, not purified virus, which may yield reasonable titers but generally poor beta-cell induction. The reason for this is not clear.
12. For storage purposes, we found that aliquots of 1 ml are optimal for the crude lysate. Frequent freezing and thawing of the virus will decrease the viral titer. No more than three cycles of freezing and thawing should be performed for the crude lysate.
13. For storage purposes, we found that aliquots of 100 μl are optimal for the purified adenovirus. Adenovirus in storage buffer is unstable once thawed. Use only freshly thawed virus for animal induction experiments. Leftovers may be refrozen and

used one more time, but expect a possible significant reduction in beta-cell induction.

14. Dissection of pancreatic tissue has to be performed as fast as possible in order to avoid disintegration by pancreatic enzymes.
15. Beta cells have an average diameter of approximately 10–15 μm. By cutting sections of 12–14 μm, we ensure that a single beta cell is only counted once if quantification of induced cells is performed.
16. For a 2 % STZ solution, add 5 ml of citrate buffer to a vial containing 100 mg STZ. Dissolved STZ must be used within 10 min.
17. It is critical that a relatively short fasting (8 h) is performed during daytime, rather than nighttime. We have found that Rag animals are sensitive to the destruction of beta cells by STZ. They are prone to developing hypoglycemia after STZ treatment which often leads to the death of animals. If animal death occurs, use the lower range of the recommended STZ dose (160 μg/kg body weight). To ensure abstinence from all food, the feeding grid as well as the bottom of the cage should be exchanged as small particles of food might adhere.
18. After STZ administration, low glucose levels are seen the morning after the injection (due to release of insulin from destroyed islets). On the second day after STZ, blood sugars should be high indicating destruction of endogenous islets.
19. For example, islet suspension = 30 ml. Average of islets/100 μl = 20. Therefore, 6,000 islets can be estimated in 30 ml islet suspension.
20. Decide on number of islets to transplant per animal, e.g., 600 islets/animal ("transplantation dose"). Add 1 ml of media for each "transplantation dose," e.g., 6,000 islets in total, desired "transplantation dose" 600 islets/animals = > add 10 ml media for resuspension. Aliquots of 1 ml will contain 600 islets each.
21. In general, transplantation works best early after STZ injection. However, at least 3 days should be waited after STZ injection to document diabetic status of animals (blood glucose >350 mg/dl).
22. The kidney is located in the corner between the rib cage and the back muscle.
23. Make sure to drip saline 0.9 % on the kidney throughout the procedure whenever the kidney starts to look dry.
24. We chose a 2-week interval between islet transplantation and viral injection. This gives animals time to recuperate from the transplantation and islets to adapt to the new environment.
25. Ideally, the interval between virus injection and nephrectomy should be as long as possible for induced cells to adopt a beta-

cell-like phenotype. We chose a 25-day interval, but a longer interval of up to 60 days can be used. Shorter intervals will subject immature induced beta cells to hyperglycemia and may lead to cell death.

26. The major risk during nephrectomy is bleeding from renal vessels.
27. While IPGTTs are normally performed in fasted animals, IPGTTs early after reprogramming can be performed in fed animals due to a tendency of hypoglycemia during that time.
28. For blood collection into capillary tubes, start at the base of the tail gently squeezing out blood from tail with fingers (preferentially do not use gloves as more injuries of the tail occur). Once the desired amount of blood is collected, push it into labeled Eppendorf tube using a P 200 pipette. Store blood on ice until spinning it down on benchtop centrifuge for 1.5–2 min. Transfer serum to another labeled Eppendorf tube using a P 20 pipette. Try to get at least 20 μl serum. Store serum on dry ice. Freeze serum at −80 °C until further processing by ELISA at the end of the test.

Acknowledgment

We thank Jennifer Hollister-Lock for revising the manuscript. This work was supported by NIDDK and HSCI. W. L. is supported by a postdoctoral fellowship from the Juvenile Diabetes Research Foundation (JDRF). C. C-W. is supported by postdoctoral fellowships from the Swiss Science Foundation (SNF) and the Swiss Foundation for Grants in Biology and Medicine (SFGBM).

References

1. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, Lin JK, Farzadfar F, Khang YH, Stevens GA, Rao M, Ali MK, Riley LM, Robinson CA, Ezzati M (2011) National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet 378(9785):31–40
2. American Diabetes Association (2014) Diagnosis and classification of diabetes mellitus. Diabetes Care 37(Suppl 1):S81–90
3. Butler AE, Janson J, Bonner-Weir S, Ritzel R, Rizza RA, Butler PC (2003) Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes. Diabetes 52(1):102–110
4. Foulis AK, Liddle CN, Farquharson MA, Richmond JA, Weir RS (1986) The histopathology of the pancreas in type 1 (insulin-dependent) diabetes mellitus: a 25-year review of deaths in patients under 20 years of age in the United Kingdom. Diabetologia 29(5):267–274
5. Dor Y, Brown J, Martinez OI, Melton DA (2004) Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. Nature 429(6987):41–46
6. Teta M, Long SY, Wartschow LM, Rankin MM, Kushner JA (2005) Very slow turnover of beta-cells in aged adult mice. Diabetes 54(9): 2557–2567
7. Nir T, Melton DA, Dor Y (2007) Recovery from diabetes in mice by beta cell regeneration. J Clin Invest 117(9):2553–2561

8. Xu X, D'Hoker J, Stangé G, Bonné S, De Leu N, Xiao X, Van de Casteele M, Mellitzer G, Ling Z, Pipeleers D, Bouwens L, Scharfmann R, Gradwohl G, Heimberg H (2008) Beta cells can be generated from endogenous progenitors in injured adult mouse pancreas. Cell 132(2):197–207
9. Inada A, Nienaber C, Katsuta H, Fujitani Y, Levine J, Morita R, Sharma A, Bonner-Weir S (2008) Carbonic anhydrase II-positive pancreatic cells are progenitors for both endocrine and exocrine pancreas after birth. Proc Natl Acad Sci U S A 105(50):19915–19919
10. Rosenberg L (1998) Induction of islet cell neogenesis in the adult pancreas: the partial duct obstruction model. Microsc Res Tech 43(4):337–346
11. Bonner-Weir S, Toschi E, Inada A, Reitz P, Fonseca SY, Aye T, Sharma A (2004) The pancreatic ductal epithelium serves as a potential pool of progenitor cells. Pediatr Diabetes 5(Suppl 2):16–22
12. Bonner-Weir S, Baxter LA, Schuppin GT, Smith FE (1993) A second pathway for regeneration of adult exocrine and endocrine pancreas. A possible recapitulation of embryonic development. Diabetes 42(12):1715–1720
13. Zhou Q, Brown J, Kanarek A, Rajagopal J, Melton DA (2008) In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. Nature 455(7213):627–632
14. Thorel F, Népote V, Avril I, Kohno K, Desgraz R, Chera S, Herrera PL (2010) Conversion of adult pancreatic alpha-cells to beta-cells after extreme beta-cell loss. Nature 464(7292):1149–1154
15. Ferber S, Halkin A, Cohen H, Ber I, Einav Y, Goldberg I, Barshack I, Seijffers R, Kopolovic J, Kaiser N, Karasik A (2000) Pancreatic and duodenal homeobox gene 1 induces expression of insulin genes in liver and ameliorates streptozotocin-induced hyperglycemia. Nat Med 6(5):568–572
16. Collombat P, Xu X, Ravassard P, Sosa-Pineda B, Dussaud S, Billestrup N, Madsen OD, Serup P, Heimberg H, Mansouri A (2009) The ectopic expression of Pax4 in the mouse pancreas converts progenitor cells into alpha and subsequently beta cells. Cell 138(3):449–462
17. Chung CH, Hao E, Piran R, Keinan E, Levine F (2010) Pancreatic beta-cell neogenesis by direct conversion from mature alpha-cells. Stem Cells 28(9):1630–1638
18. Zhou Q, Melton DA (2008) Extreme makeover: converting one cell into another. Cell Stem Cell 3(4):382–388
19. Szymczak AL, Vignali DA (2005) Development of 2A peptide-based strategies in the design of multicistronic vectors. Expert Opin Biol Ther 5(5):627–638
20. O'Dowd JF (2009) The isolation and purification of rodent pancreatic islets of Langerhans. Methods Mol Biol 560:37–42

