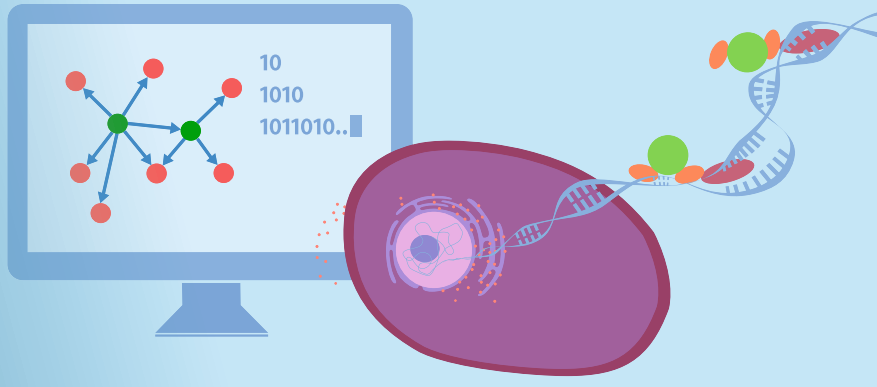
Methods in
Molecular Biology 2594

Springer Protocols



Transcription Factor Regulatory Networks

Humana Press

METHODS IN MOLECULAR BIOLOGY

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

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

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

Transcription Factor Regulatory Networks

Edited by

Qi Song

Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA

Zhipeng Tao

Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA

Humana Press

Editors
Qi Song
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA

Zhipeng Tao
Cutaneous Biology Research Center
Massachusetts General Hospital, Harvard
Medical School
Charlestown, MA, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-2814-0 ISBN 978-1-0716-2815-7 (eBook)
https://doi.org/10.1007/978-1-0716-2815-7

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

Preface

Transcription factors (TFs) are the major modulators of gene regulation in living cells. TFs regulate gene expression by preferentially binding to specific genomic regions on the chromatin. Systematic approaches are needed to characterize regulatory events at genomic scale. In recent years, a myriad of experimental and computational approaches have been developed to investigate complex associations between TFs and targeted genomic regions. This book covers various state-of-the-art techniques regarding the TFs-genes associations, with a focus on providing methodological and practical references for researchers. In this volume, we aim to cover diverse protocols and summaries of TFs including screening of TF-DNA interactions, detection of open chromatin regions, identification of epigenetic regulations, engineering TFs with genome editing tools, detection of transcriptional activities, and computational analysis of TF networks. Additionally, several chapters focused on presenting a comprehensive review of functions and druggabilities of TFs in biomedical research, such as cell fate reprogramming and development of molecular drugs. With the emerging interests of single-cell techniques, several chapters focus on applying single-cell-based approaches to promote the studies of TFs. We are hoping that this book will benefit readers who are interested in using state-of-the-art techniques to study TFs, and this volume will serve as a step-by-step protocol for performing experiments and troubleshooting in their studies.

Pittsburgh, PA, USA — *Qi Song*
Charlestown, MA, USA — *Zhipeng Tao*

Contents

Contributors

Peter D. Adams • *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA*

Zhichao Ai • *Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK*

Bastiaan O. R. Bargmann • *School of Plant and Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA*

Julio Aguila Benitez • *Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden*

Matthew D. Brooks • *Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA; USDA ARS Global Change and Photosynthesis Research Unit, Urbana, IL, USA*

Le Chang • *Department of Human Genetics, McGill University, Montreal, QC, Canada*

Wen-Chi Chang • *Institute of Tropical Plant Sciences and Microbiology, National Cheng Kung University, Tainan, Taiwan*

Zhiyong Cheng • *Food Science and Human Nutrition Department, University of Florida, Gainesville, FL, USA; Department of Human Nutrition, Foods, and Exercise, Virginia Tech, Blacksburg, VA, USA*

Chi-Nga Chow • *Institute of Tropical Plant Sciences and Microbiology, National Cheng Kung University, Tainan, Taiwan*

George M. Church • *Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA, USA; Department of Genetics, Harvard Medical School, Boston, MA, USA*

Gloria M. Coruzzi • *Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA*

Anagha Deshpande • *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA*

Aniruddha J. Deshpande • *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA*

Darren Finlay • *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA*

Haydy George • *GC Therapeutics Inc., Cambridge, MA, USA; School of Medicine, St. George's University, Grenada, West Indies*

Eva Hedlund • *Department for Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden; Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden*

Byungjin Hwang • *Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, South Korea*

Takashi Ito • *Department of Biochemistry, Kyushu University Graduate School of Medical Sciences, Fukuoka, Japan*

Jiming Jiang • *Department of Plant Biology, Michigan State University, East Lansing, MI, USA; Department of Horticulture, Michigan State University, East Lansing, MI, USA*

Parastoo Khoshakhlagh • *GC Therapeutics Inc., Cambridge, MA, USA*

Gabriel Krouk • *BPMP, Univ Montpellier, CNRS, INRA, SupAgro, Montpellier, France*

Emal Lesha • *GC Therapeutics Inc., Cambridge, MA, USA; Department of Neurosurgery, University of Tennessee Health Science Center, Memphis, TN, USA*

Qiang Li • *Hunan Provincial Key Laboratory of Forestry Biotechnology & International Cooperation Base of Science and Technology Innovation on Forest Resource Biotechnology, Central South University of Forestry and Technology, Changsha, China; Microbial Variety Creation Center, National Laboratory of Yuelushan Seed Industry, Changsha, China*

Sha Li • *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA*

Song Li • *School of Plant and Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA*

Xinxu Li • *State Key Laboratory for Crop Genetics and Germplasm Enhancement, JCIC-MCP, CIC-MCP, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China*

Gao-Qiang Liu • *Hunan Provincial Key Laboratory of Forestry Biotechnology & International Cooperation Base of Science and Technology Innovation on Forest Resource Biotechnology, Central South University of Forestry and Technology, Changsha, China; Microbial Variety Creation Center, National Laboratory of Yuelushan Seed Industry, Changsha, China*

Jiang-Shan Ma • *Hunan Provincial Key Laboratory of Forestry Biotechnology & International Cooperation Base of Science and Technology Innovation on Forest Resource Biotechnology, Central South University of Forestry and Technology, Changsha, China; Microbial Variety Creation Center, National Laboratory of Yuelushan Seed Industry, Changsha, China*

James R. Mason • *Scripps MD Anderson Cancer Center, La Jolla, CA, USA*

Fumihito Miura • *Department of Biochemistry, Kyushu University Graduate School of Medical Sciences, Fukuoka, Japan*

Alex H. M. Ng • *GC Therapeutics Inc., Cambridge, MA, USA*

Jik Nijssen • *Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden*

Emily R. Novak • *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA*

Kelsey M. Reed • *School of Plant and Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA*

Julia Salafranca • *Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK*

Christoph Schweingruber • *Department for Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden; Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden*

Limin Shi • *Food Science and Human Nutrition Department, University of Florida, Gainesville, FL, USA*

Cory J. Smith • *GC Therapeutics Inc., Cambridge, MA, USA*

Qi Song • *Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA; Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA*

Zhipeng Tao • *Department of Human Nutrition, Foods, and Exercise, Virginia Tech, Blacksburg, VA, USA; Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA*

Irina A. Udalova • *Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK*

Erinke van Grinsven • *Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK*

Lihui Wang • *Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK*

Qiang Wu • *College of Food and Chemical Engineering, Shaoyang University, Shaoyang, China; Hunan Provincial Key Laboratory of Forestry Biotechnology & International*

Cooperation Base of Science and Technology Innovation on Forest Resource Biotechnology, Central South University of Forestry and Technology, Changsha, China; Microbial Variety Creation Center, National Laboratory of Yuelushan Seed Industry, Changsha, China

Xu Wu • *Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA*

Jianguo Xia • *Department of Human Genetics, McGill University, Montreal, QC, Canada; Institute of Parasitology, McGill University, Montreal, QC, Canada; Department of Animal Science, McGill University, Montreal, QC, Canada*

Nan Cher Yeo • *Precision Medicine Institute, University of Alabama-Birmingham, Birmingham, AL, USA; Department of Pharmacology and Toxicology, University of Alabama-Birmingham, Birmingham, AL, USA*

Mark M. Zaki • *GC Therapeutics Inc., Cambridge, MA, USA; Department of Neurosurgery, University of Michigan, Ann Arbor, MI, USA*

Aicen Zhang • *State Key Laboratory for Crop Genetics and Germplasm Enhancement, JCIC-MCP, CIC-MCP, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China*

Wenli Zhang • *State Key Laboratory for Crop Genetics and Germplasm Enhancement, JCIC-MCP, CIC-MCP, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China*

Hainan Zhao • *Department of Plant Biology, Michigan State University, East Lansing, MI, USA*

Chapter 1

The TARGET System: Rapid Identification of Direct Targets of Transcription Factors by Gene Regulation in Plant Cells

Matthew D. Brooks, Kelsey M. Reed, Gabriel Krouk, Gloria M. Coruzzi, and Bastiaan O. R. Bargmann

Abstract

The TARGET system allows for the rapid identification of direct regulated gene targets of transcription factors (TFs). It employs the transient transformation of plant protoplasts with inducible nuclear entry of the TF and subsequent transcriptomic and/or ChIP-seq analysis. The ability to separate direct TF–target gene regulatory interactions from indirect downstream responses and the significantly shorter amount of time required to perform the assay, compared to the generation of transgenics, make this plant cell-based approach a valuable tool for a higher throughput approach to identify the genome-wide targets of multiple TFs, to build validated transcriptional networks in plants. Here, we describe the use of the TARGET system in *Arabidopsis* seedling root protoplasts to map the gene regulatory network downstream of transcription factors-of-interest.

Key words Transcription factor, Gene regulatory network, Protoplast, Fluorescence activated cell sorting, RNA-seq

1 Introduction

The TARGET system (transient assay reporting genome-wide effects of transcription factors) was developed to enable the rapid identification of genes directly regulated by a transcription factor (TF) of interest [1]. The plant cell-based system makes use of transient transformation of isolated plant cell protoplasts [2] with a vector (pBeacon_GR vector series, Fig. 1) containing a TF fused to the glucocorticoid receptor (GR). The GR-TF fusion protein is held in the cytoplasm by association with a heat shock protein (HSP) complex. The addition of the dexamethasone ligand displaces the GR–HSP association, which makes possible the induced nuclear translocation of the TF [3]. Concurrent application of the translational inhibitor cycloheximide ensures that only direct transcriptional targets are affected by the TF nuclear import and

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_1,

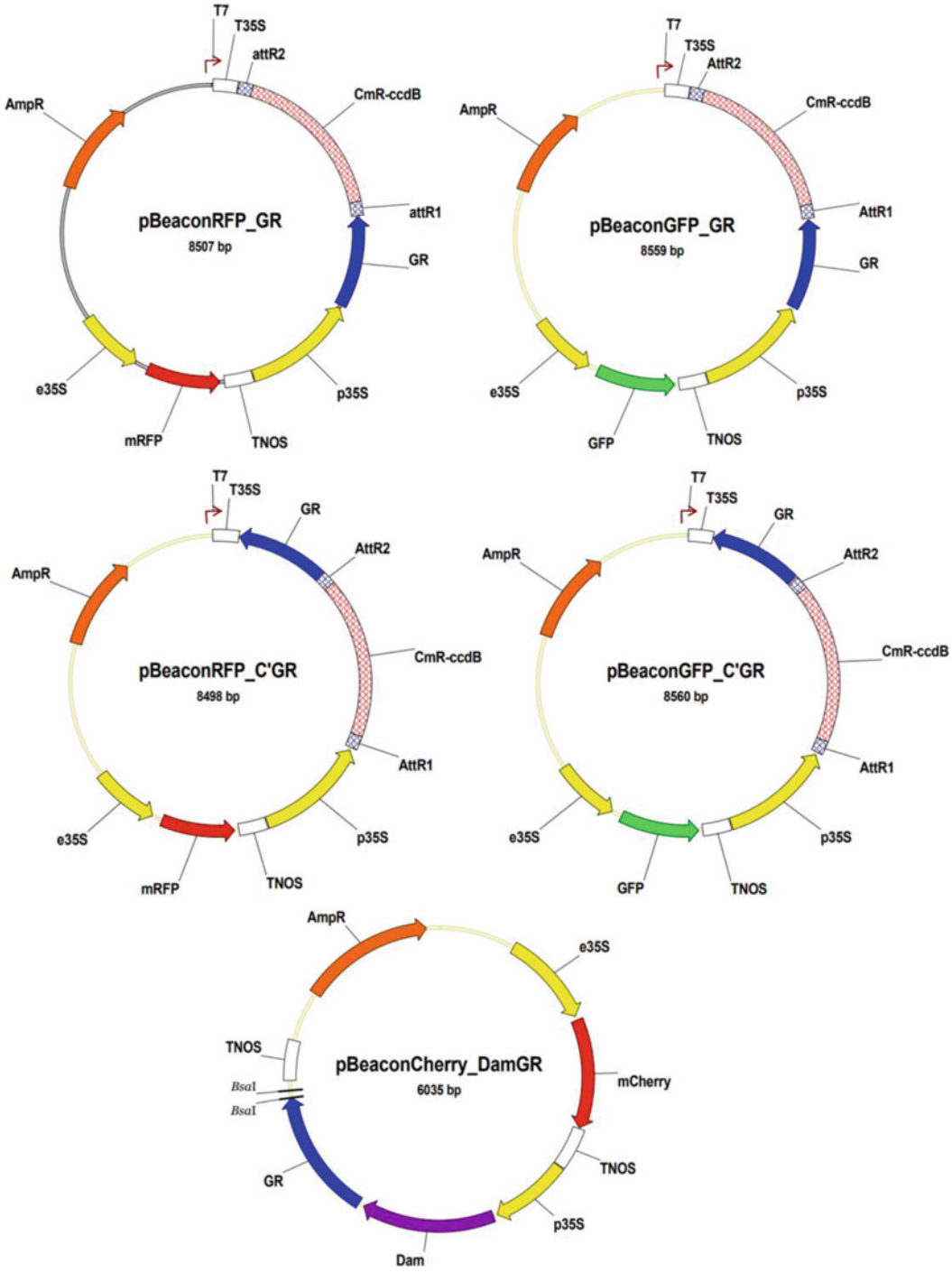


Fig. 1 The pBeacon vector series for use in the TARGET system. pBeaconRFP_GR, pBeaconGFP_GR, pBeaconRFP_C'GR, and pBeaconGFP_C'GR. These gateway-compatible vectors vary in harboring the mRFP (monomeric Red Fluorescent Protein)- or GFP (Green Fluorescent Protein)-positive fluorescent selection markers as well as either an N- or C-terminal fusion with the glucocorticoid receptor (GR). The Goldengate-compatible pBeaconCherry_DamGR contains an mCherry-positive selection marker and has an N-terminal fusion with DNA adenine methyltransferase (Dam) and GR. The TARGET vectors and full sequences are available from the VIB-UGent Gateway collection (https:/gatewayvectors.vib.be/) (see Table 1)

precludes the action of secondary TFs that may be regulated by the primary TF. In addition, the pBeacon_GR vectors make use of positive fluorescent selection, which allows for the use of fluorescence activated cell sorting (FACS) to isolate the successfully transformed plant cells. This leads to reduction of transcriptomic background noise which aids in the detection of differentially expressed genes [4, 5].

TARGET has been used to study numerous TFs, either individually or in groups [1, 6–13]. While this methods paper is focused on conducting TARGET in seedling root protoplasts, the assay has also been adapted to shoot protoplasts [10]. In addition, a new version of the TARGET assay that uses two different vectors (pBeaconRFP_GR and pBeaconGFP_GR) has led to a higher throughput TARGET assay of up to 24 TF assays/cycle [11]. Furthermore, the TARGET assay has been used in conjunction with other techniques for characterization of DNA binding, including chromatin-immunoprecipitation (ChIP-seq), DNA adenine methyltransferase identification (DamID-seq), or capture by 4tU-affinity labeling of TF-regulated nascent mRNAs [6, 12, 14, 15]. Additionally, ~50 TFs assayed in TARGET can now be integrated with a large collection of published TF–target datasets housed in the new ConnecTF software platform and database (https://connectf.org), which enables analyses, refinement, and visualization of extensive gene regulatory networks and their potential physiological relevance [13].

Aside from its use in fully sequenced model plants, like *Arabidopsis*, the TARGET system can potentially also be used in crop species as well as in species where the genome sequence is unavailable or incomplete. Furthermore, the TARGET approach, which uses transient transformation of plant cell protoplasts, may be applicable to species where transgenic approaches to study transcriptional regulation are not feasible. Importantly, the TARGET system can be deployed in a much shorter time-frame and at higher throughput compared to transgenic plant approaches.

In this chapter, we describe the use of the TARGET system (Fig. 2) in *Arabidopsis* seedling root protoplasts, reviewing protoplast isolation, transient transformation, and treatment, as well as transcript analysis. Of note, these techniques can be used in other plant tissues such as shoot protoplast [10] and other species with minor modifications.

2 Materials

Prepare all solutions using ultra-pure water (prepared by purifying deionized water, to attain a sensitivity of 18 MΩ/cm at 25 °C) and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise). Diligently follow all regulations when disposing of waste materials.