Chapter 18

Direct Reprogramming of Cardiac Fibroblasts to Cardiomyocytes Using MicroRNAs

Tilanthi Jayawardena, Maria Mirotsou, and Victor J. Dzau

Abstract

The therapeutic administration of microRNAs represents an innovative reprogramming strategy with which to advance cardiac regeneration and personalized medicine. Recently, a distinct set of microRNAs was found capable of converting murine fibroblasts to cardiomyocyte-like cells in vitro. Further treatment with JAK inhibitor I significantly enhanced the efficiency of the microRNA-mediated reprogramming (Jayawardena et al., Circ Res 110(11):1465–1473, 2012). This novel technique serves as an initial tool for switching the cell fate of cardiac fibroblasts toward the cardiomyocyte lineage using microRNAs. As the budding field of reprogramming biology develops, we hope that a thorough examination of the chemical, physical, and temporal parameters determining reprogramming efficiency and maturation will enable a better understanding of the mechanisms governing cardiac cell fate and provide new approaches for drug discovery and therapy for cardiovascular diseases.

Key words Reprogramming, Cardiac, Cardiomyocyte, MicroRNA, Fibroblast, Regeneration, Transfection

1 Introduction

Cardiac fibroblasts represent a significant fraction of the non-myocyte cell population in the heart and play a key role in pathological cardiac remodeling; as such they are attractive therapeutic targets for cardiac repair [1]. The seminal discovery of inducible pluripotent stem cells (iPS) in 2006 [2] has led to an active pursuit of using fibroblast reprogramming in cardiac regenerative medicine [3–6]. While several publications have reported iPS-derived cardiomyocytes, there are still concerns about the maturity and functional heterogeneity of these cells, their low survival and retention when delivered to the injured myocardium, as well as their potential tumorigenicity [7, 8]. Accordingly, the direct reprogramming of cardiac fibroblasts using transcription factors [5, 6, 9, 10], microRNAs [3], or a combination of both [4] is of particular interest to the field of regenerative medicine. We have

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_18, © Springer Science+Business Media New York 2014

recently developed a novel method using a single transient transfection of microRNAs 1, 133a, 208a, and 499-5p accompanied by treatment with JAK inhibitor I [3] to switch the cell fate of cardiac fibroblasts to the cardiomyocyte cell lineage. This treatment generated cardiomyocyte-like cells in vitro that expressed cardiomyocyte-specific genes and proteins and sarcomeric organization and exhibited spontaneous calcium oscillations. This method is significant as a preliminary tool for the reprogramming of cardiac fibroblasts to the cardiomyocyte lineage.

2 Materials

2.1 Isolation and Passaging of Neonatal Cardiac Fibroblasts

1. Digestion buffer—Hanks' Balanced Salt Solution (Sigma-Aldrich, St. Louis, MO) supplemented with Collagenase type II, 100 U/mL (Worthington Biochemical, Lakewood, NJ).
2. Fibroblast culture medium—Dulbecco's Modified Eagle's Medium (Cat# 30-2002, ATCC, Manassas, VA) supplemented with 15 % fetal bovine serum (not heat deactivated) (Cat# SH30071.03, HyClone/Thermo Scientific, Waltham, MA) and 1 % penicillin–streptomycin (100 U/mL penicillin and 100 μg/mL streptomycin final concentration) (Gibco/Life Technologies, Grand Island, NY).
3. Gelatin—0.2 % final concentration in PBS. Gelatin stock is a 2 % solution (Cat# G1393, Sigma-Aldrich, St. Louis, MO).
4. Cell strainer—100 μM (BD Biosciences, San Jose, CA).
5. 10× ADS Buffer—Mix 34 g NaCl, 23.8 g HEPES, 0.6 g NaH_2PO_4, 5 g glucose, 2 g KCl, and 0.5 g $MgSO_4$ and add water to a final volume of 500 mL.
6. Percoll Gradient Stock Solution—Make 9:1 dilution of Percoll (Cat# 17-0891-02, GE Healthcare, Piscataway, NJ) in 10× ADS Buffer. Increase volume according to volume of prep and make up fresh on the day of cell isolation.
7. Gradient Solution 1—Add 1 mL of 1× ADS Buffer to 4.15 mL Percoll Gradient Stock Solution (final density of 1.095 g/mL). Increase volume according to the number of samples and make up fresh on the day of cell isolation.
8. Gradient Solution 2—Add 2 mL of 1× ADS Buffer to 1.47 mL Percoll Gradient Stock Solution (final density of 1.1179 g/mL). Increase volume according to the number of samples and make up fresh on the day of cell isolation.
9. Gradient Solution 3—Add 2 mL of 1× ADS Buffer to 1 mL fibroblast culture medium and 0.807 mL Percoll Gradient Stock Solution (final density of 1.025 g/mL). Increase volume

according to the number of samples and make up fresh on the day of cell isolation.

10. Antibodies for Vimentin (Abcam, Cambridge, MA), Ddr2 (R&D, Minneapolis, MN).

2.2 Reagents for MicroRNA/JAK Inhibitor I Treatment

1. MicroRNAs—Synthetic mimics of mature microRNAs (Pre-miR™ miRNA Precursor Ambion/Life Technologies, Grand Island, NY). Hsa-miR-1/mmu-miR-1a-3p (Cat# PM10617/AM17100); hsa-miR-133a/mmu-miR-133a-3p (Cat# PM10413/AM17100); hsa-miR-208a/mmu-miR-208-3p (Cat# PM10677/AM17100); hsa-miR-499a-5p/mmu-miR-499a-5p; FAM™ Dye-Labeled Pre-miR Negative Control (AM17121). All microRNAs were resuspended to stock concentrations of 50 used at a final concentration of 50nM.
2. DharmaFECT 1—(Cat# Dharmacon/Thermo Scientific, Waltham, MA).
3. JAK inhibitor I—(Cat#420099, EMD Millipore, Billerica, MA).

2.3 Reagents for RNA Purification and cDNA Synthesis

1. RNeasy 96 and RNeasy Plus Micro Kits (Qiagen, Valencia, CA).
2. High-Capacity cDNA Reverse Transcription Kit (Cat# 4368814, Life Technologies, Grand Island, NY).

2.4 Reagents for Quantitative Real-Time PCR

1. TaqMan Gene Expression and MicroRNA Expression Assays (Applied Biosystems/Life Technologies, Grand Island, NY).
2. StepOnePlus Real-Time PCR System (Applied Biosystems/Life Technologies, Grand Island, NY).

2.5 Reagents for Immunocytochemistry, Microscopy, and FACS

1. Paraformaldehyde.
2. Antibodies for α-actinin (Sigma-Aldrich, St. Louis, MO), myosin heavy chain (Abcam, Cambridge, MA), cardiac troponin I and T (Abcam, Cambridge, MA).
3. 24-well glass bottom multi-well plates (Cat# P24G-1.0-13-F, Mattek Corporation, Ashland, MA).
4. 35 mm glass bottom dishes (Cat# P35G-0-10-C, Mattek Corporation, Ashland, MA).

2.6 Reagents for Calcium Imaging

1. Fura-2 AM, used at final concentration of 1 μM (Molecular Probes/Life Technologies, Grand Island, NY).
2. Standard Ringer Solution (140 mM NaCl, 2.8 mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 10 mM glucose, 10 mM Hepes, pH 7.4).
3. High [K^+] Ringer Solution (80 mM NaCl, 62.8 mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 10 mM glucose, 10 mM Hepes, pH 7.4).

3 Methods

3.1 Isolation and Passaging of Neonatal Cardiac Fibroblasts (Ventricular)

1. Prepare a 0.2 % gelatin (in PBS) solution (1:10 of gelatin stock in PBS). Coat flasks and place in incubator for at least 30 min prior to use. Immediately before seeding cells, aspirate any excess gelatin solution from flasks. One T25 flask (Corning 430639, Corning, NY) is sufficient for seeding the cells acquired from two neonatal hearts.
2. From 1- to 2-day-old neonatal mouse pups, remove ventricles under sterile conditions and then wash briefly in PBS. To harvest a neonatal heart, begin by first disinfecting the pup with 70 % ethanol, followed by decapitation, and cutting through the frontal rib cage to reveal a clear view of the heart and lungs (*see* **Note 1**). While in PBS, remove atria and then move the ventricles immediately to digestion buffer.
3. Once in digestion buffer, cut the tissue into very small pieces using a razor and/or scissors (*see* **Note 2**).
4. Incubate the minced tissue in 2 mL of the same digestion buffer for 10 min at 37 °C. At the end of the incubation, pass the lysate through a cell strainer (100 μm) and collect the flow-through in a 50 mL Falcon tube. Add 2 mL of fibroblast culture media to the flow-through (to deactivate collagenase) and, in parallel, continue to perform another round of digestion on any remaining undigested tissue (*see* **Note 3**). Repeat the procedure until all ventricular tissue is digested (usually 2–3 rounds). At the end, pool the flow-through from each digestion round, spin down at 3,000 × *g* for 5 min, and resuspend in 945 μL fibroblast culture medium containing in a 15 mL Falcon tube. Next, subject cells to Percoll gradient centrifugation.
5. Add 2.055 mL of Gradient Solution 1 to the 945 μL of resuspended cells from the previous step. Then, gently overlay this layer with 3 mL of Gradient Solution 2 using a 5 mL sterile, disposable transfer pipet. Carefully control both exertion of liquid (maintain tip of pipet at the meniscus of bottom layer) and positioning of recipient 15 mL Falcon tube so as to ensure that the two layers do not mix. Then using similar techniques, overlay this layer with 3 mL of Gradient Solution 3. Three well-separated layers should be observed clearly at this point. Centrifuge at 1,500 × *g* with the brake turned off for 10 min at room temperature.
6. Remove the top 2.5 mL and transfer the middle layer (~3–3.5 mL) to a separate 15 mL Falcon tube. Add fibroblast culture medium to a total volume of 15 mL and centrifuge for 5 min at 300 × *g*. Resuspend the pellet in 5 mL of culture medium and seed in gelatin-coated T25 flasks. Change the media after 24–48 h and continue growing till the flask is completely confluent.

7. Split P0 cells (from 80 to 90 % confluent T25 flasks) to P1 at a ratio of 1:6. Subsequently, split additional passages at ratios of 1:3. Do not conduct microRNA transfections in cells at P5 or beyond.
8. Verify the isolated cell population by staining for classic fibroblast markers such as Vimentin or Ddr2.

3.2 MicroRNA Transfection of Neonatal Cardiac Fibroblasts

1. Split neonatal cardiac fibroblasts to P3 or P4 24 h prior to transfection. Transfections can be performed in 96-well, 24-well, T25, or T75 format depending on the nature of the question being investigated. Follow the guidelines in the table below for the recommended seeding density. In general, the 96-well and 24-well plate formats serve as excellent screening tools, while the T25 and T75 formats generate more cells for follow-up studies. For immunocytochemistry experiments, cells can be seeded in glass bottom multi-well format (Mattek, Ashland, MA) (*see* **Notes 4** and **5**).
2. Set up the transfection reaction under completely antibiotic-free conditions (however, seeding of cells 24 h prior can be set up in antibiotic-containing media). Start by preparing two mixtures: (1) dilution of microRNAs in serum-free (SF) media and (2) dilution of DharmaFECT 1 in serum-free media. Use the proportions outlined in the table below. Incubate at room temperate for 5 min. Note that for every microRNA transfection experiment conducted, include several untransfected, mock, and nontargeting microRNA controls (FAM™ Dye-Labeled Pre-miR Negative Control, Ambion/Life Technologies) in order to compare and validate the relative induction of cardiac reprogramming in controls versus microRNA-treated samples.
3. Add Mixture 2 to Mixture 1. Incubate at room temperature for 20 min.
4. During the incubation, change media on cells to the appropriate volume of antibiotic-free culture media (DMEM with 15 % FBS) as indicated in the table below.
5. Add the appropriate remaining volume of SF media (refer the last column of the table below) to the reaction and add the entire mixture directly to the cells.
6. After 24 h, change media on the cells and replace with regular fibroblast culture medium (with antibiotics). To boost reprogramming efficiency, supplement the culture medium with JAK inhibitor I to a final concentration of 1 μM. Continue daily treatment with the inhibitor for 4 additional days.