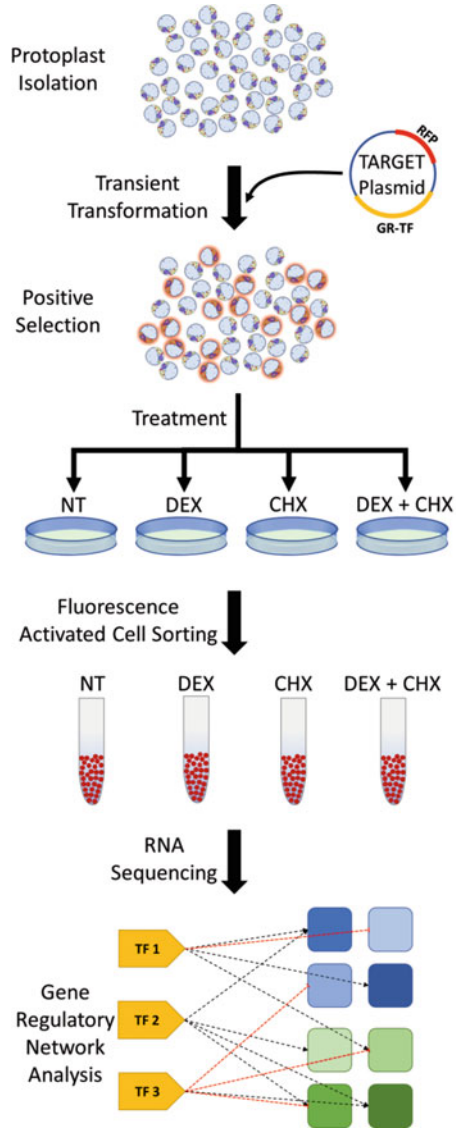


Fig. 2 Schematic overview of the TARGET system. NT, not treated; DEX, dexamethasone; CHX, cycloheximide

2.1 Plant Growth

1. Plant growth medium: 2.2 g/L MS salts + vitamins (Murashige and Skoog [16]), 1% w/v sucrose, 1% w/v agar. Adjust to pH 5.8 with KOH. Autoclave at 120 °C to sterilize. Pour into 10 × 10 cm square plates. Store at 4 °C.
2. 70% v/v ethanol, 20% v/v bleach (1.21% w/v sodium hypochlorite), and sterile water.
3. 100-μm nylon mesh (Sefar Filtration, USA), sterile transfer pipettes, ¼″ micropore tape.
4. Plant growth chamber, 100-μmol m^{-2} s^{-1} PAR and 22 °C with 18 h light/6 h dark.

2.2 Generation of Root Cell Protoplasts

1. Enzyme solution: 1.25% w/v Cellulase R-10 (Kanematsu USA), 0.3% w/v Macerozyme R-10 (Kanematsu USA), 0.4-M mannitol, 20-mM MES (2-(N-morpholino)ethanesulfonic acid), 20-mM potassium chloride, 0.1% w/v bovine serum albumin, 10-mM calcium chloride, 5-mM β-mercaptoethanol. Adjust to pH 5.7 with 1-M TRIS hydrochloride pH 7.5. Heat the solution to 55 °C for 10 min (the solution should become clear), and cool to room temperature. Store at −20 °C. Filter sterilize with 0.22-μm filters (*see* **Notes 1** and **2**).
2. Scalpel, rotary shaker, 250-mL flasks, 40-μm cell strainer (BD Falcon, USA), 15-mL conical tubes, swing-bucket centrifuge (500*G*), hemacytometer.

2.3 Transformation of Root Cell Protoplasts

1. Plasmids: pBeaconRFP_GR, pBeaconGFP_GR, pBeaconRFP_C'GR, pBeaconGFP_C'GR, pBeaconCherry_DamGR (Table 1, Fig. 1).
2. MMg solution: 0.4-M mannitol, 15-mM magnesium chloride hexahydrate, 4-mM MES. Adjust to pH 5.7 with 1-M potassium hydroxide solution. Make fresh. Filter sterilize with 0.22-μm filters.
3. MIDIprep kit (QIAGEN, USA). Store DNA at −20 °C (*see* **Note 3**).
4. PEG solution: 40% w/v polyethylene glycol 4000, 0.4-M mannitol, 0.1-M calcium chloride. Make fresh. Filter sterilize with 0.22-μm filters (*see* **Note 4**).

Table 1
Vectors compatible with the TARGET system

Vector name	Alternate names	Description	Ref.	Availability
pBeaconRFP_GR	pBOB11	RFP-positive selection, N-terminal GR fusion, gateway-compatible	[1]	VIB-UGent
pBeaconGFP_GR		GFP-positive selection, N-terminal GR fusion, gateway-compatible	[11]	VIB-UGent
pBeaconRFP_C'GR	pBOB11_C-Term	RFP-positive selection, C-terminal GR fusion, gateway-compatible	[12]	VIB-UGent
pBeaconGFP_C'GR		GFP-positive selection, C-terminal GR fusion, gateway-compatible	This work	To be deposited VIB-UGent
pBeaconCherry_DamGR	pDamBOB	mCherry-positive selection, N-terminal Dam-GR fusion, BsaI (Goldengate) cloning	[12]	VIB-UGent

5. W5 solution: 154-mM sodium chloride, 125-mM calcium chloride, 5-mM potassium chloride, 5-mM MES. Adjust to pH 5.7 with 1-M potassium hydroxide solution. Store at room temperature. Autoclave or filter sterilize with 0.22-μm filters.
6. 24-well plates, epifluorescence microscope equipped with GFP and RFP (or equivalent) filters.

2.4 Treatment of Protoplasts

1. Dexamethasone: 10-mM stock dissolved in 96% v/v ethanol. Store at −20 °C.
2. Cycloheximide: 35-mM stock dissolved in 96% v/v ethanol. Store at −20 °C.

2.5 Sorting of Protoplasts

1. FACSAria (BD, USA) or equivalent sorter with PBS (phosphate-buffered saline) as a sheath fluid and a 100-μm nozzle.
2. RNeasy micro kit (QIAGEN, USA).

2.6 Transcript Analysis

1. Dynabeads mRNA Purification Kit (Invitrogen/Thermo Fisher, USA).
2. NEBNext Ultra II RNA Library Prep Kit (New England BioLabs, USA).

3 Methods

Carry out all procedures at room temperature unless otherwise specified. Use aseptic techniques for all procedures (*see* **Note 2**).

3.1 Plant Growth

1. Sterilize 1-mL of dry *Arabidopsis* seed (approximately 35×10^3 seeds for Col-0) in a 50-ml tube by a 5-min incubation with 10-mL 70% ethanol, followed by a 10-min incubation with 10-mL 20% bleach, and rinse three times with 50-mL sterile water.
2. Plate seeds in two rows on top of 100-μm nylon mesh in square plates with plant growth medium using a sterile transfer pipette (Fig. 3). 1-mL of seed is divided over ten plates. Plates are sealed with micropore tape and placed vertically in a growth chamber.

3.2 Generation of Protoplasts

1. Harvest the roots of 7- to 10-day-old seedlings using a scalpel, and transfer to a 250-mL flask with 50-mL enzyme solution.
2. Shake the flask with the roots in enzyme solution at 75 rpm at room temperature for 3 h.
3. Filter the protoplasts by passing them over a 40-μm cell strainer into a fresh flask (*see* **Note 5**).

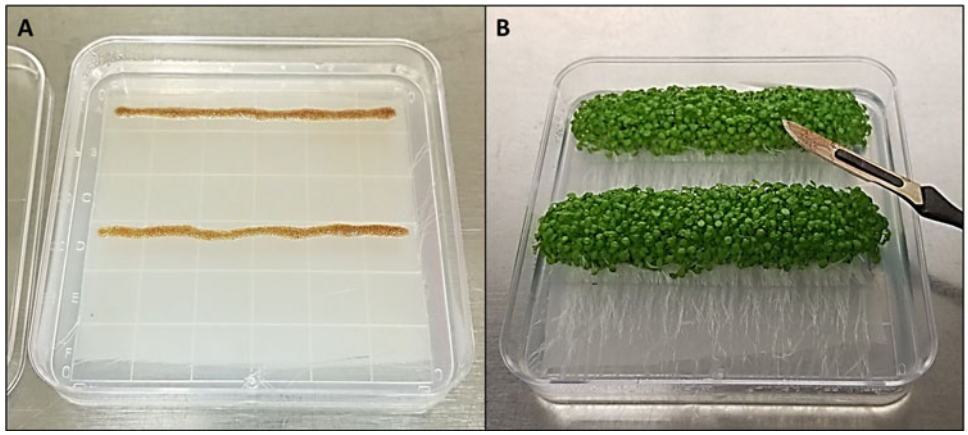


Fig. 3 *Arabidopsis* seedling growth for harvesting of roots. (**a**) Sterilized seeds are plated on top of 100-μm nylon mesh in a 10 × 10 cm square plate with solidified plant growth medium. The plates are placed vertically in a growth chamber. (**b**) After 1 week, the roots are harvested using a scalpel and used for the preparation of protoplasts. NB. The TARGET assay can also be performed on shoots [10]

4. Transfer the filtered protoplast suspension to 15-mL conical tubes, and spin them down for 5 min at 500*G* with no brake. Remove the supernatant by aspirating (*see* **Note 6**).

3.3 Transformation of Protoplasts

1. Wash the protoplasts by resuspending the protoplast pellet in 15-mL MMg solution and spinning them down for 5 min at 500*G* with no brake. Remove the supernatant by aspirating.
2. Resuspend the protoplasts in a volume appropriate for quantification of protoplast density using a hemacytometer. Assess the protoplast density, and adjust the volume of the MMg solution to achieve a density of 4×10^6 protoplasts per ml (*see* **Note 7**).
3. Prepare a 15-mL conical tube for each transformation (and one mock transformation with no DNA) by labeling them and adding 50-μg plasmid DNA to the bottom of the tube.
4. Add 250-μl (1×10^6) protoplasts to each tube. Add 250-μl PEG solution to each tube and mix well by vortexing for 5 s (*see* **Notes 8** and **9**).
5. Wash the protoplasts by adding 15-ml W5 solution to each tube and spinning them down for 5 min at 500*G* with no brake. Remove the supernatant by aspirating (*see* **Note 10**).
6. Resuspend the protoplasts in 1 mL W5 and transfer them to a 24-well plate.
7. Incubate the protoplasts overnight at room temperature in the dark while shaking at 50 rpm.
8. Inspect the protoplasts with an epifluorescence microscope equipped with GFP and RFP (or equivalent) filters to check for successful transformation (*see* **Note 11**).

3.4 Treatment of Root Cell Protoplasts

1. Divide each independent transformation over 4 wells; for mock, dexamethasone, cycloheximide, and dexamethasone + cycloheximide treatment (*see* **Notes 12–15**).
2. Treat the protoplasts with 35-μM cycloheximide and/or 10-μM dexamethasone. Cycloheximide is administered with a 20-min pretreatment, and the subsequent dexamethasone treatment is incubated for 3 h. Stagger the start of treatments to account for the time it takes to sort individual samples (5–15 min).

3.5 Sorting of Protoplasts

1. Set up the FACS with PBS as a sheath fluid and a 100-μm nozzle (*see* **Note 10**).
2. Set up a dotplot for green (GFP) (488-nm excitation, 530/30 emission) vs. red (RFP) (561-nm excitation, 583/30 emission) fluorescence emission (*see* **Note 16**).
3. Use the mock-transformed protoplasts to set up gates for RFP- and GFP-positive cells (Fig. 4) (*see* **Note 17**).
4. Sort 20×10^3 protoplasts into 350-μL RNA extraction buffer (RLT). Freeze samples (−20 °C) upon completion of the sort.
5. Extract the RNA according to the manufacturer's instructions, eluting with 50 μL of nuclease-free H_2O (*see* **Note 18**).

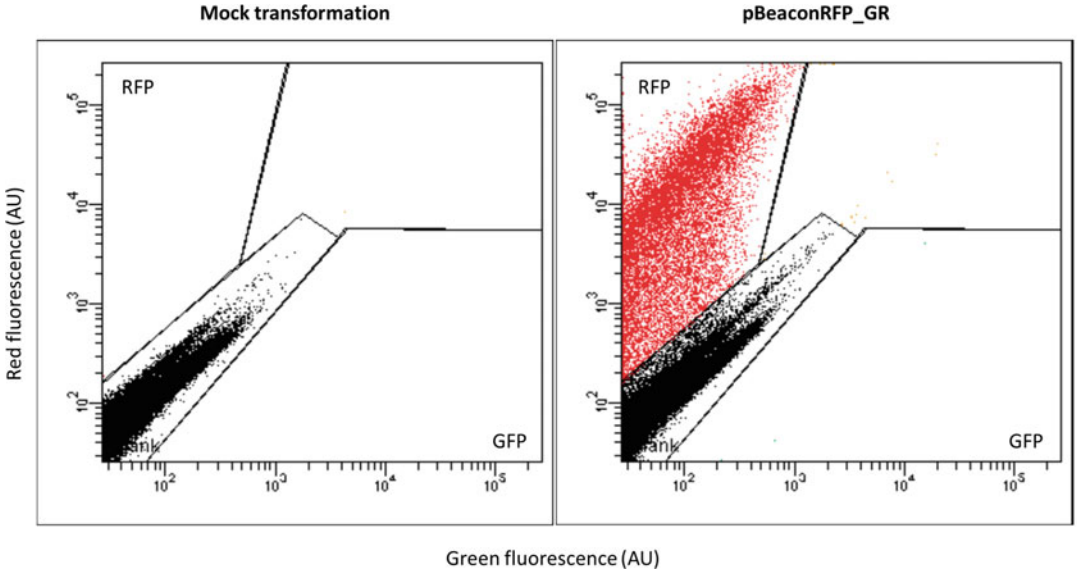


Fig. 4 Cytometric FACS analysis of TARGET vector transformed root cell protoplasts. Dotplots representing a mock transformation and protoplasts successfully transformed with pBeaconRFP_GR vector are shown. Red fluorescence is indicated on the *Y*-axis (in arbitrary units) and green fluorescence on the *X*-axis. Each dot represents an individual detection event. This data was generated with a FACSAria cell sorter

3.6 Transcript Analysis

1. Perform polyA mRNA purification from total RNA using the Dynabeads mRNA Purification Kit. For total RNA from 20×10^3 cells, we typically use 15 μL of Oligo $(dT)_{25}$ beads per sample, combining the beads needed for all samples in one tube for the initial bead wash step. Beads are resuspended in binding buffer, and an equal volume (50 μL) is added to each total RNA sample.
2. Make sequencing libraries according to manufacturer's instructions using the NEBNext Ultra II RNA Library Prep Kit (*see* **Note 19**).
3. Pool and sequence samples on an Illumina sequencer according to the manufacturer's specifications for single-end reads. Read lengths of 50 bp are sufficient to identify differentially expressed genes in *Arabidopsis*. Aim for between five and ten million reads per library (*see* **Note 20**).
4. Remove reads and bases of low quality, and remove adapter sequences from raw reads. Align FASTQ files to the *Arabidopsis* genome (TAIR10) using a short-read mapping package (e.g., HISAT2 [17] or STAR [18]) (*see* **Note 21**).
5. Quantify read counts per gene using a program such as HTSeq-count [19] or featureCounts [20].
6. Identify genes that are differentially expressed in response to dexamethasone using DESeq2 [21] or another appropriate package for transcriptome analysis. Genes that respond to dexamethasone in the absence of cycloheximide include both indirect and direct target genes of the TF, while genes that respond to dexamethasone in the presence of cycloheximide are direct TF target genes (*see* **Note 15**).
7. TF–target genes can be compared to published TF–target gene interactions using the ConnecTF platform [13] either via the public website or setting up a private instance of the tool.

4 Notes

1. The use of β-mercaptoethanol is optional. It has been reported to improve yield and/or viability, but we have found no significant effect.
2. These experiments can be conducted under non-sterile conditions, since the experiment only takes 24–36 h. This may speed up the process. However, for reproducibility, and especially if you are studying factors involved in plant pathogenesis responses, it is recommended to perform the experiment using aseptic techniques. The enzyme solution may clog the 0.22-μm filters and require several extra filters to be used.

3. Resuspend the DNA pellet in a small volume of sterile de-ionized water (50 μL). Alternatively, use 10-mM Tris–HCl pH 8.0 (DNA pellets may dissolve better in slightly buffered water, and note that EDTA inhibits transfections and it is in almost all kit elution buffers (TE)). It is critical to get a concentration between 1 and 4 μg/μL. Be sure to dilute plasmid DNA at least ten-fold before measuring concentration for accuracy. Other plasmid purification kits have been successfully used (e.g., ZymoPURE Plasmid Midiprep or Maxiprep kits (Zymo Research), PureLink™ HiPure Expi Plasmid Megaprep kit (Invitrogen)); however, the quality of the DNA is important, and other purification methods/kits may need to be tested on a small scale first.
4. PEG of differing average molecular weight has been used successfully in our hands (1500–8000). Due to its viscosity, filter sterilization of the PEG solution may take more time.
5. Cell strainers can be washed, sterilized with 70% ethanol, and reused.
6. The use of 15-mL conical tubes ensures the formation of a compact pellet and prevents loss of protoplasts. Similarly, the use of centrifugation with no (or minimal) braking ensures maximum recovery of protoplasts.
7. Be gentle when resuspending the protoplasts. Large orifice pipette tips can be used to minimize shearing forces. A viability stain (e.g., fluorescein diacetate) can be used for a more accurate count of live protoplasts.
8. Plasmid preps that contain too much salt will cause a precipitation of the DNA when mixed with the PEG solution.
9. Traditionally, the protoplast/PEG/DNA solution was incubated at room temperature for 15 min, but we have found that brief vortexing at moderate speed and immediate washing can work better and save time.
10. W5 solution contains relatively high levels of calcium chloride which can, in the case of some cytometers, cause issues due to precipitation with the phosphate in the PBS sheath fluid that lead to clogging. An alternative is to use the enzyme solution base (enzyme solution without the enzymes added) as a wash and incubation solution. Another option is to use plain saline solution (or filtered tap water) as the sheath fluid.
11. We generally see protoplast transformation efficiencies ranging between 5% and 20% using *Arabidopsis* seedling root protoplasts.
12. For statistical analysis, we recommend a minimum of three independent transformations per tested TF.