3.3 Evaluation of Transfection Efficiency

1. Lyse cells 48 h post-transfection (refer to the next section on recommended methods using the RNeasy 96 Kit from Qiagen, Valencia, CA) and assess knockdown of Twinfilin-1, *twf1* (mRNA target of miR-1), and/or collagen 16, *col16a1*

(mRNA target of miR-133a). The expected knockdown efficiency is in the range of 65–75 %.

2. Alternatively, the small RNA population can be harvested using the *mir*Vana™ miRNA Isolation Kit (Ambion/Life Technologies, Grand Island, NY) and processed to cDNA using the TaqMan® MicroRNA Reverse Transcription Kit (Applied Biosystems/Life Technologies, Grand Island, NY). Levels of the transfected microRNAs can then be assessed using TaqMan® MicroRNA Assays specific for each microRNA concerned.
3. The transfection can also be monitored qualitatively using the FAM™ Dye-Labeled Pre-miR Negative Control.

3.4 RNA Harvest, cDNA Synthesis, and Quantitative Gene Expression

1. Lyse cells in RLT buffer and harvest RNA according to manufacturer's recommendations (Qiagen, Valencia, CA) and either freeze lysate at −80 °C or proceed immediately to RNA purification. Depending on the format with which the experiment was conducted, purify RNA using either RNeasy 96 or RNeasy Plus Micro Kits (Qiagen, Valencia, CA).
2. Prepare cDNA using High-Capacity cDNA Synthesis Kit according to manufacturer's recommendations (Applied Biosystems/Life Technologies, Grand Island, NY) (*see* **Note 6**).
3. Assess for gene expression or microRNA expression using the TaqMan system (Applied Biosystems/Life Technologies, Grand Island, NY) using either GAPDH or beta-actin as housekeeping genes for normalization.

3.5 Evaluation of Cardiac Reprogramming Readout I: Day 3

1. In transfected neonatal cardiac fibroblasts exhibiting efficient knockdown of known targets such as *Twf1*, harvest RNA and synthesize cDNA as outlined in the section above. For the first evidence of cardiac reprogramming, assess expression using quantitative real-time PCR of the following genes: *Mef2c*, *Tbx5*, *Hand2*, *Nkx2.5*, *Gata4*, *Ddr2*, and *Vim* (Fig. 1). With the exception of the latter two genes, a two- to threefold upregulation in expression is expected relative to nontargeting microRNA and untransfected controls (*see* **Note 7**).

3.6 Evaluation of Cardiac Reprogramming Readout II: Day 6

1. Continue evaluation of reprogramming parameters by day 6 post-transfection using quantitative real-time PCR of the same genes as assessed in the previous section as well as *αMHC* and *Tnni3*. A positive reprogramming experiment at this timepoint will exhibit a continuing trend in the upregulation of genes of cardiac differentiation (*see* **Note 7**).
2. In parallel, from cells seeded on glass bottom 24-well plates (refer table above) at days 6–7 post-transfection, fix in 4 % paraformaldehyde and proceed to staining with cardiac troponin I/T, α-myosin heavy chain (Abcam, Cambridge, MA), and sarcomeric actinin (Sigma-Aldrich, St. Louis, MO).

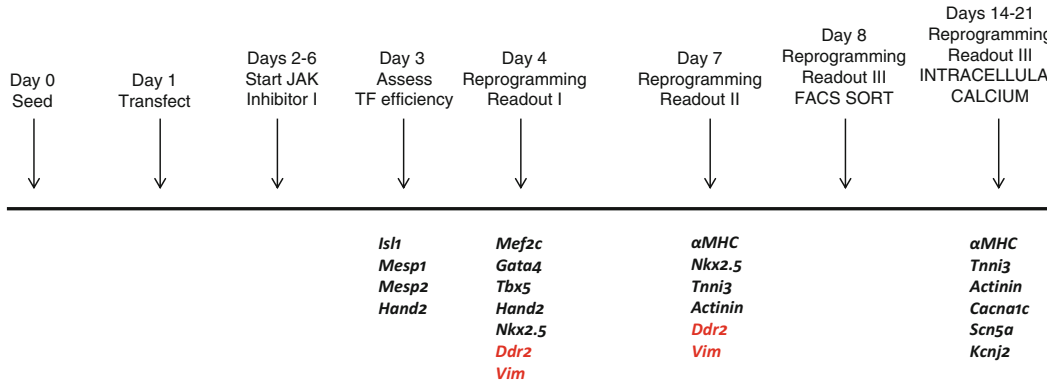


Fig. 1 Experimental outline for a typical microRNA-mediated reprogramming experiment including stages of seeding, transfection, and JAK inhibitor I treatment. Also indicated is the expected timeline for induction of markers of cardiac differentiation (in *black*) or downregulation of fibroblastic markers (in *red*), which can be assessed by quantitative real-time PCR expression and/or immunocytochemistry between 3 and 21 days post-transfection

3.7 Evaluation of Cardiac Reprogramming Readout III: Days 7–21

1. Continue evaluation of later markers of cardiac reprogramming via immunocytochemistry (see previous section) 1–4 weeks post-transfection. Positive reprogramming experiments at this timepoint will start to exhibit expression of markers such as cardiac troponin and sarcomeric actinin between the 1–2-week period following microRNA transfection (*see* **Note 7**).
2. Neonatal cardiac fibroblasts that have been isolated from a transgenic system where the expression of a fluorescent reporter is driven by a cardiomyocyte-specific promoter can be exploited for the selection of cells that have switched fate toward the cardiomyocyte lineage. Transfections should be conducted exactly as outlined in Table 1 and, starting from day 7 (*see* **Notes 8** and **9**), can be FACS sorted (sorted population consists of all cells that express CFP above gates set by untransfected controls) and analyzed in a variety of ways. One would be by quantitative gene expression, but additionally, selected cells could be maintained in culture for extended periods of time (2–4 weeks post-FACS) and tested at weekly intervals for the expression of calcium signaling properties that resemble those of developing cardiomyocytes. To this end, FACS-selected ("reprogrammed") cells are seeded on 35 mm glass bottom dishes (5,000 cells/well) (only on the glass portion of each dish) and maintained in culture. Cultures are then mounted on a Nikon TE2000 inverted microscope equipped with a Photometrics CoolSNAP camera, xenon arc lamp, and lambda DG-4 rapid filter changer (Sutter). Image acquisition, analysis, and processing are carried out using MetaFluor/MetaMorph software (Molecular Devices, Sunnyvale, CA). Fura-2 fluorescence is measured by alternate excitation at 340

Table 1
Summary of the amounts of microRNA, lipid carrier (DharmaFECT 1), media, and fibroblast seeding densities required for setting up microRNA-mediated reprogramming experiments in 96-well, 24-well, T25 flask, or T75 flask formats

Format	Seeding density	Total transfection volume	Mixture 1	Mixture 2	Volume of antibiotic-free culture media	Volume of SF[a] media (final step)
96 well	1,500 cells	100 μL	0.1 μL mi RNA[b] 9.9 μL SF media	0.1 μL Dhl[c] 9.9 μL SF media	30 μL	50 μL
24 well	9,000 cells	600 μL	0.6 μL mi RNA[b] 59.4 μL SF media	0.6 μL Dhl[c] 59.4 μL SF media	180 μL	300 μL
T25 flask	100,000 cells	5 mL	0.1 μL mi RNA 9.9 μL SF media	0.1 μL Dhl 9.9 μL SF media	1.5 mL	2.5 mL
T75 flask	300,000 cells	15 mL	0.1 μL mi RNA 9.9 μL SF media	0.1 μL Dhl 9.9 μL SF media	4.5 mL	7.5 mL

[a]SF media = serum-free media
[b]MicroRNA volumes are added from a 50 μM stock concentration. In the case of the microRNA combo (miR-1, miR-133a, miR-208a, miR-499a-5p), add ¼ the volume indicated for each microRNA. For example, when setting up Mixture 1 for the 96-well format, add the equivalent of 0.025 μL (for a total volume of 0.1 μL) of each microRNA that has been resuspended to a final concentration of 50 μM. A total volume of 0.1 μL of a 50 μM stock of the recommended nontargeting microRNA control (FAM™ Dye-Labeled Pre-miR Negative Control; AM17121) should be used as comparison
[c]Dh1 = DharmaFECT 1

and 380 nm and emission at 510 nm and the background subtracted 340/380 ratio used to express the calcium levels for each cell [11, 12]. Several parameters that characterize the calcium signaling properties of cardiomyocytes can be analyzed, including basal calcium levels, the occurrence and frequency of spontaneous calcium oscillations, and responses to agonists and depolarization with high $[K^+]$ (*see* **Note 10**).

4 Notes

1. If these steps are performed within a minute, the heart can be easily distinguished via its pumping action and, accordingly, separated and placed in PBS.
2. Mincing of the tissue in a small drop of digestion buffer (a few 100 μL) permits easier handling.
3. It is very important at this phase to intermittently pipet the digestion mixture up and down both in between 37° C incubation periods and in the middle of them. For this purpose, using

disposable 5 mL transfer pipets is helpful (Thermo Scientific, Waltham, MA).

4. On the day of transfection, cells should be evenly spread and be 55–60 % confluent. Optimal seeding density and transfection conditions might require fine-tuning to account for differences cell quality and handling.
5. For more uniform seeding (especially in the case of multiwall formats like the 96 well), seed the cells into wells containing a small-volume growth media instead of directly onto plastic. For example, when seeding a 96-well plate, seed 70 μL of cell suspension into wells containing 30 μL of fibroblast culture media.
6. In the case of T25 flask formats and larger, 1μg of total RNA was used in each cDNA synthesis reaction. For significantly smaller formats, such as the 96-well format, it is helpful to pool the lysates from multiple wells and purify as one using the RNeasy 96 system (Qiagen, Valencia, CA). Normalize by amount of input total RNA and/or relative expression levels of a housekeeping gene such as GAPDH. Alternatively RNA can be amplified using commercially available reagents.
7. Figure 1 is an indication of how microRNA-mediated cardiac reprogramming progresses in neonatal cardiac fibroblasts and the temporal expression induction patterns of a variety of cardiac markers following microRNA introduction. However, it should be stressed that these patterns of expression as well as the strength of their induction can vary from one reprogramming experiment to the next. MicroRNA-mediated cardiac reprogramming appears to be governed by several parameters including cell health, passage, type of serum, serum concentration, and seeding at the time of transfection. These factors can vary the temporal induction patterns in reprogrammed cells quite significantly and can be the cause of high levels of experimental variability. We therefore recommend closely following cardiac fibroblasts before and during reprogramming experiments and always evaluating transfection efficiency 48 h post-transfection. Additionally, evaluation of this expression can be very challenging due to the fact that the efficiency of reprogramming is typically low (1.13–5.28 % in non-JAK inhibitor I-treated cells [3]). Therefore, assessing the induction of expression in nonselected cell populations can be difficult. When assessing nonselected, transfected cell populations by a method such as quantitative real-time PCR for instance, the effects induced by the microRNAs are diluted by the presence of untransfected fibroblasts. The latter appear to proliferate faster than transfected cells, and as a result, this issue is exacerbated at later timepoints. We find that our most reliable quantitative real-time PCR data are generated from nonselected populations between days 3 and 6 post-transfection.

8. The T75 format is recommended when selection of the reprogrammed population is desired. This can be achieved by using fibroblasts isolated via fluorescent reporters driven by a cardiomyocyte-specific promoter (e.g., α-myosin heavy chain driving CFP or GFP).
9. Sorted population consists of all cells that express CFP above gates set by untransfected controls (use the same batch of fibroblasts in the untransfected controls and microRNA transfection samples).
10. The protocol described above works also well for transfection and reprogramming of adult cardiac fibroblasts. However, the efficiency and the maturity of the induced cardiomyocyte-like cells are reduced compared to neonatal cells.

References

1. Yoshida Y, Yamanaka S (2012) An emerging strategy of gene therapy for cardiac disease. Circ Res 111(9):1108–1110
2. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126(4):663–676
3. Jayawardena TM et al (2012) MicroRNA-mediated in vitro and in vivo direct reprogramming of cardiac fibroblasts to cardiomyocytes. Circ Res 110(11):1465–1473
4. Nam YJ et al (2013) Reprogramming of human fibroblasts toward a cardiac fate. Proc Natl Acad Sci U S A 110(14):5588–5593
5. Song K et al (2012) Heart repair by reprogramming non-myocytes with cardiac transcription factors. Nature 485(7400): 599–604
6. Qian L et al (2012) In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. Nature 485(7400):593–598
7. Mummery C (2011) Induced pluripotent stem cells—a cautionary note. N Engl J Med 364(22):2160–2162
8. Yamashita T et al (2011) Tumorigenic development of induced pluripotent stem cells in ischemic mouse brain. Cell Transplant 20(6):883–891
9. Christoforou N et al (2013) Induced pluripotent stem cell-derived cardiac progenitors differentiate to cardiomyocytes and form biosynthetic tissues. PLoS One 8(6):e65963
10. Protze S et al (2012) A new approach to transcription factor screening for reprogramming of fibroblasts to cardiomyocyte-like cells. J Mol Cell Cardiol 53(3):323–332
11. Stiber J et al (2008) STIM1 signalling controls store-operated calcium entry required for development and contractile function in skeletal muscle. Nat Cell Biol 10(6):688–697
12. Eckel J et al (2011) TRPC6 enhances angiotensin II-induced albuminuria. J Am Soc Nephrol 22(3):526–535

Chapter 19

Reprogramming Somatic Cells into Pluripotent Stem Cells Using miRNAs

Frederick Anokye-Danso

Abstract

Reversal of terminally differentiated somatic cells to ground-state pluripotency has rejuvenated our hopes of generating patient-specific stem cells for therapeutic use in regenerative medicine and drug screening. Originally generated using defined exogenous protein-coding DNA, several methods have been described in reprogramming somatic cells into iPSC. Majority of published methods seek to improve or refine the techniques of reprogramming. This chapter describes reprogramming to pluripotency using miRNAs.