13. In order to minimize the number of RNA-sequencing samples, the dexamethasone alone and mock treatments can be omitted. Include a control with an empty vector (e.g., pBeaconRFP_GR) or one with an insert without transcriptional activity (e.g., pBeaconRFP_GR-GUS).
14. For scaled-up TARGET assays (e.g., [11]), two separate protoplast transformations (one with a TF in pBeaconRFP_GR and one with another TF in pBeaconGFP_GR) can be combined and treated and sorted in unison. In this scaled-up assay, comparison of TF-transfected cells to Empty Vector—all in +DEX only—allows one to perform 24 TF in one assay. Vectors are available from the VIB-UGent Gateway vector collection (https://gatewayvectors.vib.be/).
15. If the experimental design includes an empty vector or GUS control in lieu of a combination of treatments, the model design for identifying TF regulated targets is only the vector used (e.g., empty vector vs. TF). When multiple TFs are examined across multiple days with the same treatments, the inclusion of the control allows the data to be analyzed simultaneously by including a Batch factor (e.g., TF + Batch) in the model, as in Brooks et al. [11].
16. We have successfully used 488-nm excitation for RFP detection. The excitation and emission frequencies listed here are not strictly the only ones that can be used.
17. A detailed methodology for setting up a FACS for sorting protoplasts is available [5].
18. We generally get about 85-ng of total RNA from 20×10^3 root cells. A Bioanalyzer (Agilent, USA) with a 6000 RNA Pico Kit can be used to accurately assess RNA quantity and quality.
19. Library preparation kits from other manufactures can also be used, but ensure that they are optimized for low RNA input. Kits designed for 3′ RNA sequencing can be used and do not require the purification of mRNA from total RNA (Subheading 3.6, **step 1**).
20. If using a kit designed for 3′ RNA sequencing, the number of reads required to identify differentially expressed genes will be lower, between three and six million reads per sample, and allows pooling of more samples per run.
21. Standard quality control should be performed on each step in the bioinformatics analysis, beginning with the raw sequence data using a program such as MultiQC [22]. It should be noted that sequencing results from TARGET experiments often contain a large percentage of duplicate reads.

Acknowledgments

This work was supported by NIH Grant R01-GM121753 to G.M. C., NIH NIGMS Fellowship F32GM116347 to M.D.B., and USDA-NIFA Hatch VA-160133 and Multistate VA-136377 projects to B.O.R.B.

References

1. Bargmann BO, Marshall-Colon A, Efroni I et al (2013) TARGET: a transient transformation system for genome-wide transcription factor target discovery. Mol Plant 6:978–980
2. Yoo SD, Cho YH, Sheen J (2007) Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat Protoc 2:1565–1572
3. Schena M, Yamamoto KR (1988) Mammalian glucocorticoid receptor derivatives enhance transcription in yeast. Science 241:965–967
4. Bargmann BO, Birnbaum KD (2009) Positive fluorescent selection permits precise, rapid, and in-depth overexpression analysis in plant protoplasts. Plant Physiol 149:1231–1239
5. Bargmann BO, Birnbaum KD (2010) Fluorescence activated cell sorting of plant protoplasts. J Vis Exp. https://doi.org/10.3791/1673
6. Para A, Li Y, Marshall-Colón A et al (2014) Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in Arabidopsis. Proc Natl Acad Sci 111:10371–10376
7. Sparks EE, Drapek C, Gaudinier A et al (2016) Establishment of expression in the SHORTROOT-SCARECROW transcriptional cascade through opposing activities of both activators and repressors. Dev Cell 39:585–596
8. Medici A, Marshall-Colon A, Ronzier E et al (2015) AtNIGT1/HRS1 integrates nitrate and phosphate signals at the Arabidopsis root tip. Nat Commun 6:6274
9. Safi A, Medici A, Szponarski W et al (2021) GARP transcription factors repress Arabidopsis nitrogen starvation response via ROS-dependent and -independent pathways. J Exp Bot. https://doi.org/10.1093/jxb/erab114
10. Varala K, Marshall-Colón A, Cirrone J et al (2018) Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. Proc Natl Acad Sci 115:6494–6499
11. Brooks MD, Cirrone J, Pasquino AV et al (2019) Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. Nat Commun 10:1569
12. Alvarez JM, Schinke A-L, Brooks MD et al (2020) Transient genome-wide interactions of the master transcription factor NLP7 initiate a rapid nitrogen-response cascade. Nat Commun 11:1–13
13. Brooks MD, Juang C-L, Katari MS et al (2021) ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks. Plant Physiol 185:49–66. https://doi.org/10.1093/plphys/kiaa012
14. Doidy J, Li Y, Neymotin B et al (2016) "Hit-and-Run" transcription: de novo transcription initiated by a transient bZIP1 "hit" persists after the "run". BMC Genomics 17:92
15. Para A, Li Y, Coruzzi GM (2018) μChIP-Seq for genome-wide mapping of in vivo TF-DNA interactions in Arabidopsis root protoplasts. In: Ristova D, Barbez E (eds) Root development: methods and protocols. Springer, New York, pp 249–261
16. Murashige T, Skoog F (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol Plant 15:473–497
17. Kim D, Paggi JM, Park C et al (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915
18. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21
19. Anders S, Pyl PT, Huber W (2015) HTSeq – a python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169
20. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930
21. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550
22. Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048

Chapter 2

The Method of Screening and Identification of Transcription Factor in *Klebsiella*

Qiang Wu, Gao-Qiang Liu, Jiang-Shan Ma, and Qiang Li

Abstract

This protocol describes a method for verifying the specific transcription factor regulating glycerol dehydratase (GDH) expression in *Klebsiella*. DNA pull-down accompanied with mass spectrometry is used to screen and identify the transcription factor interacting with the promoter region of the key gene in *Klebsiella*. EMSA method is used to validate the specific binding of the transcription factor to the promoter region in vitro. In addition, the target DNA fragments are constructed by fusion PCR to prepare competent cells from *Klebsiella* for electrical transformation and further transformed to obtain key gene deletion strains to verify the transcription factor responsible for the target gene expression in *Klebsiella*.

Key words Transcription factor, DNA pull-down, EMSA, Gene knockout

1 Introduction

DNA pull-down is a powerful tool to study the interaction between DNA and transcription factor in vitro. During DNA pull-down experiment, biotin-labeled DNA fragments are bound to streptavidin magnetic beads and then incubated with the nuclear protein to purify the transcription factor interacting with DNA fragment [1]. The target transcription factor obtained by washing and elution is detected by Western blot to verify specificity between transcription factor and DNA fragment. Mass spectrometry can be used to identify the nominated transcription factors that may interact with DNA fragments [2]. The interaction between biotin and streptavidin is the most intense non-covalent interaction known [3]. The activated biotin can be conjugated with almost all known biological macromolecules mediated by protein crosslinking agents [4]. Therefore, after labeling DNA with biotin, various biomolecular complexes can be purified with DNA. Pull-down experiment is generally used for in vitro transcription or translation systems, but it cannot truly reflect their interaction, because they may not combine

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_2,

spatially under physiological conditions in vivo. Therefore, other methods are needed to verify the results of DNA pull-down. Electrophoretic mobility assay (EMSA) is a technique for studying the interaction between DNA-binding protein and the associated DNA-binding sequence and can be used for qualitative and quantitative analysis [5]. EMSA technique has been used to study the interactions between DNA-binding proteins and their associated DNA-binding sequences, for DNA qualitative and quantitative analysis, and the interaction between RNA binding protein and specific RNA sequence [6].

Gene knockout is the most direct method to study gene function. The original RecA system was used generally in the traditional method of gene knockout, depending on specific restriction endonuclease restriction sites, and requires homologous arm length to operate as complex hybridization [7, 8]. The emergence of Red homologous recombination method makes genome modification more rapid and simple, and its application in bacterial gene knockout is more extensive and mature. It is important for the preparation of the competent cell during state electrical conversion of Red homologous recombination. At present, the commonly used methods for the preparation of competent cells include traditional $CaCl_2$ method and electrical conversion method. Compared with $CaCl_2$ method, the electrical conversion method has the advantages of simple operation and high conversion efficiency [9]. Here, we focus on the screening and identification methods of transcription factor that our laboratories have routinely used for studying transcription factor regulating of key gene expression. DNA pull-down technology and mass spectrometry were used to screen and identify proteins binding to key gene promoters. EMSA technology was used to verify the specific binding of DNA to proteins in vitro, and then Red homologous recombination method was used to knockout the identified transcription factor.

2 Materials

2.1 Materials for DNA Pull-down

1. THES buffer: 50 mM Tris–HCl (pH 7.5). 10 mM EDTA. 20% sucrose (mass/vol). 140 mM NaCl. 0.7% protease inhibitor cocktail II (vol/vol). 0.1% phosphatase inhibitor cocktail II (vol/vol).
2. 5× BS buffer: 50 mM HEPES. 25 mM $CaCl_2$. 250 mM KCl. 60% glycerol.
3. BS/THES binding washing buffer: 44.3% THES buffer. 20.0% 5× BS buffer. 35.7% nuclease-free water.
4. 2× B/W buffer: 10-mM Tris–HCl (pH 7.5). 1 mM EDTA. 2 M NaCl.

5. Other kits and reagents: High-fidelity prime-star DNA polymerase PCR kit. Highly purified PCR product kit. M-280 immunomagnetic beads. ECL kit. BCA protein assay kit. DNA marker. Protease K. Lysozyme. Yeast extract. Tryptone.

2.2 Reagent Preparation for Western Blot

1. 30% acrylamide: Dissolve 1.0 g methylene acrylamide and 29.0 g acrylamide to 100 mL, and filter by 0.45 filter membrane. Store at 4 °C.
2. 1× TAE buffer: Dissolve 121.0 g Tris base, 18.6 g Na_2 EDTA•$2H_2O$, and 28.55 mL acetic acid to 500 mL using ultra-pure water. Autoclave at 121 °C for 30 min, and store at 4 °C. Before use, dilute 50 times with sterile ultra-pure water to obtain 1× TAE buffer.
3. 1× electrophoresis buffer: Dissolve 30.3 g Tris base, 144.2 g glycine, and 10.0 g SDS to 1000 mL using ultra-pure water. Autoclave at 121 °C for 30 min and store at 4 °C. Before use, dilute 10 times with sterile ultra-pure water to obtain 1× electrophoresis buffer.
4. 5× sample loading buffer: Dissolve 3.0 g SDS, 150.0 mg bromophenol blue, and 7.5 mL 1 M Tris–HCl buffer (pH 6.8), 15.0 mL glycerol, and 1.5 mL β-mercaptoethanol to 30.0 mL using ultra-pure water. Store at 120 °C for later use.
5. 1× membrane transfer buffer: Dissolve 144.1 g glycine and 30.4 g Tris base to 1000 mL to obtain 10× membrane transfer buffer. Before use, mix 100.0 mL of 10× membrane transfer buffer and 200.0 mL methanol, and then dilute the mixture volume to 1000 mL with ultra-pure water.
6. 1× TBST buffer: Dissolve 60.55 g Tris base and 87.5 g sodium chloride to 1000 mL to obtain 10× TBS buffer. Before use, mix 100 mL 10× TBS buffer, and add 1.0 mL 0.1% tween using ultra-pure water.

2.3 Materials for ESMA

1. Materials for user: pET-28a vector. Competent *E. coli* BL21. Ni^{2+}-NTA His protein purification column.
2. LB medium: Dissolve 5 g yeast extract, 10 g peptone, 10 g NaCl, 20 g agar to 1 L, pH 7.0. Autoclave at 121 °C for 30 min and store at 4 °C.
3. 5× TBE buffer: 0.45 M Tris, 0.45 M boric acid, and 10 mM Na-EDTA. pH 8.3. Before use, dilute 10 times with sterile ultra-pure water to obtain 0.5× TAE buffer.
4. Kits and other reagents: High-fidelity prime-star DNA polymerase PCR kit. Highly purified PCR product kit. Plasmid small amount extraction kit. Genome extraction kit. Gel cutting recovery kit. ESMA kit. DNA marker. 5× nucleic acid loading buffer. rTaq DNA polymerase enzyme. T4 ligase.

Restriction enzyme including Sma I, BgI II, BamH I, Xho O, Nhe I, Not I, spe I, and EcoR I. IPTG. Erythrocin. Kanamycin. Levogyre.

2.4 Materials for Gene Knockout

1. Materials for user: the plasmids including pKD46, pKD3, and pCP20.
2. LB solid medium: Dissolve 5 g yeast extract, 10-g peptone, 10 g NaCl, and 20-g agar to 1 L, pH 7.0. Autoclave at 121 °C for 30 min.
3. LB liquid medium: Dissolve 5 g yeast extract, 10 g peptone, and 10 g NaCl to 1 L, pH 7.0. Autoclave at 121 °C for 30 min and store at 4 °C.
4. NB medium: Dissolve 10 g peptone, 3 g beef extract, and 5 g NaCl to 1 L, pH 7.2. Autoclave at 121 °C for 30 min and store at 4 °C.
5. Kits and the main reagents: Plasmid small amount extraction kit. Genome extraction kit. Gel cutting recovery kit.

2.5 Instrument and Equipment

1. Magnetic grate.
2. PCR.
3. Protein electrophoresis, nucleic acid electrophoresis, and high-speed refrigerated centrifuge.
4. Nucleic acid electrophoresis.
5. NanoDrop1000.
6. Gel imager system.
7. ECL Western blot detection system.
8. Protein electrophoresis.
9. Ultrasonic crushing instrument.
10. High-speed refrigerated centrifuge.
11. Mass spectrometer.
12. Biological safety cabinet.
13. Electroporator.

3 Methods

3.1 DNA Pull-down

3.1.1 Probe Design and Labeling

1. Based on the sequences of target gene (described as "M") and promoter, design 200–300 bp of DNA probe in size (*see* **Note 1**). Ensure that transcription factor can bind to the middle of the probe. Label biotin at the 3′ terminal of the primers upstream of the probe. In our lab, in order to screen the transcription factor regulating glycerol dehydratase (GDH) expression in *Klebsiella pneumoniae* (*K. pneumoniae*), the primers of DNA probe are as follows:

P_{GDG}-F: TTGATTTATATCATTGCGGGCGATCACATTTT TTATTTTTGCCGCCGGAGTAAAGT-Biotin.

P_{GDG}-F: ACTTTACTCCGGCGGCAAAAATAAAAAATGT GATCGCCCGCAATGATATAAATCAA.

2. Boil 1 mL of *K. pneumoniae* strain liquid for 5 min, and centrifuge at 7500*g* force for 5 min. Use the supernatant as the template of PCR for promoter. The PCR conditions are as follows: 98 °C, denaturation for 10 s; 58 °C, annealing for 20 s; 72 °C, extension for 1 min; 40 cycles.
3. Purify the labeled probe by highly purified PCR product kit, and store it at −20 °C for later use.

3.1.2 Pull-down

1. Pre-mix 5 μg of biotin-labeled DNA probe and 500 μg of the extracted nuclear protein of *K. pneumoniae* on ice.
2. Wash 100 μL of streptavidin-agarose G magnetic beads with 500 μL cold 2× B/W buffer, and centrifuge at 5000*g* for 30 s at 4 °C.
3. Add the magnetic beads into the pre-mixed of DNA and nuclear protein on ice. Remove the supernatant with a pipette gun.
4. Wash the precipitate using 500 μL cold BS/THES buffer. Repeat 2–3 times.
5. Add with 30 μL protein loading buffer to re-suspend the precipitation, and then boil for 10 min.

3.1.3 Western Blot

1. Electrophoresis: Prepare the concentrated gel (5%) and the separated gel (8–12%). Load with 40–60 μg total protein in each well, and set the initial voltage at 100 V. After reaching the separated gel, adjust the voltage to 120 V.
2. Membrane transfer: According to the molecular weight of the target protein, set the membrane transfer time for 60–120 min at 200 mA.
3. Sealing: Place the protein membrane in 1× TBST solution, and rinse it for 1–2 min to wash off the membrane transfer fluid on the membrane. Then, seal it in 5% skim milk powder solution at room temperature for 60 min.
4. Incubation with primary antibody: Cut the PVDF membrane according to the protein marker, and put into the corresponding primary antibody following incubation at 4 °C overnight by shaking. Then, wash it with TBST solution for three times by shaking every 15 min.

5. Incubation with secondary antibody: Dilute the secondary antibody labeled with horseradish peroxidase (HRP) d according to the certain concentration, and incubate with the PVDF membrane for 1 h at room temperature by shaking. Then, wash with TBST solution for three times by shaking every 15 min.
6. Color: The final immunoreactive proteins were analyzed by a gel imager system and an ECL Western blot detection system.

3.1.4 MS Identification

1. Electrophoresis: Prepare the concentrated gel (4%) and the separated gel (8–12%). Load with 40–60 μg total protein in each well, and set the initial voltage at 100 V. After reaching the separated gel, adjust the voltage to 120 V. Then, stain the proteins with Coomassie brilliant blue (*see* Fig. 1). The formulas of the concentrated gel and the separated gel are shown in Table 1.
2. Gel cutting: Cut the colloidal particles with a diameter of 1–2 mm using a blade, and then place it in 1.5 mL EP tube.
3. Cleaning: Clean the colloidal particles using 200 μL for two times every 10 min.