Key words Reprogramming, Induced pluripotent stem cells (iPSCs), microRNAs (miRNAs), Fibroblast, Lentivirus

1 Introduction

Somatic can be reprogrammed into pluripotent state by (a) nuclear transfer or (b) fusion with enucleated oocytes or (c) by forced expression of transcription factors [1]. Nuclear transfer and fusion are laborious and technically challenging. Reprogramming somatic cells into induced pluripotent stem cells (iPSC) is cheaper and easier to create. Oct4, Sox2, Myc, and Klf4 are the commonly used transcription factors in generation of iPSC [1–3]. Nanog and Lin28 in addition to Oct4 and Sox2 have also been used to generate iPSC [4]. Involvement of oncogenes Klf4 and Myc in the reprogramming cocktail makes it less attractive for therapeutic use due to their association with tumors and cancer. To overcome the use of these oncogenes, a number of small molecules from screening of pharmacologically active compounds have been identified that can substitute for these genes. Inhibitors of histone deacetylase such as valproic acid can substitute for Myc [5]. A combination of BIX-01294 and BayK8644 and other small molecules can replace Klf4 and Sox2 [6–9]. Such approach reduces the number of exogenous factors required for reprogramming into iPSC. Currently, there are no known chemical substitutes capable of replacing Oct4.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_19, © Springer Science+Business Media New York 2014

Investigators have also taken advantage of endogenous expression of some of these factors in specific tissues to reprogram into iPSC. Neural progenitor cells [10] and melanocytes and melanoma cells [11] can be induced into pluripotent stem cell state in the absence of exogenous Sox2. Other techniques of reprogramming have sought to eliminate use of viral vectors due their stable integration into the genome of recipient cells. Messenger RNAs synthesized in vitro [12] and cell-penetrating recombinant proteins [13] of OCT4, SOX2, MYC, and KLF4 have been successfully used to reprogram somatic cells into iPSC. This method also eliminates reaction of the cells to viral infection. Based on the assumption that high levels of the four transcription factors in embryonic stem cells reprogram to pluripotency, forced expression of miRNAs has been shown to convert somatic cells into iPSC [14–16]. Although there are variations in the combination of families of miRNAs that reprogram to pluripotency, miR-302 family seems to be central. It has been suggested that stem cell-specific miRNAs reprogram and promote self-renewal by inhibiting accumulation of genes that interfere with pluripotent stem cell identity [17, 18].

2 Materials

Carry out all procedure under sterilized condition. Wear personal protective equipment when working with virus. All tools and surfaces that come in contact with lentivirus must be decontaminated using Wescodyne solution. Containers used for collecting waste viral suspension must contain bleach at all times. Spray items that go in and out of the laminar flow bench with 70 % ethanol.

2.1 Components of Media, Reagents, and Plastic Wares

1. Sodium pyruvate (100 mM).
2. Leukemia inhibitory factor (LIF, 1×10^7 U/mL).
3. Knockout serum replacement.
4. Penicillin-streptomycin (10,000 U/mL).
5. MEM nonessential amino acids (100×).
6. L-glutamine (200 mM).
7. β-Mercaptoethanol.
8. Valproic acid (2-propylpentanoic acid).
9. High-glucose Dulbecco's Modified Eagle's Medium (DMEM).
10. Fetal bovine serum (FBS).
11. 0.05 % trypsin/EDTA.
12. 1× Dulbecco's Phosphate-Buffered Saline (DPBS).
13. Dimethyl sulfoxide (DMSO).
14. Transfection reagents.

15. FuGENE 6 HD transfection reagent.
16. Polybrene (10 mg/mL).
17. pMD.G packaging vector.
18. psPAX2 packaging vector.
19. miRNA-302/367 lentiviral vector.
20. 0.22 μm filter unit.
21. 0.45 μm filter unit.
22. Amicon Ultra-15 centrifugal filter unit.
23. 96-, 48-, and 6-well-coated tissue culture plates.
24. 10 cm coated tissue culture dishes.
25. Sterilized 1.5 mL Eppendorf tubes.
26. 293 T cell line.
27. Irradiated mouse fibroblast (feeder cells).
28. 70 % ethanol.
29. 15 and 50 mL disposable centrifuge tubes.
30. Disposable serological pipettes.

2.2 Media Preparation

1. Fibroblast/293 T cell growth media (500 mL): 50 mL fetal bovine serum, 5 mL penicillin-streptomycin, and top with DMEM to 500 mL (*see* **Note 1**). Filter solution through a 0.22 μm filter. Store at 4 °C until ready to use.
2. Mouse ES media (500 mL): 75 mL knockout serum replacement, 5 mL penicillin-streptomycin, 5 mL sodium pyruvate, 5 mL MEM nonessential amino acids, 5 mL L-glutamine, 50 μL LIF, 4 μL β-mercaptoethanol, 160 μL valproic acid, and top with DMEM to 500 mL (*see* **Note 2**). Filter through a 0.22 μm filter. Store at 4 °C until ready to use.

3 Method

Besides incubation at 37 °C with 5 % carbon dioxide (CO_2) in a humid chamber, all other procedures are done at room temperature.