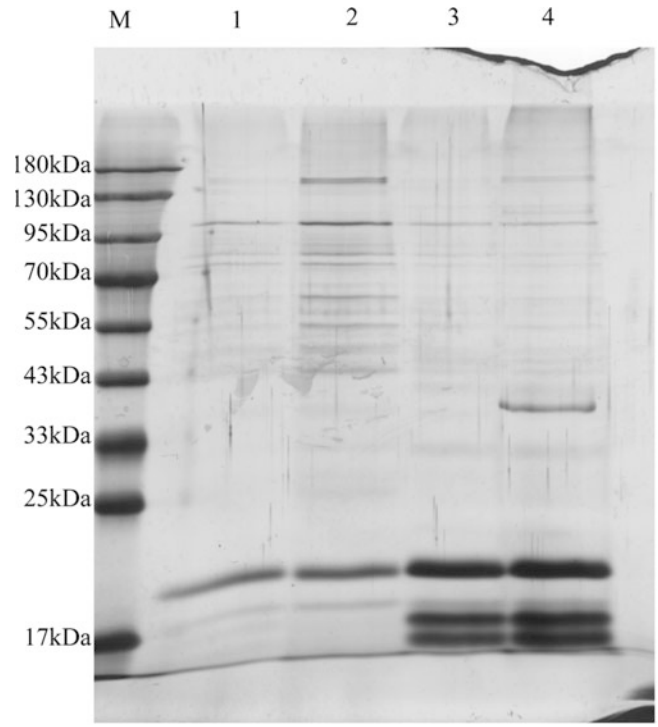


Fig. 1 Electrophoretogram of DNA pull-down. Lane 1: Unlabeled probe + *K. pneumoniae*. Lane 2: Biotin-labeled probe + *K. pneumoniae*. Lane 3: Residual binding protein after elution in lane 1. Lane 4: Residual binding protein after elution in lane 2

Table 1
The formulas of the concentrated gel and the separated gel

Reagent	4% concentrated gel	12% separated gel
ddH_2O (mL)	3.04	3.37
30% acrylamide (mL)	0.65	4.00
Tris–HCl pH 8.8 (mL)	–	2.50
Tris–HCl pH 6.8 (mL)	1.25	–
10% SDS (μL)	50	100
20% AP (μL)	12.5	25
TEMD (μL)	5	5
Total (mL)	5	10

4. Decolorization: Destain the colloidal particles with 200 μL of 50% acetonitrile containing 25 mM NH_4HCO_3 at 37 °C for 20 min, and repeat decolorization 2–3 times.
5. Dehydration: Add 100 μL acetonitrile for dehydration until the colloidal particles become whiter, and remove the acetonitrile.
6. Cleaning: Clean the colloidal particles with 200 μL MilliQ for two times every 10 min followed by cleaning with 200 μL 50% acetonitrile for two times every 10 min.
7. Dehydration: Add 100 μL acetonitrile for dehydration until the colloidal particles become whiter, and remove the acetonitrile.
8. Enzymolysis: Dilute trypsin with 25 mM NH_4HCO_3 to 12.5 mg/mL, and mix the colloidal particles with 10 μL trypsin solution. Place the mixture at 4 °C for 30 min. After removing the excess enzyme solution, add 20 μL 25 mM NH_4HCO_3 into the mixture, and place at 37 °C overnight.
9. Mass identification: Detect the target protein in the separated colloidal particle online by mass spectrometer (*see* **Note 2**). Set the parameters as reflection mode. The conditions used in first mass spectrometry are shown in Table 2, and those used in first second spectrometry are shown in Table 3.
10. Database search: Process the original mass files, and convert them by MM File Conversion software to obtain MGF format files. Then, use ProteinPilot™ 4.5 (Version 1656, AB Sciex) to search uniprot database. The beads+ protein group without DNA probe treatment are the control group (*see* **Note 3**). The retrieval parameters are as shown in Table 4.

Table 2
The used conditions of first mass spectrometry

Program	Parameter
Resolution	70,000
AGC target	3e6
Maximum IT	40 ms
Scan range	350–1800 m/z

Table 3
The used conditions of secondary mass spectrometry

Program	Parameter
Resolution	17,500
AGC target	1e5
Maximum IT	60 ms
TopN	20
NCE/stepped NCE	27

Table 4
The retrieval parameters of the original database search of mass spectrometry

Program	Parameter
Detected protein threshold [unused ProtScore (Conf)] >	0.05 (10.0%)
Competitor error margin (ProtScore)	2.00
Paragon™ algorithm	4.5.0.0, 1654
Cys. alkylation	MMTS
Digestion	Trypsin
Instrument	Orbi MS (1–3 ppm), Orbi MS/MS
ID focus	Biological modifications
Search effort	Thorough
FDR analysis	Yes
User modified parameter files	No

3.2 EMSA

3.2.1 Plasmid Construction of p-ET-28a (+)-M

1. Based on the sequence of the gene M in GenBank, design the primers using Primer Premier 5.0 software.
2. Use the genome DNA in *K. pneumoniae* as the template of PCR for promoter. The PCR conditions are as follows: 98 °C, denaturation for 10 s; set the annealing temperature according to the primer for 20 s; 72 °C, extension for 1 min; 40 cycles. Purify the labeled probe by highly purified PCR product kit, and store it at −20 °C for later use.
3. Digest the purified PCR fragment and pET-28A empty vector with the corresponding restriction endonuclease. Collect the target fragment and the vector fragment, and then ligate by T4 ligase overnight at 16 °C.
4. Transform the ligase product into the competent *E. coli* BL21 (DE3). Select positive transformed colonies to inoculate on LB medium containing 50 μg/mL kanamycin for culture at 37 °C.
5. Using the transformed bacterial solution of *K. pneumoniae* as template, amplify the target gene.

3.2.2 Induction Expression and Purification of Recombinant pET-28a (+)-M

1. Culture *E. coli* BL21 containing the expression plasmid pET-28A(+) M in 10 mL LB medium (containing 50 μg/mL kanamycin) at 37 °C on a 170 g force shaker until the absorbance value at 620 nm is 0.5.
2. Induce with IPTG at final concentration of 0.5 mM at 20 °C on the 110 g force shaker for 8 h overnight.
3. Collect bacterial precipitate followed by washing for three times with elution solution. Treat the precipitate using ultrasonic crushing instrument, and centrifuge at 11,200*g* force for 10 min.
4. Filter it by using a 0.22 μm filter membrane to obtain the supernatant. Purify the target recombinant protein on a Ni^{2+}-NTA His protein purification column, and elute with 500 mM imidazole eluent.

3.2.3 Gel Preparation for EMSA

1. Prepare 15 mL 6% non-denatured polyacrylamide gel (*see* **Note 4**), according to the formula in Table 5.

3.2.4 EMSA Conjugation Reaction

1. The conditions for negative control reaction are set as following in Table 6.
2. The conditions for sample reaction are set according to Table 7.
3. The conditions for the competitive reaction probe are shown in Table 8 (*see* **Note 6**).
4. After adding with these above-mentioned reagents (except labeled probe), place the mixture at room temperature (20–25 °C) for 10 min to eliminate possible non-specific

Table 5
The formulas of 6% non-denatured polyacrylamide gel (15 mL) in EMSA conjugation reaction

Reagent	Volume
ddH_2O	7.15 mL
30% acrylamide (W/V)	2 mL
5 × TBE	1 mL
Glycerol	175 μL
10%APS	75 μL
TEMED	8 μL
Total volume	15 mL

Table 6
The conditions for negative control reaction in EMSA conjugation reaction

Reagent	Volume
Nuclease-free water	7 μL
EMSA/gel-shift binding buffer (5×)	2 μL
Bacterial nuclear protein or purified recombinant protein	0 μL
Labeled probe	1 μL
Total volume	10 μL

Table 7
The conditions for sample reaction EMSA conjugation reaction

Reagent	Volume
Nuclease-free water	5 μL
EMSA/gel-shift binding buffer (5×)	2 μL
Bacterial nuclear protein or purified recombinant protein	2 μL
Labeled probe	1 μL
Total volume	10 μL

Table 8
The conditions for the competitive reaction probe in EMSA conjugation reaction

Reagent	Volume
Nuclease-free water	4 μL
EMSA/gel-shift binding buffer (5×)	2 μL
Purified recombinant protein	2 μL
Labeled probe	1 μL
Unlabeled probe	1 μL
Total volume	10 μL

binding between probes and proteins. Then, add the labeled probe, and place at room temperature (20–25 °C) for 20 min. Finally, add 1 μL EMSA/gel-shift sampling buffer (colorless, 10×) into the mixture.

3.2.5 Electrophoretic Analysis

1. Electrophoresis: Use the diluted 0.5× TBE buffer as the electrophoresis solution for re-electrophoresis at 10 V per 1 cm for 10 min. Load 10 μL of the above-mentioned sample treated with colorless 1 × EMSA/gel-shift loading buffer (*see* **Note 7**). Besides, load 10 μL of 1 × EMSA/gel-shift loading buffer (blue) to observe the electrophoresis terminus (*see* **Note 8**).
2. Transfer membrane: Soak the nylon membrane with appropriate size to the EMSA gel in 0.5 × TBE buffer for 15 min. Place sponge, filter paper, gel, nylon membrane, filter paper, and sponge in turn according to the sandwich method. Remove all bubbles in the sandwich carefully. Transfer the membrane at 300 mA for 30 min.
3. Ultraviolet crosslinking: Take the transferred membrane, and expose it under a hand-held UV lamp equipped with a 254 nm bulb for 5–10 min.
4. Chromogenic reaction: Put the crosslinked membrane into the 15 mL blocking buffer in EMSA kit for 15 min. Then, add 50 μL stabilized streptavidin-horseradish peroxidase conjugate reagent in EMSA kit, and let react for 15 min. Wash the membrane 4 times using wash buffer in EMSA kit every 5 min. Finally, shake the membrane in 15 μL substrate equilibration buffer for 10 min, and detect by ECL Western blot detection system (*see* Fig. 2).

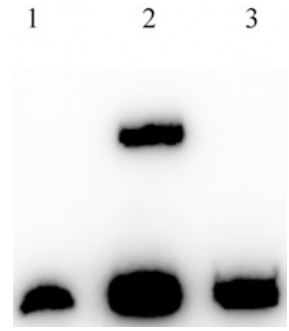


Fig. 2 Specific binding of recombinant proteins screened by EMSA and DNA probes verified in vitro. Lane 1: Biotin-labeled probe. Lane 2: Biotin-labeled probe + recombinant protein. Lane 3: Biotin-labeled probe + recombinant protein +200 times cold competitive probe

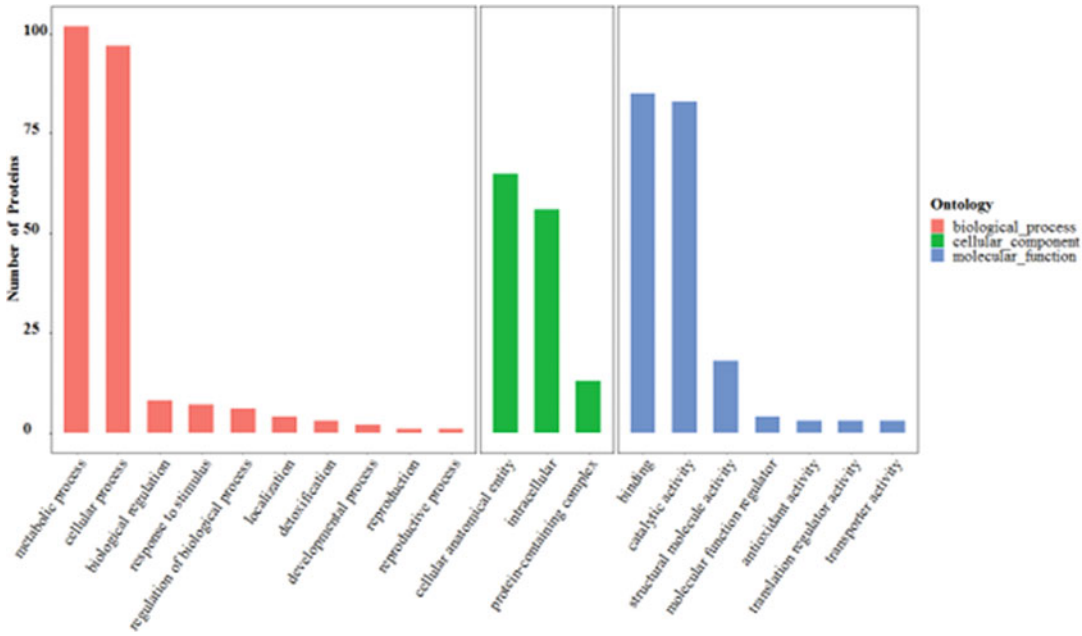


Fig. 3 GO analysis of the screened transcription factor

3.3 Bioinformatics Analysis

The tools and websites for bioinformatics analysis including KEGG analysis, GO analysis (*see* Fig. 3), and string analysis of the selected transcription factor used are as follows:

1. For protein subcellular localization: https://www.genscript.com/psort.html
2. For protein, the physical and chemical property analysis: https://web.expasy.org/protparam/
3. For conservative protein sequence analysis: https://consurf.tau.ac.il/

4. For protein secondary structure prediction: https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html
5. For protein tertiary structure prediction: https://swissmodel.expasy.org/
6. For protein sequence analysis: https://predictprotein.org/
7. For protein structure domain analysis: https://www.ncbi.nlm.nih.gov/

3.4 Gene Knockout

3.4.1 Primer Design

1. Design the primers MU and MD based on 600–800 bp upstream and downstream of the gene *GDH* in *K. pneumoniae* (*see* **Note 9**).
2. Design the primers K1 and K2 according to the internal sequence of kan^r for subsequent transformation validation.

3.4.2 Preparation of Competent Target Strain

1. Plate streaking: Culture *K. pneumoniae* strain on a non-resistant LB solid medium at 37 °C for 15–18 h (*see* **Note 10**).
2. Inoculation: Inoculate single colony of the target strain into 20 mL LB liquid medium, and culture at 30 °C on a shake at 190 g force overnight.
3. Expanded culture: Add the above bacteria liquid of *K. pneumoniae* into 200 mL LB medium at a dosage of 1% (v/v) at 30 °C on a shake at 190 g force. When the OD_{600} value reaches 0.4–0.6, place the bacteria liquid of *K. pneumoniae* on ice for pre-cooling for 20 min.
4. Collecting bacteria: Centrifuge the pre-cooled bacteria liquid of *K. pneumoniae* at 4700 g force at 4 °C for 12 min to collect *K. pneumoniae* strain, and discard the supernatant.
5. Washing bacteria: Wash *K. pneumoniae* strain with pre-cooling ddH_2O for one time and pre-cooling 10% glycerol for two times (*see* **Note 11**). After centrifugation, resuspend it using pre-cooling 10% glycerol, put into liquid nitrogen immediately, and store at −80 °C.

3.4.3 Electrical Transformation

1. Thaw 100 μL competent cells on ice for a few minutes, and add 5 μL pKD46 plasmid.
2. After co-incubation on ice for 1.5 h, place the cells in a pre-cooled 0.2 cm electric transfer cup. Set the electric transfer parameter as follows: 2500 V (*see* **Note 12**), 200 Ω, and 25 mF for 5 ms.
3. After the electrical transfer, add immediately 900 μL NB medium into the incubator at 30 °C for 140 g force, and resuscitate for 3 h.

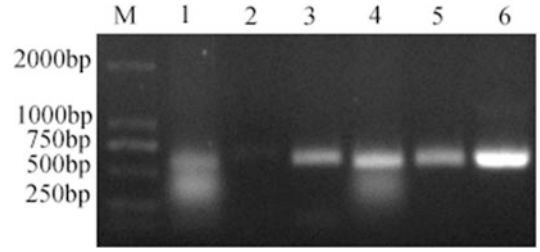


Fig. 4 PCR validation of kanamycin resistant fragment knockout in *K. pneumoniae*. Lane 1: Wild-type control. Lane 2: Negative control. Lanes 3–6: The transformants

4. Centrifuge the transferred bacteria at 5600*g* force for 5 min, and resuspend the precipitate. Then, coat it on LB solid medium containing kanamycin at 30 °C, and screen the transformant.

3.4.4 The Removal of Plasmid and Resistance

1. Inoculate the transformed strain on LB liquid medium containing kanamycin at 42 °C for 24 h. Sequentially, select single colony to inoculate on the LB solid mediums containing kanamycin or ampicillin, respectively, at 37 °C overnight.
2. Screen *K. pneumoniae* strain growing on LB solid medium containing kanamycin but not on ampicillin, and further identify by PCR, which is the clone of the removed pkD46 plasmid.
3. Transform the pCP20 plasmid into the resistant gene deletion strain by electric transformation method without adding arabinose. Screen the positive clone by LB solid medium containing ampicillin. Screen the negative clone on the LB solid medium. After culturing at 42 °C, remove the pCP20 plasmid, and further identify by PCR (*see* Fig. 4).

4 Notes

1. DNA double strand is usually used as probe.
2. The number of proteins identified by MS depends on the strength of protein–DNA binding and the abundance of proteins.
3. It is possible to identify proteins in the control group (beads+ protein). Although there are no probes in the control group, a small number of proteins can combine with magnetic beads. Hence, the proteins identified in the control group are non-specific background proteins.
4. The purified protein and cell crude extract are usually incubated together with biotin-labeled DNA or RNA probes, and the complex and unbound probes are separated by non-denatured polypropylene gel electrophoresis.

5. When detecting DNA-binding proteins such as transcriptional regulators, the purified proteins or the nuclear cell extracts can be used.
6. The specificity of DNA binding proteins is determined by using DNA fragments containing protein-binding sequences and oligonucleotide fragments (specific) in competitive experiments. In the presence of competing specific and non-specific fragments, specific binding is determined according to the characteristics and strength of the complex.
7. Bromophenol blue can affect the binding of protein and DNA. It is recommended to use colorless EMSA/gel-shift loading buffer.
8. DNA complexes move more slowly than unbound probes or bromophenol blue in EMSA/gel-shift loading buffer (blue).
9. If the fusion PCR is not successful, 40 bp homologous arm to the primer may be helpful for amplification.
10. There are differences in preparation methods of the competent cell among different strains.
11. The preparation of the competent cell is important in electrical conversion. If the bacterial strain with thick cell wall, it can be pre-suspended with ddH_2O for 20–30 min before adding glycerin resuspension, so that the bacteria cell wall become thinner and easy to transfer.
12. The voltage of electric shock conversion can be adjusted according to the thickness of cell wall of the target strain.

Acknowledgments

This study was supported by the Outstanding Youth Project of Hunan Education Authority (19B505), the National Natural Science Foundation of China (32071673, 31772374, and 31900087), and the Program of Hunan Science and Technology Innovation Team (2021RC4063).

References

1. Qiong W, Amrutkar SM, Shao F (2018) Sulfinate based selective labeling of 5-hydroxymethylcytosine: application to biotin pull down assay. Bioconjug Chem 29:245–249
2. Kliszczak AE, Rainey MD, Harhen B, Boisvert FM, Santocanale C (2011) DNA mediated chromatin pull-down for the study of chromatin replication. Sci Rep 1:95
3. Ozawa M, Ozawa T, Nishio M, Ueda K (2017) The role of CH/π interactions in the high affinity binding of streptavidin and biotin. J Mol Graph Model 75:117–124
4. Arakaki A, Hideshima S, Nakagawa T, Niwa D, Tanaka T, Matsunaga T, Osaka T (2004) Detection of biomolecular interaction between biotin and streptavidin on a self-assembled monolayer using magnetic nanoparticles. Biotechnol Bioeng 88:543–546
5. Ruscher K, Reuter M, Kupper D, Trendelenburg G, Dirnagl U, Meisel A (2000)

A fluorescence based non-radioactive electrophoretic mobility shift assay. J Biotechnol 78: 163–170

6. Fajardo T, Sung PY, Celma CC, Roy P (2017) Rotavirus genomic RNA complex forms via specific RNA-RNA interactions: disruption of RNA complex inhibits virus infectivity. Viruses 9:167
7. Sancar A, Rupp WD (1979) Physical map of the recA gene. Proc Natl Acad Sci U S A 76:3144–3148
8. Chen W, Chen R, Cao J (2021) Rapid genome modification in *Serratia marcescens* through Red homologous recombination. Appl Biochem Biotechnol 193:2916–2931
9. Chai D, Wang G, Fang L, Li H, Liu S, Zhu H, Zheng J (2020) The optimization system for preparation of TG1 competent cells and electrotransformation. Microbiology 9:e1043

Chapter 3

Genome-Wide Identification of Open Chromatin in Plants Using MH-Seq

Aicen Zhang, Xinxu Li, Hainan Zhao, Jiming Jiang, and Wenli Zhang

Abstract

Functional *cis*-regulatory elements (CREs) act as precise transcriptional switches for fine-tuning gene transcription. Identification of CREs is critical for understanding regulatory mechanisms of gene expression associated with various biological processes in eukaryotes. It is well known that CREs reside in open chromatin that exhibits hypersensitivity to enzyme cleavage and physical shearing. Currently, high-throughput methodologies, such as DNase-seq, ATAC-seq, and FAIRE-seq, have been widely applied in mapping open chromatin in various eukaryotic genomes. More recently, differential MNase (micrococcal nuclease) treatment has been successfully employed to map open chromatin in addition to profiling nucleosome landscape in both mammalian and plant species. We have developed a MNase hypersensitivity sequencing (MH-seq) technique in plants. The MH-seq procedure includes plant nuclei fixation and purification, differential treatments of purified nuclei with MNase, specific recovery of MNase-trimmed small DNA fragments within 20~100 bp in length, and MH-seq library construction followed by Illumina sequencing and data analysis. MH-seq has been successfully applied for global identification of open chromatin in both *Arabidopsis thaliana* and maize. It has been proven to be an attractive alternative for profiling open chromatin. Thus, MH-seq is expected to be valuable in probing chromatin accessibility on a genome-wide scale for other plants with sequenced genomes. Moreover, MHS data allow to implement footprinting assays to unveil binding sites of transcription factors.

Key words Chromatin accessibility, Open chromatin, MNase-hypersensitive sites, MH-seq, Plants

1 Introduction

Precise spatiotemporal physical interactions between *cis*-regulatory elements (CREs) and transcription factors (TFs) form the hub of complex transcriptional regulatory networks, which play key roles in fine-tuning gene transcription during normal growth and development and in response to internal and external stimuli in plants [1]. Active CREs reside in open chromatin accessible for recruiting *trans*-factors. Open chromatin is loosely packed and is usually nucleosome-depleted or in low nucleosome occupancy in various eukaryotic genomes [2–4]. Therefore, identification of open

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_3,

chromatin is a key step toward deciphering regulatory genomic loci, which lead to deep understanding of transcriptional regulatory mechanisms underlying various biological processes in eukaryotes. The currently available methods for genome-wide mapping of chromatin accessibility can be classified into indirect and direct assays. Enzyme- or chemical-based nucleosome mapping, like MNase-seq [5, 6], MPE-seq [7], and copper ion-mediated Fenton reaction coupled with sequencing [8], can be used to indirectly profile accessible chromatin across the genome. Direct assays include epitope-dependent ChIP-seq assay [9], including TFs, Pol II, silencer, or insulator ChIP-seq [10–13], and a number of antibody-free methods, including DNase-seq [14–16], FAIRE-seq [17], MH-seq [18], ATAC-seq [19, 20], and NOMe-seq [21], have been successfully applied to comprehensively measure chromatin accessibility across various eukaryotic genomes, ranging from yeast and humans to plants.

MNase was first isolated from bacterium *Staphylococcus aureus*, exhibiting endo- and exonuclease activity to naked DNA [22]. At chromatin level, MNase preferentially cleaves linker DNA between neighboring nucleosomes relative to DNA sequences tightly wrapped around nucleosomes. Recovery of MNase-trimmed mononucleosomal DNA fragments (approximately 150 bp) combined with high-throughput sequencing, referred as to MNase-seq, is commonly used for genome-wide mapping of nucleosomes landscape [23–25]. MNase accessibility (MACC) or differential MNase treatment can simultaneously measure nucleosome and accessible chromatin regions in bulk cells in *Drosophila* [5] or a single human cell [26]. Similar to DNase hypersensitive sites (DHSs), chromatin-accessible regions are more sensitive to MNase treatment as compared to regions associated with nucleosomes. A methodology with light MNase cleavage, termed as MNase hypersensitivity sequencing (MH-seq), was successfully developed to specifically identify open chromatin, MNase hypersensitive sites (MHSs), across maize [18] and *Arabidopsis* [27] genomes. In this chapter, we describe a detailed and robust MH-seq methodology for global mapping of open chromatin in plants. The MH-seq procedure consists of nuclei fixation and purification, differential cleavage of purified nuclei with MNase, specific recovery of MNase-trimmed DNA fragments ranging from 20 to 100 bp, MH-seq library preparation coupled with Illumina sequencing, and data analysis for identification of MHSs across the genome. This protocol has been successfully employed in both *Arabidopsis* [27] and maize [18] for mapping functional genomic loci or CREs and can also be adapted for other plant species with a sequenced genome.

2 Materials

2.1 Plants

Plant species with a sequenced genome are suited for MH-seq-based identification of open chromatin. Plants grown to a certain developmental stage under artificially controlled environment or in the field can be used for nuclei preparation for MH-seq experiments (*see* **Note 1**).

2.2 Isolation and Purification of Nuclei

1. Beaker.
2. Scissor.
3. Vacuum pump.
4. Mortar and pestles.
5. 50-mL Corning conical tubes.
6. Centrifuge with swing bucket rotor with cooling system.
7. Miracloth.
8. Funnel.
9. Pre-chilled spatula.
10. Fixation buffer: 20 mM HEPES, 1 mM EDTA, 100 mM NaCl, 1 mM PMSF, and 1% formaldehyde.
11. 2 M glycine.
12. Nuclei isolation buffer (NIB): 10 mM Tris–HCl (pH 8.0), 80-mM KCl, 10 mM EDTA, 1 mM spermidine, 1 mM spermine, 0.5 M sucrose, pH 9.5. Store at 4 °C. Add fresh β-mercaptoethanol to achieve a final concentration of 0.15% before use.
13. Nuclei washing buffer (NWB): NIB containing 0.5% of freshly prepared Triton X 100.
14. MNase digestion buffer (MNB): Mix 1 mL of 50% sucrose, 250 μL of 1 M Tris–HCl (pH 7.5), 20 μL of 1 M $MgCl_2$, 5 μL of 1 M $CaCl_2$; add ddH_2O to make up to 5 mL.

2.3 MNase Digestion and DNA Extraction

1. Micrococcal nuclease (MNase) (NEB, cat # M0247S), RNase A, and proteinase K.
2. Pipettes with various scales.
3. 1.5 mL cap-locked Eppendorf tubes.
4. 37 °C, 65 °C, and 55 °C water bath.
5. Stop solution: 0.5 M EDTA.
6. 5 M NaCl.
7. 20% SDS.
8. 1× TE buffer.
9. Tris-saturated pure phenol solution (pH 8.0) and chloroform.
10. Glycogen, 5 mg/mL stock solution.

11. Sodium acetate (NaOAc, 3 M, pH 5.2).
12. Ethanol (100% and 70%).
13. EB buffer: 10 mM Tris–HCl, pH 8.0.

2.4 Recovery of Small-Sized DNA Fragments

1. Certified low range ultra agarose.
2. 1× TBE buffer: 90 mM Tris-borate, 2 mM EDTA, pH 8.3.
3. GeneRed (TIANGEN, cat # RT211).
4. 100 bp ladder.
5. Electrophoresis apparatus and Gel Doc XR+.
6. Gel knife.
7. QIAquick Gel extraction kit (QIAGEN, cat # 28704).

2.5 MH-seq Library Preparation

1. Thermocycler.
2. 0.2-mL PCR strip tubes.
3. Centrifuge with 24 × 2 mL rotor at RT.
4. Pipettes with various scales.
5. Magnetic stand (Alpaqua, cat # A001219).
6. 0.5 mL/1.5 mL/2 mL cap-locked Eppendorf tubes.
7. Gel knife.
8. Electrophoresis apparatus and Gel Doc XR+.
9. Thermo Labquake™ rotator.
10. Spin-X filter (Sigma, 0.45 μm, cat # CLS8162).
11. 21 gauge needle.
12. NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, cat # E7645S) including the following components:
    NEBNext Ultra II End Prep Enzyme Mix;
    NEBNext Ultra II End Prep Reaction Buffer;
    NEBNext Ultra II Ligation Master Mix;
    NEBNext Ligation Enhancer;
    NEBNext Ultra II Q5 Master Mix.
13. AMPure XP Beads (Bechman Coulter, A63881).
14. 80% freshly prepared ethanol.
15. 1× TBE: 90 mM Tris-borate, 2 mM EDTA, pH 8.3.
16. EB buffer: 10 mM Tris–HCl, pH 8.0.
17. 50 bp DNA ladder.
18. NEBNext Multiplex Oligos for Illumina (96 Index Primers, NEB, cat # E6609) including the following components:
    NEBNext Adaptor for Illumina;
    USER Enzyme;
    NEBNext Index/Universal Primer Mix Plate.

3 Methods

3.1 Preparation of Purified Nuclei for MNase Digestion

1. Cut the collected fresh tissues (~2 g) into a length of 1–2 cm pieces, and then immerse them in excess fixation buffer under a vacuum for 10 min at room temperature (RT) (*see* **Note 2**).
2. Quench the excessive formaldehyde by adding 2 M glycine to a final concentration of 0.125 M under a vacuum for another 5 min.
3. Discard the liquid, wash the materials three times using sterilized water followed by absorbing residual water with absorbent papers, freeze the materials in liquid nitrogen for 5 min followed by transferring to −80 °C for a long-term storage, or directly use in the next step.
4. Grind the cross-linked tissues into fine powder in liquid nitrogen, and then transfer the ground powder into a 50 mL ice-cold Corning centrifuge tube. Store the powder at −80 °C if not used immediately (*see* **Note 3**).
5. Add an equal volume of ice-cold NIB containing 0.15% fresh β-mercaptoethanol to 5 mL of powder, and completely mix with a chilled spatula.
6. Gently shake the tube on ice for 6 min to make the powder and liquid mixed intensively, and then filter the slurry solution through four-layer Miracloth into a new 50 mL Corning tube.
7. Centrifuge the filtered solution at 1000 × *g* for 10 min at 4 °C, and discard the supernatant as much as possible.
8. Add 10 mL of cold NWB and resuspend the pellets using a soft nylon paintbrush, and then centrifuge at 1000 × *g* for 10 min at 4 °C to pellet the nuclei (*see* **Note 4**).
9. Decant the supernatant, and then repeat **step 8** 2–3 times until the nuclei become white or yellowish.
10. Resuspend the pellet with 5 mL of MNB, and centrifuge at 1000 × *g* for 10 min at 4 °C; purified nuclei can be obtained after removing supernatant.

3.2 MNase Digestion and Purification of Digested DNA Fragments

1. Gently resuspend the purified nuclei in 1.2 mL of cold MNB using a paintbrush.
2. Equally aliquot the suspension into six 1.5 mL prechilled Eppendorf tubes with 200 μl per aliquot; all samples need to be placed on ice.
3. Add various amount of MNase with a specific enzyme unit to each tube; mix well by gently inverting the tubes several times (*see* **Note 5**).
4. Incubate the tubes in a 37 °C water bath for 10 min, and gently mix every 3 min.

5. Stop the reaction by adding 16 μl of 0.5 M EDTA (pH 8.0), mix well by inverting the tubes, and place samples on ice.
6. Equally divide each sample into two parts, and store one cross-linked part at 4 °C overnight; add 16 μL of 5 M NaCl, 8 μL of 20% SDS, and 160 μL of 1 × TE to the second part of each sample, and incubate the mixture at 65 °C overnight for reverse cross-linking.
7. Add 20 μg of RNase A to each tube, and incubate for 30 min in a 37 °C water bath; then add 100 μg of proteinase K, and incubate for 2 h in a 55 °C water bath. Make the total volume in each tube approximately up to 400 μL by adding 1 × TE.
8. Extract DNA using an equal volume of phenol (400 μL), phenol:chloroform mixture (1:1), chloroform, respectively. Vibrate violently, and transfer upper liquid after centrifugation at 15,000 × *g* for 10 min at RT for each time.
9. Pipet the upper supernatant into a new 1.5 mL Eppendorf tube after the last round of extraction; add 20 μg glycogen, 1/10 volume of 3 M NaOAc (pH 5.2), and 2 × volumes of ice-cold 100% ethanol; gently invert for several times; and store at −20 °C for at least 1 h to precipitate DNA.
10. Centrifuge at 15,000 × *g* for 15 min at 4 °C and remove supernatant.
11. Wash the DNA pellet with 500 μL of 70% ethanol, and centrifuge at 15,000 × *g* for another 5 min; decant all residual liquid.
12. Air-dry the DNA for 5 min at RT, and then dissolve the DNA in 20 μL EB.

3.3 Recovery of MNase-Cleaved Small-Sized DNA Fragments

1. Prepare 2% of agarose gel containing GeneRed with 1 × TBE (*see* **Note 6**).
2. Load 10 μL of each reverse cross-linked DNA to each well, use the corresponding cross-linked DNA as a control, and run agarose gel electrophoresis at 100 V for 1 h.
3. Recover DNA fragments <100 bp from MNase-trimmed nuclei under an appropriate MNase unit (*see* Fig. 1).
4. Purify the DNA using a QIAGEN gel extraction kit (QIAGEN, part # 28704), and elute DNA with 30 μL EB.

3.4 MH-seq Library Preparation According to the Manual for the Kit

3.4.1 End Repair and Addition of "A" Base

1. Use 5–20 ng fragmented DNA dissolved in EB (or 1 × TE) as starting material. Add EB to a final volume of 50 μL.
2. Make 60 μL of mix in a PCR tube. Gently pipette up and down for at least 10 times to mix thoroughly (*see* **Note 7**).

   NEBNext Ultra II End Prep Enzyme Mix: 3 μL

   NEBNext Ultra II End Prep Reaction Buffer: 7 μL

   Fragmented DNA: 50 μL

   The total volume: 60 μL

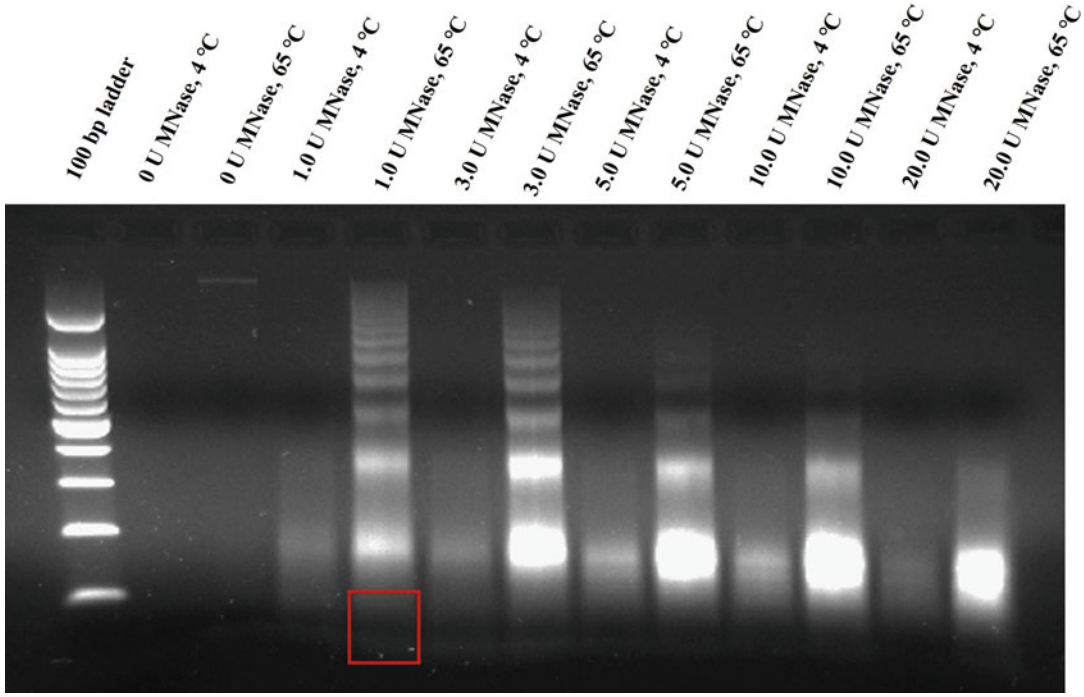


Fig. 1 Agarose gel image showing separation of MNase-trimmed nucleosomal DNA in 2% of agarose gel. Nuclei were purified from formaldehyde-fixed rice seedling tissues and cleaved by various amounts of MNase as indicated on the top. DNA was purified from MNase-trimmed nuclei with (MNase-trimmed fixed nuclei incubated at 65 °C overnight) or without (MNase-trimmed fixed nuclei kept at 4 °C overnight) reverse-cross-linking. The DNA fragments with size less than 100 bp indicated by red box were recovered from fixed nuclei with 1.0 U MNase treatment for MH-seq library preparation. The optimal MNase digestion should be determined by the sequencing data

3. Place the tube containing the mixture in a thermocycler, with the heated lid set to 80 °C. The running program is as follows: 30 min at 20 °C; 30 min at 65 °C; hold at 4 °C.