3.1 Preparation of Primary Mouse Embryonic Fibroblast

Euthanize a pregnant mouse at E12.5–13.5 by CO_2. Briefly dip the euthanized mouse in a beaker of 70 % ethanol. Remove embryos from uterus using sterilized desertion kit. Place embryos in Petri dish containing cold DPBS. Remove the limbs, head, and internal organs. Transfer the resultant carcass into a fresh DPBS and wash. Cut carcass into pieces and place in a 50 mL disposable centrifuge tube. Add 5 mL of 0.05 % trypsin/EDTA and incubate at 37 °C on a shaker for 15 min. Pipette up and down vigorously

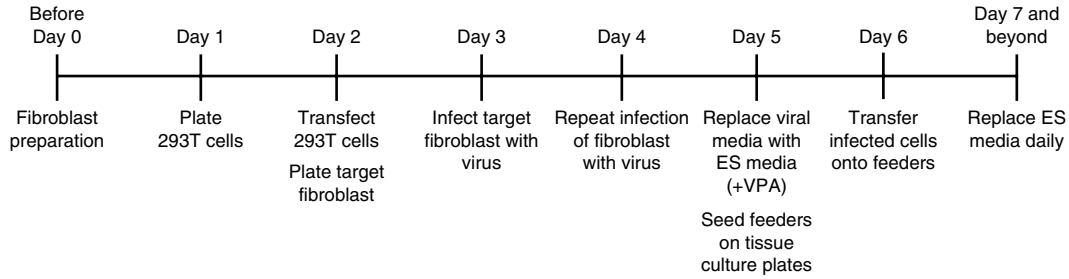


Fig. 1 A time line for reprogramming mouse embryonic fibroblast (MEF) using miRNAs

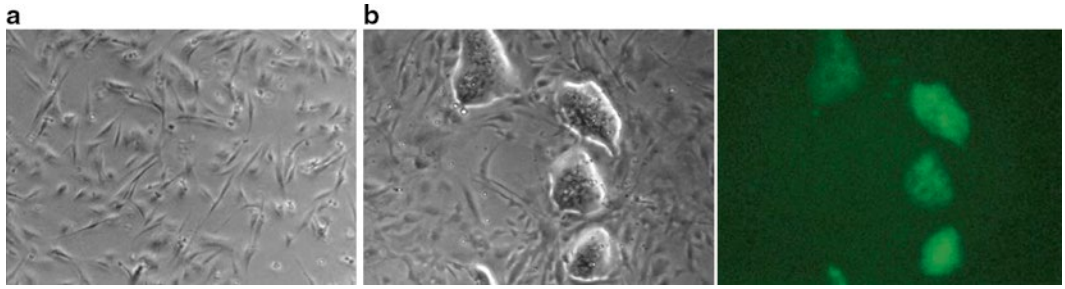


Fig. 2 Morphologies of MEF (**a**) and iPSC (**b**). The right image on panel (**b**) is a reprogrammed cell from Oct4-gfp MEF

to break loose tissues. Add 5 mL of 0.05 % trypsin/EDTA and return the tube to 37 °C and shake for an additional 15 min. Pipette 10 mL of fibroblast growth media into the tube and vigorously pipette. Allow the suspension to settle at bottom of the tube for 5 min at room temperature. Remove large chunks of tissues, count the number of cells, and plate appropriate number of cells depending on the size of Petri dish. Incubate at 37 °C with 5 % CO_2 in a humid chamber overnight. Remove undissociated tissues. Prepare freezes for storage when cells become confluent.

3.2 Plating of Target Fibroblasts to Be Reprogrammed

Fibroblast (*see* **Note 3**) must be plated on day 2 of lentiviral production (Fig. 1). This is to ensure that both fibroblast and lentivirus are ready for use at the same time.

1. Thaw a vial of mouse embryonic fibroblasts (Fig. 2a).
2. Transfer the cell suspension into a 15 mL tube containing 10 mL of fibroblast growth media.
3. Spin at 200 × *g* for 4 min.
4. Discard supernatant and resuspend cells in fibroblast growth media.
5. Plate predetermined number of fibroblasts in each well of a 6-well plate (approximately 20,000 cells).
6. Incubate at 37 °C with 5 % CO_2 overnight.

3.3 Production of Virus and Infection of Target Cells

Day 1

1. In 10 cm dish, plate 1.5×10^6 293T cells.
2. Incubate at 37 °C with 5 % CO_2 in a humid chamber overnight (*see* **Note 4**).

Day 2

1. Aliquot 0.6 mL DMEM into 1.5 mL Eppendorf tube.
2. Add 54 μL of FuGENE directly into the DMEM (*see* **Note 5**).
3. Finger-tap the tube and incubate at room temperature for 10 min.
4. On the inner surface of the tube's cap, pipette 5 μg of pMD.G, 5 μg of psPAX2, and 10 μg of miR302/367 vector plasmids.
5. Carefully cap the tube (without spilling the DNA) and mix by finger-tapping.
6. Incubate at room temperature for 15 min.
7. Replace the media on the 293T cells (plated on day 1) with fresh fibroblast growth media.
8. Add the DNA-FuGENE mixture to the 293 T cells dropwise throughout the entire plate. Rock the plate back and forth and sideways to ensure even distribution.
9. Return the plate into the incubator and allow cells to grow for 24 h.

Day 3

1. Collect the media from 293T plate and filter through a 0.45 μm filter (*see* **Note 6**).
2. Add fresh fibroblast growth media to the 293T cells and return to 37 °C and 5 % CO_2 to be used for virus collection on day 4.
3. Transfer the collected media into the Amicon centrifugal tube and spin at $3{,}600 \times g$ for 15 min.
4. Collect the concentrated viral suspension into a fresh 50 mL tube containing 10 mL of fibroblast growth media. Rinse column with same media 1–2 times to ensure you have all of the viral suspension.
5. Add 5 μL of polybrene to the viral suspension.
6. Invert the tube a couple of times.
7. Discard the fibroblast growth media on the target fibroblasts to be reprogrammed.
8. Pipette 2 mL of viral suspension on to the plated cells in 6-well plate or 10 mL in 10 cm plate.
9. Incubate at 37 °C with 5 % CO_2 overnight.

Day 4

1. Repeat steps **1**, **3**–**8** on day 3.
2. Decontaminate 293T virus-producing cells with bleach.

Day 5

1. Replace the viral media with fresh mouse ES growth media (+VPA) and incubate at 37 °C with 5 % CO_2 overnight.

3.4 Plating of Infected Cells on Feeders

Feeder cells must be plated and incubated on day 5 for use on day 6 (Fig. 1).

1. Thaw a vial of irradiated mouse feeder cells in a 37 °C bath. Decontaminate the vial with 70 % ethanol.
2. Transfer the cell suspension into a 15 mL tube containing 10 mL fibroblast growth media.
3. Spin at $200 \times g$ for 4 min.
4. Discard supernatant and resuspend cells in fibroblast growth media.
5. Count and plate 1×10^6 cells in one 10 cm dish (*see* **Note 7**).
6. Incubate at 37 °C with 5 % CO_2 overnight.

Day 6

1. Discard the fibroblast growth media on the infected target fibroblast cells and wash once with DPBS.
2. Discard the DPBS and add 200–300 μL 0.05 % trypsin/EDTA per well of a 6-well plate or 2 mL 0.05 % trypsin/EDTA into 10 cm plate. Incubate at 37 °C for 2 min and tap lightly to loosen cells.
3. Inactivate trypsin/EDTA with 2 mL ES media and spin at $200 \times g$ for 4 min.
4. Discard the media and add 10 mL of ES media (+VPA).
5. Pipette up and down a few times to dissociate cells.
6. Replace the fibroblast growth media on the feeder cells with the suspended cells in the ES growth media (+VPA).
7. Return the plates to 37 °C incubator.