3.4.2 Adaptor Ligation

1. Add 33.5 μL of the following components to the 60 μL of mixture obtained in the last step. Gently mix well using pipettes (*see* **Note 8**).

   NEBNext Ultra II Ligation Master Mix: 30 μL

   NEBNext Ligation Enhancer: 1 μL

   NEBNext Adaptor for Illumina: 2.5 μL

   The total volume: 93.5 μL

2. Incubate at 20 °C for 15 min in a thermocycler with the heated lid off.
3. Add 3 μL of USER™ Enzyme to the adaptor–ligation mixture; the total volume is 96.5 μL. Gently mix well using pipettes.
4. Incubate at 37 °C for 15 min in a thermocycler with the heated lid set to 50 °C.

3.4.3 Cleanup of Adaptor-Ligated DNA

1. Add 87 μL (0.9×) resuspended AMPure XP beads, which are prewarmed at RT for at least 30 min before use, to 96.5 μL mixture, mix well, and incubate at RT for 10 min.
2. Put the tube on a magnetic stand for 5 min to completely separate the beads from the supernatant. Gently move the tube to converge beads if necessary.
3. Carefully discard the supernatant, wash beads using 200 μL of freshly prepared 80% ethanol for two times, and incubate at RT for 30 s each time. All operations are performed on the magnetic stand.
4. Remove the supernatant as much as possible, air-dry the beads on the magnetic stand until all visible liquid has been evaporated, but do not overdry the beads (*see* **Note 9**).
5. Remove the tube from the magnetic stand. Elute DNA with 15–17 μL EB, pipet up and down to mix well, and incubate at RT for 10 min.
6. Place the tube back on a magnetic stand for 5 min at RT, transfer the supernatant into a new tube after the liquid becomes clear, do not contain the beads in the final solution, and extend the time on the magnetic stand to thoroughly converge beads if necessary.

3.4.4 Enrichment of Adaptor-Ligated DNA by PCR

1. Any adaptor-ligated DNA fragments can be amplified by PCR. PCR cocktail:

   Adaptor-ligated DNA fragments: 15 μL

   NEBNext Ultra II Q5 Master Mix: 25 μL

   Index/universal primer: 10 μL

   The total volume: 50 μL
2. Run PCR program with the following parameters: 98 °C × 30 s; 11 cycles of 98 °C × 10 s, 65 °C × 75 s; 65 °C × 5 min; then hold on at 4 °C (*see* **Note 10**).

3.4.5 Cleanup of PCR Products

1. Prewarm the AMPure XP beads at RT for at least 30 min.
2. Add 45 μL (0.9×) resuspended beads to the PCR products. Mix well and incubate at RT for 10 min, and then place the tube on a magnetic stand for 5 min to converge beads.
3. Carefully decant the supernatant without disturbing the beads.
4. Wash the beads for two times using 200 μL freshly prepared 80% ethanol, incubate at RT for 30 s each time, and remove all the residual liquid.
5. Air-dry the beads on the magnetic stand until all visible liquid has been evaporated, but do not overdry the beads.

6. Remove the tube from the magnetic stand. Elute DNA with 15–17 μL EB, mix well, and incubate at RT for 10 min.
7. Place the tube back on the magnetic stand for 5 min, and transfer the supernatant into a new liquid. The final liquid can be stored at −20 °C for later use (*see* **Note 11**).

3.4.6 Purification of Bead-Purified PCR Products Using Polyacrylamide Gel Electrophoresis (PAGE) (See ***Note 12***)

1. Prepare 15% TBE polyacrylamide gel (PAGE); make sure it is fully solidified before use.
2. Premix the 16 μL of PCR products with 4 μL 5 × loading buffer. Load the mix into one well of PAGE gel; add 1 μL of 50 bp DNA marker to another single well as control.
3. Run the PAGE gel at 100 V for about 3 h in 1 × TBE buffer.
4. Stain the gel in a petri dish (150 × 15 mm) containing 5 mL of sterile water or 1 × TE plus 10 μL of GeneRed (TIANGEN). Observe the gel under the UV light.
5. Cut the DNA band with size ranging from 150 bp to 250 bp with a clean gel knife (*see* Fig. 2); place the gel pieces into a 0.5 mL Eppendorf tube with bottom punched by a 21 gauge needle.

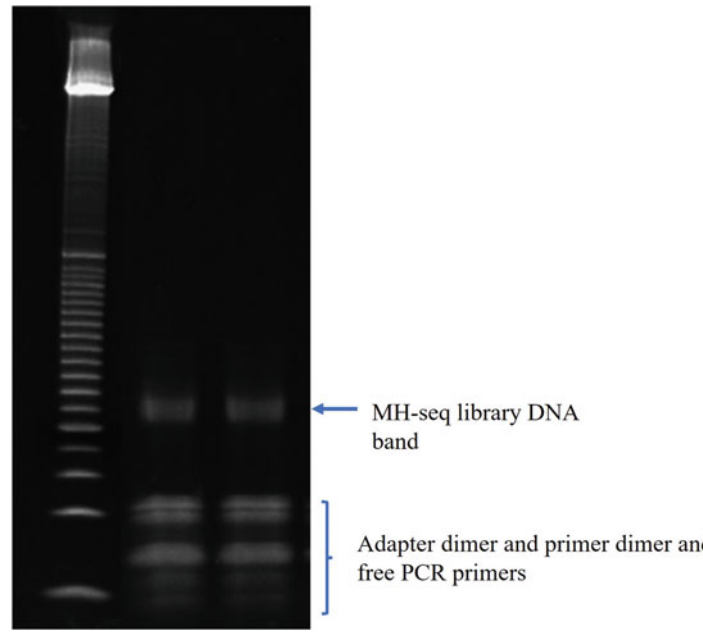


Fig. 2 PAGE image showing the final PCR enrichment of MH-seq library. Approximately 25–30 μl of the PCR product with loading dye were loaded in each lane. The DNA band of interest with sizes ranging from 150 to 220 bp is indicated by the blue arrow. The single bracket indicates the contamination of adapter and PCR primer dimer and free PCR primers that need to be removed after purification

6. Put the 0.5 mL tube into a 2 mL Eppendorf collection tube with round bottom, and centrifuge the tube at 15,000 × *g* for 5 min at RT to crash the gel through the hole.
7. Add 400 μL of 1 × TE buffer into the 2 mL tube, and elute the DNA by gently rotating the tube using Thermo Labquake™ rotator at RT for 2 h.
8. Transfer the crashed gel mixture to the column of the Spin-X filter (Sigma, 0.45 μm) using a cut pipet tip. Spin at 15,000 × *g* for 3–5 min, and transfer the filter liquid into a new 1.5 mL Eppendorf tube.
9. Repeat the same procedures as **steps 8–11** in Subheading 3.2 to precipitate the DNA.
10. Air-dry the DNA for 5 min at RT, and then add 15–17 μL of EB to dissolve DNA for 5–10 min.
11. Rerun 1–2 μL of recovered DNA on 15% TBE PAGE gel to verify the quality of the MH-seq library (*see* Fig. 3).
12. Quality control and estimate the concentration of MH-seq library by using BioAnalyzer before performing Illumina sequencing (*see* Fig. 4).

3.4.7 Bioinformatics Analysis of MH-seq Datasets

The raw sequencing reads are trimmed using Cutadapt [28] and are mapped to the *Arabidopsis* genome (TAIR 10) and maize genome (B73_RefGen_v4) by Bowtie 2 [29], respectively. MACS2 [30] with default parameters is used to call peaks for identification of MHSs across the whole genome. IGV [31] is used for displaying MHS peaks, and the profiles of open chromatin could be plotted by R. More details for MH-seq data analysis can be found in previous publications [27].

3.4.8 MHS-Related Footprinting Assay

FIMO from meme suite [32] is used to search motifs on a genome-wide scale, which are located within binding sites of transcription factors (TFs). TF data are derived from the current database of plant transcription factors and newly generated transcription factor-related ChIP-seq/DAP-seq datasets. All TF-binding motifs identified within DHSs or MHSs are used for downstream footprinting assays. The motif center is used as a reference point; the average DNase or MNase cut frequencies with 5 bp windows are computed around the reference point. The conspicuous dip corresponding to the motif center represents the presence of footprinting (*see* Fig. 5), which is caused by the protection of TF protein binding, therefore resulting in less DNase or MNase cuts occurred relative to immediately flanking regions.

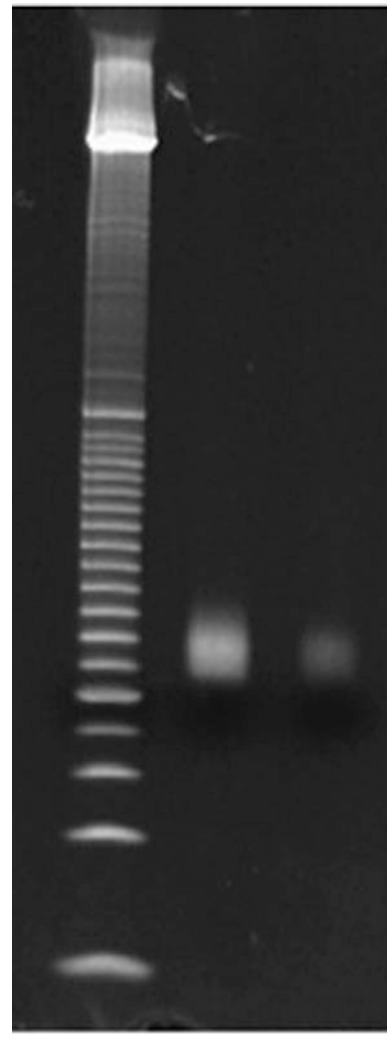


Fig. 3 Confirmation of purified MH-seq library using 1-μl re-running 15% PAGE. Only a single DNA band with the expected sizes, ranging from 150 to 200 bp, is observed in the gel

4 Notes

1. This protocol has been successfully applied in model plant species, *A. thaliana* and maize. It should be applicable to all plant varieties with sequenced genomes.
2. Formaldehyde can create covalent bonds between DNA and proteins to stabilize their interactions in vivo. The final concentration of formaldehyde and incubation time need to be adjusted according to species and tissue types. Excessive cross-linking will impact solubility of protein–DNA complexes; it is better to control the cross-linking time under vacuum within 10 min.

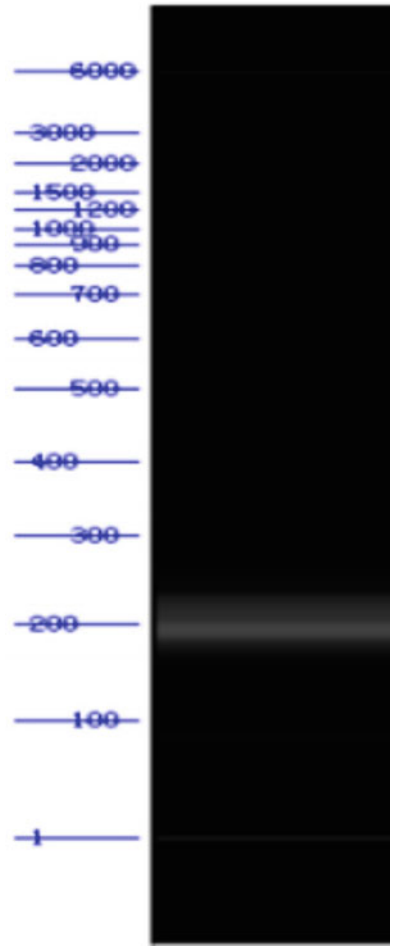


Fig. 4 A BioAnalyzer gel image showing the quality of MH-seq library. An ideal MH-seq library with sizes ranging from 170 to 200 bp, reflecting inserted DNA fragments with sizes ranging from 50 to 100 bp

3. Young tissues from plant seedlings can result in a higher yield of nuclei. Fresh plant tissues should be ground into powder as fine as possible in liquid nitrogen. Keep the powder frozen before adding ice-cold NIB.
4. Triton X-100 (0.5% final concentration) is used to remove chloroplast and mitochondria contamination by breaking the membrane. The concentration of Triton X-100 needs to be adjusted to achieve the best results. The nuclei should become yellowish or white after washing for 3 times; otherwise, the concentration of Triton X-100 needs to be increased (about 0.8–1%) to limit the wash time within three.
5. The extent of MNase digestion can be determined by digestion time and enzyme concentration. The concentration of MNase used in experiments would depend on the number of nuclei

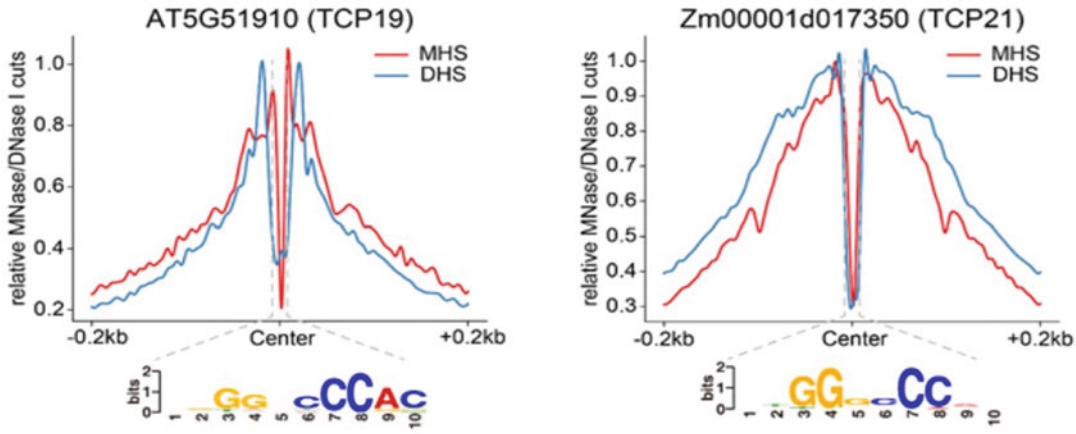


Fig. 5 Curve plot showing the presence of MHS/DHS-related footprint corresponding to the motif of transcription factors, TCP19 and TCP21. The binding motif is predicted using 50 bp DNA sequences in the MHS/DHS peak center. MNase/DNase cutting sites are calculated using MHS data with a bin window as 5 bp

and varies among different tissues or species. To make the operation easier and time saving, we recommend to test a series of enzyme concentrations with a fixed digestion time, by which at least one sample can be recovered for MH-seq.

6. Make the agarose powder fully dissolved in 1 × TBE solution; the final gel with a thickness of 0.5 cm is recommended. The resolution of the DNA band may be affected if the gel is too thick.
7. Gently pipet up and down for at least 10 times to mix well when preparing reaction solution during library construction; try to avoid foam and bubbles.
8. The adaptor needs to be diluted (by mix of 10 mM Tris–HCl, 10 mM NaCl, pH 8.0) if the sample input is <10 ng; the dilution ratio is typically 1:10 or 1:25 when the input is between 5 ng and 100 ng or less than 5 ng, respectively. Excess adaptor should be removed before PCR reaction.
9. AMPure XP Beads should be prewarmed at RT for at least 30 min before use. The orientation of tubes placing on a magnetic stand can be switched to adequately converge the beads. Overdried beads will result in lower recovery of DNA; thus it is necessary to elute DNA timely when the beads are still dark grown but all external liquid evaporates.
10. This step is for enrichment of adaptor-ligated DNA. The number of PCR cycles is determined based on the amount of starting DNA; it should be high enough to generate sufficient DNA fragments, but avoid over-amplification; the cycle number is always set between 3 and 15 cycles.

11. For more details for library preparation for Illumina, please refer to the manual of NEBNext Ultra II DNA Library Prep Kit for Illumina.
12. This step is optional. It can be skipped if the final MH-seq library lacks adaptor–dimer contamination.

Acknowledgments

We thank the Bioinformatics Center in Nanjing Agricultural University for providing facilities to assist sequencing data analysis. This work was supported by grants from the National Natural Science Foundation of China for W.Z. (32070561, 31571579) and the Fundamental Research Funds for the Central Universities (KYYJ201808).

References

1. Khan ZH, Kumar B, Dhatterwal P, Mehrotra S, Mehrotra R (2017) Transcriptional regulatory network of *cis*-regulatory elements (Cres) and transcription factors (Tfs) in plants during abiotic stress. Int J Plant Biol Res 5:1064
2. Klemm SL, Shipony Z, Greenleaf WJ (2019) Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 20:207–220
3. Tsompana M, Buck MJ (2014) Chromatin accessibility: a window into the genome. Epigenet Chromatin 7:33
4. Zhang W, Zhang T, Wu Y, Jiang J (2014) Open chromatin in plant genomes. Cytogenet Genome Res 143:18–27
5. Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S, Deaton AM, Urban JA, Larschan E, Park PJ et al (2016) MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. Nat Commun 7:11485
6. Mueller B, Mieczkowski J, Kundu S, Wang P, Sadreyev R, Tolstorukov MY, Kingston RE (2017) Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. Genes Dev 31:451–462
7. Ishii H, Kadonaga JT, Ren B (2015) MPE-seq, a new method for the genome-wide analysis of chromatin structure. Proc Natl Acad Sci U S A 112:E3457–E3465
8. Voong LN, Xi L, Sebeson AC, Xiong B, Wang JP, Wang X (2016) Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. Cell 167:1555–1570
9. Narlikar L, Ovcharenko I (2009) Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic 8:215–230
10. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19:24–32
11. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. Science 316:1497–1502
12. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436:876–880
13. Gheorghe M, Sandve GK, Khan A, Cheneby J, Ballester B, Mathelier A (2019) A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res 47:e21
14. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132:311–322
15. Zhang W, Wu Y, Schnable JC, Zeng Z, Freeling M, Crawford GE, Jiang J (2012) High-resolution mapping of open chromatin in the rice genome. Genome Res 22:151–162
16. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al (2012) The

accessible chromatin landscape of the human genome. Nature 489:75–82

17. McKay DJ (2019) Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to identify functional regulatory DNA in insect genomes. Methods Mol Biol 1858:89–97
18. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016) Open chromatin reveals the functional maize genome. Proc Natl Acad Sci U S A 113:E3177–E3184
19. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10:1213–1218
20. Lu Z, Hofmeister BT, Vollmers C, DuBois RM, Schmitz RJ (2017) Combining ATAC-seq with nuclei sorting for discovery of *cis*-regulatory regions in plant genomes. Nucleic Acids Res 45:e41
21. Nordstrom KJV, Schmidt F, Gasparoni N, Salhab A, Gasparoni G, Kattler K, Muller F, Ebert P, DEEP Consortium et al (2019) Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data. Nucleic Acids Res 47:10580–10596
22. Klein DC, Hainer SJ (2020) Genomic methods in profiling DNA accessibility and factor localization. Chromosom Res 28:69–85
23. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898
24. Voong LN, Xi L, Wang JP, Wang X (2017) Genome-wide mapping of the nucleosome landscape by micrococcal nuclease and chemical mapping. Trends Genet 33:495–507
25. Zhang W, Jiang J (2018) Application of MNase-Seq in the global mapping of nucleosome positioning in plants. Methods Mol Biol 1830:353–366
26. Lai B, Gao W, Cui K, Xie W, Tang Q, Jin W, Hu G, Ni B, Zhao K (2018) Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. Nature 562:281–285
27. Zhao H, Zhang W, Zhang T, Lin Y, Hu Y, Fang C, Jiang J (2020) Genome-wide MNase hypersensitivity assay unveils distinct classes of open chromatin associated with H3K27me3 and DNA methylation in *Arabidopsis thaliana*. Genome Biol 21:24
28. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12
29. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359
30. 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:R137
31. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192
32. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–W208

Chapter 4

Post-bisulfite Adaptor Tagging with a Highly Efficient Single-Stranded DNA Ligation Technique

Fumihito Miura and Takashi Ito

Abstract

Post-bisulfite adaptor tagging (PBAT) is a procedure for efficiently preparing a sequencing library for whole-genome bisulfite sequencing (WGBS). The original version of the PBAT protocol was highly efficient, such that it helped realize library preparation from samples of limited amounts. However, two rounds of random priming reactions employed in the original protocol limited further improvement of the PBAT protocol in terms of read length and mapping rate. In this chapter, an improved version of the PBAT protocol called tPBAT is described.

Key words Whole-genome bisulfite sequencing (WGBS), Methylome, Post-bisulfite adaptor tagging (PBAT), DNA methylation, Single-stranded DNA ligation, TACS ligation

1 Introduction

Whole-genome bisulfite sequencing (WGBS) is a method for measuring cytosine 5-methylation levels at both the whole-genome scale and single nucleotide resolution. Despite the unsurpassed spec, the cost of sequencing and difficulties in library preparation and data analysis have made WGBS impractical for many laboratories. However, the recent reduction in sequencing cost made WGBS a practical choice for methylome analysis. This chapter describes one of the most efficient protocols for library preparation of WGBS, termed tPBAT. Post-bisulfite adaptor tagging (PBAT) was originally developed in 2012 [1], and the protocol has been described in detail [2]. tPBAT is an improved version of PBAT with a highly efficient single-stranded DNA ligation technique, which we reported in 2019 [3].

Bisulfite treatment has been the sole principle for discriminating the 5-methylation status of individual cytosines by sequencing. However, bisulfite treatment severely damages DNA, causing fragmentation and loss of DNA. Therefore, methods that use bisulfite

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_4,

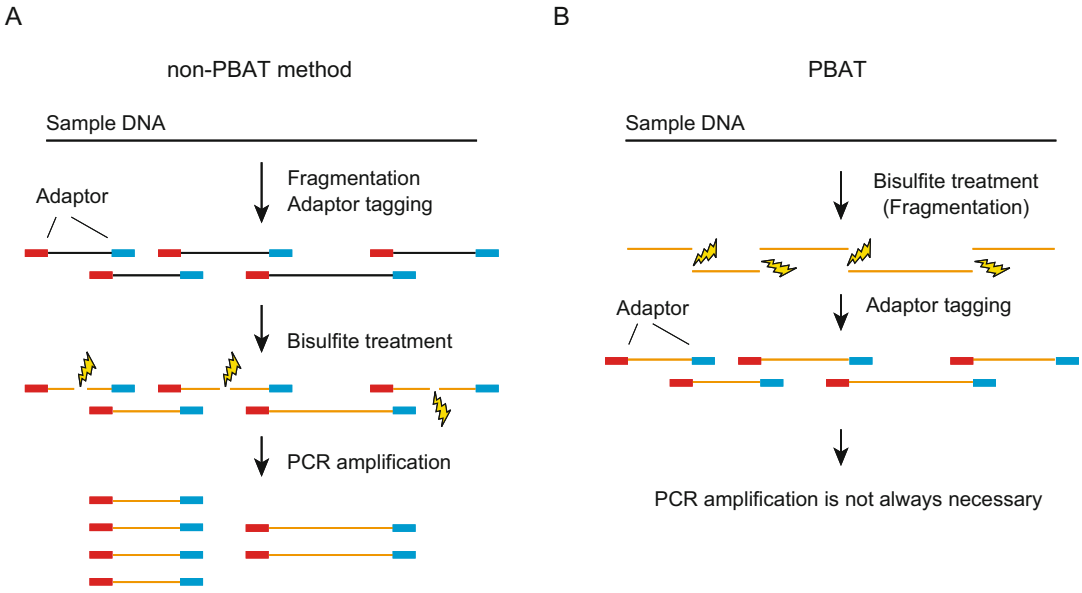


Fig. 1 The principle of post-bisulfite adaptor tagging (PBAT). (**a**) Procedures before PBAT. The bisulfite treatment causes a fatal loss of library structure. (**b**) Adaptor tagging after bisulfite treatment can avoid the loss of library structure

treatment must consider this limitation. WGBS was first reported by two independent groups in 2009 [4, 5]. However, these protocols required additional micrograms of DNA, which meant that they were extremely inefficient in library preparation. As a library of molecules, DNA should be attached to two different adaptors at both ends. However, the bisulfite treatment of adaptor-tagged DNA destroys this fundamental structure, resulting in a low yield (*see* Fig. 1a). To avoid this yield-reducing effect of the previous procedures, we devised a plan to invert the order of bisulfite treatment and adaptor tagging. Of course, bisulfite treatment causes DNA fragmentation, but because adaptor tagging is performed after the bisulfite treatment, no destructive effect of bisulfite treatment on adaptor-tagged DNA would occur, which is the core concept of PBAT (*see* Fig. 1b) [1].

Since bisulfite-treated DNA (BS-DNA) is single-stranded, adaptor tagging with conventional T4 DNA ligase-based or tagmentation-based methods cannot be used for the implementation of PBAT. Therefore, we designed the initial version of the PBAT protocol with two rounds of random priming (RP) on BS-DNA (rPBAT; *see* Fig. 2a) [1]. However, RP has several drawbacks that need to be addressed (*see* Fig. 3). First, RP causes shrinkage of the library insert (*see* Fig. 3a). A primer binds to the target DNA somewhere else during the RP reaction, and DNA polymerase extends the primer. While the sequence information downstream of the binding site should be passed to the library,

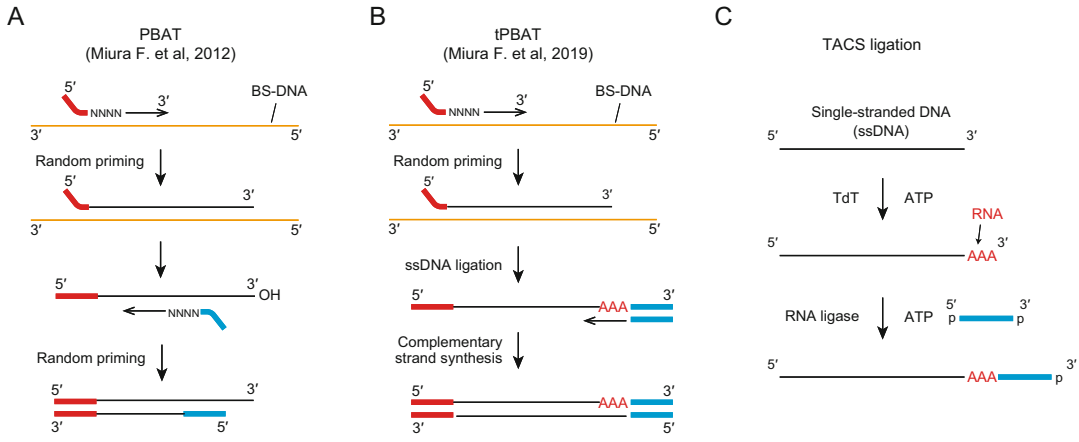


Fig. 2 TACS ligation. (**a**) The original PBAT protocol performs two rounds of random priming. (**b**) In the tPBAT, the second random priming is replaced with a ssDNA ligation technique called TACS ligation. (**c**) The principle of TACS ligation. By combining TdT-mediated ribotailing and RNA ligase, a highly efficient ligation of ssDNA was realized

the upstream sequence would never be sequenced. Thus, the more RP is repeated, the shorter the insert becomes (*see* Fig. 3a). The mean insert size of the rPBAT library is approximately 140 nucleotides, whereas the commonly used length of current Illumina sequencers is 300 nucleotides.

Therefore, to fully exploit the ability of current DNA sequencers, the insert length must first be improved. Second, since the RP occurs between DNA fragments, chimeric sequences would be generated to increase the number of reads unmappable to the reference genome (*see* Fig. 3b). The mapping rate of rPBAT reads is usually approximately 70%, whereas the rates of WGBS reads produced by methods other than rPBAT typically reach 80%. In addition, RP tends to prime more GC-rich regions, causing a GC-dependent bias in genomic coverage (*see* Fig. 3c).

To overcome the drawbacks of RP, we developed a novel procedure for highly efficient adaptor ligation of ssDNA, called TACS ligation (*see* Fig. 2c). The tPBAT is an improved protocol that replaced one of the two RPs in rPBAT with TACS ligation (*see* Fig. 2b). While the GC-dependent mapping bias has not been resolved by tPBAT because it still employs RP, the insert length and mapping rate of reads were greatly improved in tPBAT [1]. This new protocol has been applied for several methylome analyses [6–13], proving its usefulness in practice.

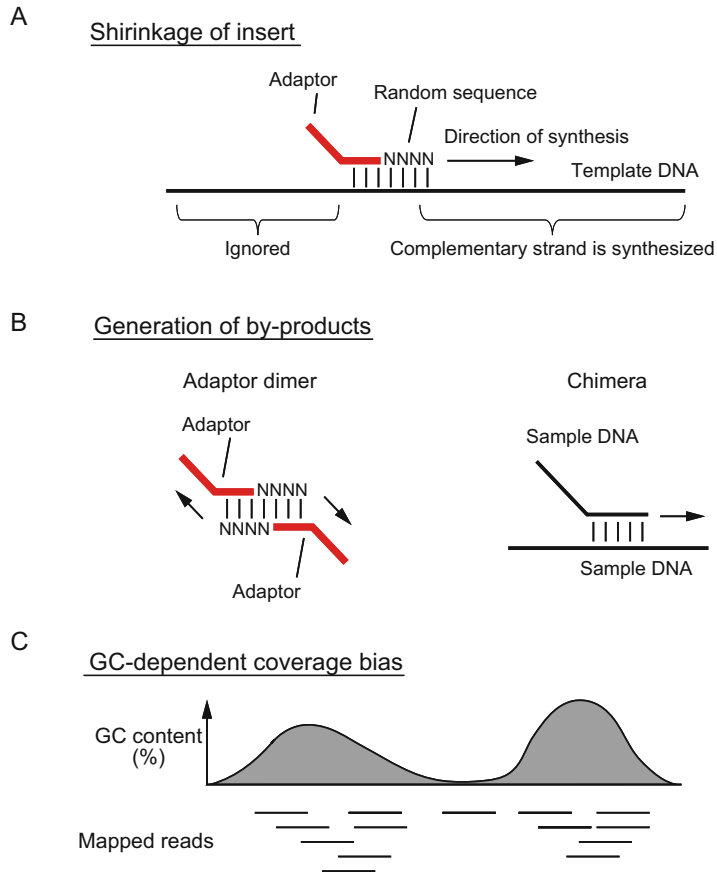


Fig. 3 The three drawbacks of random priming. (**a**) Random priming causes the loss of target sequence. (**b**) Random priming produces chimera sequences. (**c**) Random priming tends to collect GC-rich sequences

2 Materials

2.1 Commercial Kits, Enzymes, and Consumables

1. Zymo Research EZ DNA methylation-Gold Kit.
2. New England Biolabs Klenow Fragment (3–5 exo-) (50,000 unit/mL) (*see* **Note 1**).
3. Nacalai Tesque polyethylene glycol (PEG) 400.
4. Nacalai Tesque polyethylene glycol (PEG) 6000.
5. Takara Bio Terminal Deoxynucleotidyl Transferase (TdT).
6. Epicentre CircLigase II.
7. Qiagen Protease K (20 mg/mL).
8. Promega rATP (10 mM).
9. Promega Unmethylated lambda DNA.
10. Beckman Coulter AMPure XP.

11. Thermo Fisher Scientific Qubit ssDNA Assay Kit.
12. Thermo Fisher Scientific Qubit dsDNA Assay HS Kit.
13. Takara Bio Library quantitation kit.

2.2 Solutions

1. 300 bp cutoff solution: 19% (v/v) PEG 400, 10 mM Tris–HCl (pH 8.0), 1 M NaCl (*see* **Note 2**).
2. Buffer B2: 3 M guanidine hydrochloride, 20% (v/v) Tween 20.
3. 2.5 × TACS buffer: 125 mM HEPES-KOH, pH 7.5, 12.5 mM $MgCl_2$, 1.25% (v/v) Triton X-100, 50% (w/v) PEG 6000 (*see* **Note 3**).
4. dNTP solution: 2.5 mM dATP, 2.5 mM dCTP, 2.5 mM dGTP, 2.5 mM dTTP.
5. Denaturing loading dye: 95% (v/v) formamide, 0.01% (w/v) bromophenol blue.

2.3 Oligonucleotides (OPC Grade)

1. PEA2-N4: 5′-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN N-3′.
2. PA-anti-PEA1 P: 5′-phosphate-AGA TCG GAA GAG CAC ACG TCT GAA CTC CAG TCA C-phosphate-3′.
3. Primer-3: 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CT-3′.
4. Indexing primer: 5′-CAA GCA GAA GAC GGC ATA CGA GAT [Index Sequence] GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC-3′ (*see* **Note 4**).

3 Methods

3.1 Bisulfite Treatment

1. Prepare CT Conversion Reagent with adding 900 μL water, 300 μL of M-Dilution Buffer, and 50 μL M-Dissolving Buffer to a tube of CT Conversion Reagent. Completely dissolve the powder of CT Conversion Reagent by mixing at room temperature for at least 10 min. Prepare prior to use.
2. Mix 130 μL of CT conversion reagent, 100 ng of DNA (*see* **Note 5**), and 1 ng of unmethylated lambda DNA (*see* **Note 6**) in a PCR tube, and adjust the total volume to 150 μL with water.
3. Incubate the tubes at 98 °C for 10 min and at 64 °C for 150 min.
4. Load 600 μL of M-Binding buffer on a Zymo Spin column.
5. Load the sample to the Zymo Spin column.
6. Mix the sample and M-Binding buffer by inverting the Zymo Spin column.

7. Spin at 10,000 × *g* for 1 min and discard the flowthrough.
8. Wash the column with 200 μL of M-Wash.
9. Add 200 μL M-Desulfonation buffer to the column, and incubate at room temperature for 15 min.
10. Wash the column with 200 μL of M-Wash.
11. Repeat the previous step again.
12. Transfer the column to a new tube.
13. Add 20–40 μL of M-Elution buffer directly to the matrix of the column.
14. Centrifuge for 30 s at full speed to elute the DNA.
15. Use 1 μL eluent to measure the amount of DNA using the Qubit ssDNA Assay Kit (*see* **Note 7**).