Day 7 and Beyond

Change the ES growth media (+2 mM final conc. VPA) daily until colonies appear (Fig. 1). Colonies usually appear 6–8 days post-lentiviral infection. Continue culturing cells with VPA until clones are well established (approximately 2 weeks postinfection).

3.5 Picking Colonies

Allow colonies to grow big enough to be visible to the naked eyes but before visible signs of differentiation (*see* **Note 8**). Prior to the day of picking colonies, plate feeder cells in a 48-well plate.

1. Aliquot 25 μL of 0.05 % trypsin/EDTA into each well of 96-well plate.
2. Replace the ES media on the cells with DPBS.
3. Carefully use a sterilized pipette tip to draw a circle around single colony.
4. Gently nudge the colony on the sides.
5. Slowly suck the colony into the pipette tip. Make sure the colony is in the pipette tip.
6. Transfer the colony into the 0.05 % trypsin/EDTA in the 96-well plate and incubate at 37 °C.
7. Inactivate the enzyme by adding 150 μL ES media.
8. Pipette up and to dissociate the colony.
9. Aspirate the fibroblast media in the 48-well plate and fill each well with 400 μL of ES media.
10. Transfer the dissociated cells into the 48-well plate and incubate at 37 °C with 5 % CO_2 till confluence.
11. Cell can be expanded or freeze for storage.

3.6 Passaging Reprogrammed Cells

1. To passage iPS cells (Fig. 2b), aspirate ES media, gently wash with DPBS, and aspirate wash.
2. Add enough trypsin/EDTA to cover surface of the cells. The amount of the trypsin/EDTA depends on the vessels containing the cells. Incubate for 1–2 min at 37 °C or until cells loosen from dish.
3. Add 10 mL ES growth media. Triturate the cell suspension to get medium-small fragments.
4. Change ES media every day until cells become near confluent. Then repeat splitting procedures.

3.7 Thawing Reprogrammed Cells

1. Thaw cells in 37 °C water bath and transfer thawed cells to 10 mL ES growth media in a 50 mL tube.
2. Spin 200 × *g* for 4 min.
3. Aspirate and resuspend in 10 mL of fresh ES media.
4. Pipette gently with a serological pipette before plating the entire contents onto a plate of feeder cells (feeder cells should be plated a day prior).

4 Notes

1. Penicillin-streptomycin can be omitted when culturing 293T cells for lentiviral production. I am not sure of the effect of penicillin-streptomycin on viral titer.

2. Omit valproic acid when culturing fully reprogrammed cells.
3. Do not use fibroblast beyond passage 3. Fibroblast passaged more than three times senesce and stop dividing. It is advisable to use Oct4-gfp MEF for identification of fully reprogrammed cells.
4. Do not use confluent or near-confluent 293T cells for transfection. This would negatively affect the titer value of the virus.
5. Dispensing undiluted FuGENE along the wall of the tube would affect efficiency of transfection. This would result in low-quality virus.
6. Check the integrity of the filter. Use of broken filters would contaminate the target MEF with 293T cells.
7. Plate cells at different densities for easy picking of colonies.
8. To avoid picking more than one colony at a time, try picking well-spaced colonies.

Acknowledgment

This work was funded by NIH grants.

References

1. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–676
2. Maherali N, Ahfeldt T, Rigamonti A, Utikal J, Cowan C, Hochedlinger K (2008) A high-efficiency system for the generation and study of human induced pluripotent stem cells. Cell Stem Cell 3:340–345
3. Seki T, Yuasa S, Oda M, Egashira T, Yae K, Kusumoto D, Nakata H, Tohyama S, Hashimoto H, Kodaira M et al (2010) Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. Cell Stem Cell 7:11–14
4. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R et al (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318:1917–1920
5. Huangfu D, Maehr R, Guo W, Eijkelenboom A, Snitow M, Chen AE, Melton DA (2008) Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat Biotechnol 26:795–797
6. Shi Y, Desponts C, Do JT, Hahm HS, Schöler HR, Ding S (2008) Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. Cell Stem Cell 3:568–574
7. Ichida JK, Blanchard J, Lam K, Son EY, Chung JE, Egli D, Loh KM, Carter AC, Di Giorgio FP, Koszka K et al (2009) A small-molecule inhibitor of tgf-Beta signaling replaces sox2 in reprogramming by inducing nanog. Cell Stem Cell 5:491–503
8. Lyssiotis CA, Foreman RK, Staerk J, Garcia M, Mathur D, Markoulaki S, Hanna J, Lairson LL, Charette BD, Bouchez LC et al (2009) Reprogramming of murine fibroblasts to induced pluripotent stem cells with chemical complementation of Klf4. Proc Natl Acad Sci U S A 106:8912–8917
9. Zhu S, Li W, Zhou H, Wei W, Ambasudhan R, Lin T, Kim J, Zhang K, Ding S (2010) Reprogramming of human primary somatic cells by OCT4 and chemical compounds. Cell Stem Cell 7(6):651–655
10. Eminli S, Utikal J, Arnold K, Jaenisch R, Hochedlinger K (2008) Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous Sox2 expression. Stem Cell 26:2467–2474
11. Utikal J, Maherali N, Kulalert W, Hochedlinger K (2009) Sox2 is dispensable for reprogramming

of melanocytes and melanoma cells into induced pluripotent stem cells. J Cell Sci 122:3502–3510

12. Warren L, Manos PD, Ahfeldt T, Loh Y-H, Li H, Lau F, Ebina W, Mandal PK, Smith ZD, Meissner A et al (2010) Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. Cell Stem Cell 7:618–630
13. Zhou H, Wu S, Joo JY, Zhu S, Han DW, Lin T, Trauger S, Bien G, Yao S, Zhu Y et al (2009) Generation of induced pluripotent stem cells using recombinant proteins. Cell Stem Cell 4:381–384
14. Lin S-L, Chang DC, Lin C–H, Ying S-Y, Leu D, Wu DTS (2011) Regulation of somatic cell reprogramming through inducible mir-302 expression. Nucleic Acids Res 39:1054–1065
15. Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, Zhang Y, Yang W, Gruber PJ, Epstein JA et al (2011) Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell Stem Cell 8:376–388
16. Miyoshi N, Ishii H, Nagano H, Haraguchi N, Dewi DL, Kano Y, Nishikawa S, Tanemura M, Mimori K, Tanaka F et al (2011) Reprogramming of mouse and human cells to pluripotency using mature microRNAs. Cell Stem Cell 8:633–638
17. Anokye-Danso F, Snitow M, Morrisey EE (2012) How microRNAs facilitate reprogramming to pluripotency. J Cell Sci 125: 4179–4787
18. Subramanyam D, Lamouille S, Judson RL, Liu JY, Bucay N, Derynck R, Blelloch R (2011) Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. Nat Biotechnol 29:443–448

Index

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6, © Springer Science+Business Media New York 2014

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W