3.2 Random Priming

1. Mix the bisulfite-treated DNA, 5 μL of 10 × NEBuffer 2, 4 μL of dNTP solution, and 1 μL of 100 μM PEA2 N4 in a PCR tube, and adjust the total volume to 50 μL with water.
2. Incubate at 95 °C for 3 min and at 4 °C for 5 min.
3. Add 1 μL of Klenow fragment (3′–5′ exo-) to the reaction.
4. Incubate the mixture at 4 °C for 15 min, and increase the temperature at a rate of +1 °C/min to 37 °C. After maintaining the reaction at 37 °C for 30 min, inactivate the enzyme by heating at 70 °C for 10 min.
5. Add 50 μL AMPure XP to the reaction, keep the tube stand at room temperature for 5 min, place the tube on a magnetic stand to collect the beads, and remove the supernatant.
6. Add 200 μL of 300 bp cutoff solution, resuspend the beads, place the tube on a magnetic stand, and remove the supernatant.
7. Repeat the previous step once.
8. After rinsing the beads with 200 μL 70% (v/v) ethanol, elute the DNA with 12 μL of 10 mM Tris–HCl, pH 8.5.
9. Take 1 μL DNA for measuring DNA amount with Qubit dsDNA Assay HS Kit (*see* **Note 8**).

3.3 TACS Ligation

1. Mix 10 μL of 2.5× TACS reaction buffer, 11 μL of the purified DNA in the previous step, 1 μL of 30 μM PA-anti-PEA1 P, and 1 μL of 10 mM ATP in a PCR tube.
2. Incubate at 95 °C for 5 min and at 4 °C for 5 min.
3. Add the 1 μL of 15 U/μL TdT and 1 μL of 100 U/μL CircLigase II to the reaction.
4. Incubate at 37 °C for 30 min, 65 °C for 120 min, and 95 °C for 5 min.

3.4 Primer Extension (1)

1. Add 5 μL of 10 × Gene Taq Universal Buffer, 4 μL of 2.5 mM dNTPs, 1 μL of Indexing primer, 1 μL of 2.5 U/μL Hot Start GeneTaq, and 14 μL of water to the reaction after TACS ligation.
2. Incubate at 95 °C for 3 min, 45 °C for 3 min, and 72 °C for 30 min.
3. Add 20 μL of Buffer B2 and 5 μL of 20 mg/mL proteinase K to the reaction.
4. Incubate at 50 °C for 15 min.
5. Add 50 μL AMPure XP, and then incubate for 5 min at room temperature.
6. Place the tube on a magnetic stand, wait until the beads are collected, and remove the supernatant.
7. Wash the beads with 200 μL of 300 bp cutoff solution.
8. Repeat the previous step once.
9. Rinse the beads with 200μL 70% (v/v) ethanol.
10. Remove the residual solution completely.
11. Add 40 μL of 10 mM Tris–acetate to resuspend the beads, place the tube on the magnetic stand to separate the beads, and transfer the supernatant to a new PCR tube.
12. Measure the DNA concentration by using 1 μL of purified DNA and Qubit dsDNA HS kit (*see* **Note 9**).

3.5 Primer Extension (2)

1. Add 5 μL of 10 × Gene Taq Universal Buffer, 4 μL of 2.5 mM dNTPs, 1 μL of 60 μM Primer-3, and 1 μL of 5 U/μL Hot Start GeneTaq to 39 μL of the purified DNA in the previous step.
2. Incubate at 94 °C for 3 min, 45 °C for 5 min, and 72 °C for 30 min.
3. Add 50 μL AMPure XP and incubate for 5 min at room temperature.
4. Place the tube on a magnetic stand and collect the beads.
5. Wash the beads with 200 μL of 300 bp cutoff solution.
6. Repeat the previous step once.
7. Rinse the beads with 200 μL 70% (v/v) ethanol.
8. Remove the residual solution completely.
9. Add 26 μL of 10 mM Tris–acetate to resuspend the beads, place the tube on the magnetic stand to separate the beads, and transfer the supernatant to a new PCR tube.
10. Take 1 μL of the purified DNA to measure the concentration with Qubit dsDNA HS kit.

3.6 Library QC

1. Thaw the contents of the Library Quantitation kit at room temperature (*see* **Note 10**).
2. Calculate the number of wells required as following (*see* **Note 11**):

Number of wells
= (Number of samples + number of standards + negative control + 1)
× multiplying factor

3. Prepare a master mix solution by mixing 10 μL/well of Terra PCR Direct TB Green Premix (2×), 4 μL/well of 5 × Primer Mix, 0.4 μL/well of 50 × ROX Reference Dye, and 3.6 μL/well of water to prepare a master mix solution. Multiply the volume of each reagent with the number of wells calculated in **step 2** (*see* **Note 12**).
4. Dispense 18 μL of the master mix into PCR tubes.
5. Dilute the libraries with 10 mM Tris–HCl (pH 0.0) at appropriate dilution rates (*see* **Note 13**).
6. Add either 2 μL of templates, i.e., standard, non-templated control, or diluted libraries, to a PCR tube containing the master mix.
7. Cap and place the PCR tubes in a real-time PCR machine.
8. Perform PCR amplification with the following program: 95 °C for 1 min; 35 cycles of three-step incubations at 95 °C for 10 s, 60 °C for 15 s, and 68 °C for 45 s; melt curve analysis.
9. Prepare electrophoresis device.
10. Mix 1 μL of PCR-amplified DNA with 10 μL of denaturing loading dye, incubate at 70 °C for 5 min, and load 5 μL of sample on a 6% Novex TBE-Urea gel.
11. Run electrophoresis at 300 V until the blue dye reaches two-thirds of the gel.
12. Stain the gel with SYBR Gold nucleic acid gel stain, and take a photograph (*see* **Note 14**).

3.7 Sequencing

1. Mix the libraries appropriately (*see* **Note 15**).
2. Adjust the concentration of the library mixture (*see* **Note 16**).
3. Run sequencer (*see* **Note 17**).

4 Notes

1. NEB provides a Klenow fragment (3–5 exo-) at two different concentrations. The concentrated version (50,000 units/mL) is appropriate for this protocol.
2. The results of size cutoff may vary depending on the preparation of the cutoff solutions. PEG 400 is a viscous liquid, which may cause difficulties in the accurate quantitation of liquids. Therefore, an evaluation of the cutoff solution at least once after each preparation is strongly recommended. An example of such an evaluation is shown in Fig. 4.
3. A concentrated PEG 6000 solution is highly viscous, and dissolving it at a high concentration requires long incubation with heating. Overnight incubation at 50 °C is sufficient to dissolve PEG 6000. Since the air bubbles held by the flakes of PEG 6000 sometimes cause inefficient dissolution, removing the air bubbles by brief centrifugation before heating would help the dissociation (*see* Fig. 5a). After preparation, the 2.5 × TACS buffer was cloudy, but it became transparent during storage (*see* Fig. 5b). The highly viscous 2.5 × TACS buffer is difficult to dispense with the usual micropipettes. For this purpose, we use

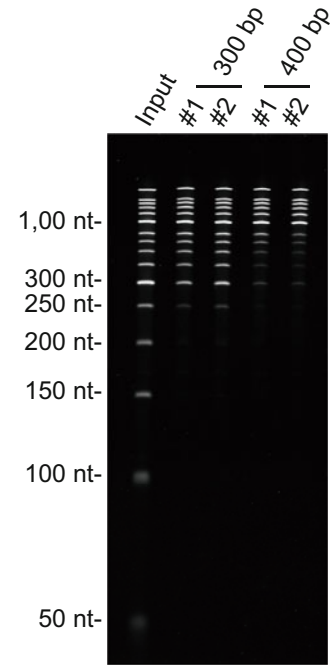


Fig. 4 Evaluation of the size cutoff solution. Two different lots of size–cutoff solutions of 300 bp and 400 bp were compared. For the model, a 50 bp DNA ladder from Takara Bio Inc. was used. Slight differences in the yields of the bands were observed

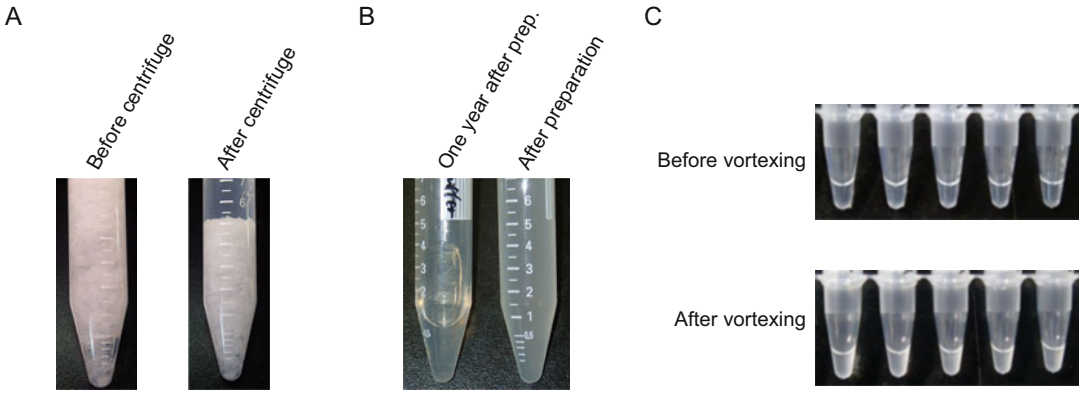


Fig. 5 Preparation of solution that contains 50% (w/v) PEG#6000. (**a**) Air bubbles (left) can be removed by brief centrifugation (right). (**b**) Just after the preparation of 2.5 × TACS buffer, it is cloudy (right), but it becomes transparent during storage (left). (**c**) The reaction is transparent after assembling (top). It becomes cloudy after mixing (bottom)

Table 1
Index numbers and index sequences

Index number	Index sequence	Index number	Index sequence	Index number	Index sequence
1	CGTGAT	9	CTGATC	18	GCGGAC
2	ACATCG	10	AAGCTA	19	TTTCAC
3	GCCTAA	11	GTAGCC	20	GGCCAC
4	TGGTCA	12	TACAAG	21	CGAAAC
5	CACTGT	13	TTGACT	22	CGTACG
6	ATTGGC	14	GGAACT	23	CCACTC
7	GATCTG	15	TGACAT	25	ATCAGT
8	TCAAGT	16	GGACGG	27	AGGAAT

a positive displacement pipette Microman E from Gilson. The 2.5 × TACS buffer is stored at room temperature. The solution is transparent, but occasionally becomes cloudy, especially when the room temperature is low. The solution containing 1 × TACS buffer after assembling the reaction is transparent but becomes slightly cloudy upon vortexing (*see* Fig. 5c). In our experience, TACS ligation proceeded successfully with a cloudy reaction.

4. For indexing primers, insert one of the index sequences listed in Table 1 into the Index Sequence.
5. As a starting material, 100 ng of DNA is recommended. From this DNA amount, sufficient reads to cover the mammalian genome can be obtained without PCR amplification. In

addition, the yield of DNA can be measured at each step with Qubit-based measurements if library preparation is started with more than 100 ng of DNA.

6. In most methylome analyses, ~1% (w/w) of unmethylated lambda DNA was spiked into the sample DNA to calculate the conversion rate of bisulfite treatment.
7. Usually, approximately 70% of the input DNA is recovered from human and mouse samples. This quantitation step is essential because if DNA is not detected in this step, good library preparation cannot be expected.
8. Typical yields of this step are usually 40–80% of the input DNA.
9. The yields of this step are 10–30% of the input DNA.
10. We evaluated three commercially available kits from Kapa Biosciences, Takara Bio, and Toyobo, and all the kits worked well.
11. For reliable quantitation, multiplication of wells is recommended. In our experience, duplexing is enough for the quantitation; the multiplying factor in the equation is set to 2.
12. Preparation of master mix is strongly recommended. It is an effective for reproducible quantitation.
13. Usually, a 1000 fold dilution is appropriate for the quantitation tPBAT library.
14. For the analysis of the size distribution of libraries, electrophoresis of PCR-amplified libraries is recommended. This is because the library before amplification contains many DNA by-products, and the analysis of libraries before PCR amplification sometimes causes misinterpretations of their size distribution. Denaturing gel electrophoresis is recommended because the library after the PCR plateau is difficult to analyze on a native gel. Typical electrophoresis patterns are shown in Fig. 6.
15. As the nucleotide composition of DNA after bisulfite treatment is extremely biased and the amount of C is quite limited, adding a C-containing sequencing library to the WGBS library to compensate for the low C signal is very important. Without adding such a library, one might fail to obtain a sufficient amount of sequence data. The addition of the PhiX control at 20% of the input library is widely used for this compensation. The spiking ratio can be reduced to 5% using a library with high GC content [14]. Alternatively, if two WGBS libraries of different topologies can be prepared, library spiking can be eliminated [3].
16. The concentration of the library prepared with tPBAT is sometimes low, and condensation might be required. For this, the AMPure XP-based procedure is easy to perform. Mix the same volume of AMPure XP with the library mixture, incubate at room temperature for 5 min, remove on a magnetic stand,

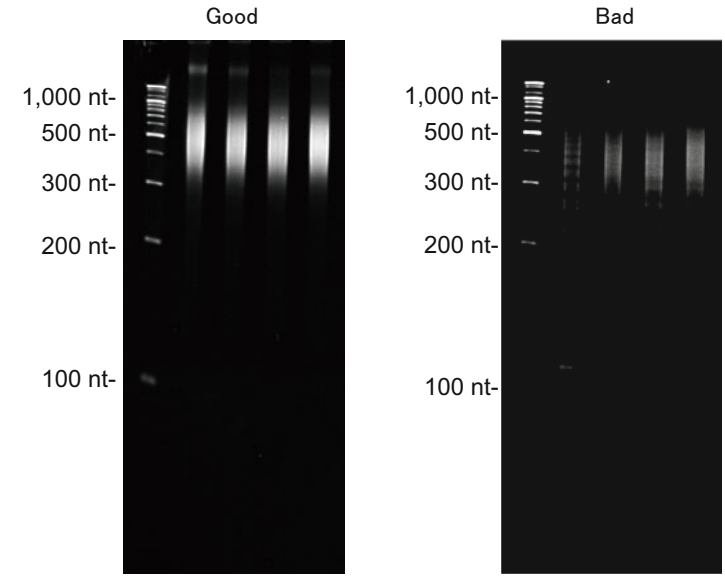


Fig. 6 The size distribution of amplified libraries. The amplified DNA fragments were analyzed by denaturing gel electrophoresis. 1 μL of amplified DNA was mixed with 10 μL of formamide and heat-denatured at 70 °C for 5 min. The sample was then loaded onto a 6% Novex TBE-Urea Gel (Thermo Fisher Scientific). After separation, the gel was stained with SYBR Gold Gel Stain (Thermo Fisher Scientific) and photographed. The successfully prepared libraries showed smear patterns on the left, whereas unsuccessful ones showed ladder patterns

rinse with 70% ethanol, and elute the library mixture with an appropriate volume of 10 mM Tris–acetate. Usually, there is no need for re-quantitation with qPCR.

17. The libraries prepared with the tPBAT protocol were successfully sequenced on Illumina sequencers, including MiSeq, NextSeq, HiSeq, and NovaSeq.

Acknowledgments

We are grateful for Yukiko Shibata and Miki Miura for their technical supports. This work was supported by a JSPS KAKENHI grant to F.M. (20H03243) and T.I. (17H06305). This work was also supported by the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP20am0101103.

References

1. Miura F, Enomoto Y, Dairiki R, Ito T (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res 40(17):e136. https://doi.org/10.1093/nar/gks454
2. Miura F, Ito T (2018) Post-bisulfite adaptor tagging for PCR-free whole-genome bisulfite sequencing. Methods Mol Biol 1708:123–136. https://doi.org/10.1007/978-1-4939-7481-8_7
3. Miura F, Shibata Y, Miura M, Sangatsuda Y, Hisano O, Araki H, Ito T (2019) Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res 47(15):e85. https://doi.org/10.1093/nar/gkz435
4. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452(7184):215–219. https://doi.org/10.1038/nature06745
5. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(3):523–536. https://doi.org/10.1016/j.cell.2008.03.029
6. Araki H, Miura F, Watanabe A, Morinaga C, Kitaoka F, Kitano Y, Sakai N, Shibata Y, Terada M, Goto S, Yamanaka S, Takahashi M, Ito T (2019) Base-resolution methylome of retinal pigment epithelial cells used in the first trial of human induced pluripotent stem cell-based autologous transplantation. Stem Cell Rep 13(4):761–774. https://doi.org/10.1016/j.stemcr.2019.08.014
7. Matsuda T, Irie T, Katsurabayashi S, Hayashi Y, Nagai T, Hamazaki N, Adefuin AMD, Miura F, Ito T, Kimura H, Shirahige K, Takeda T, Iwasaki K, Imamura T, Nakashima K (2019) Pioneer factor neuroD1 rearranges transcriptional and epigenetic profiles to execute microglia-neuron conversion. Neuron 101 (3):472-485:e477. https://doi.org/10.1016/j.neuron.2018.12.010
8. Hamazaki N, Kyogoku H, Araki H, Miura F, Horikawa C, Hamada N, Shimamoto S, Hikabe O, Nakashima K, Kitajima TS, Ito T, Leitch HG, Hayashi K (2021) Reconstitution of the oocyte transcriptional network with transcription factors. Nature 589(7841): 264–269. https://doi.org/10.1038/s41586-020-3027-9
9. Sangatsuda Y, Miura F, Araki H, Mizoguchi M, Hata N, Kuga D, Hatae R, Akagi Y, Amemiya T, Fujioka Y, Arai Y, Yoshida A, Shibata T, Yoshimoto K, Iihara K, Ito T (2020) Base-resolution methylomes of gliomas bearing histone H3.3 mutations reveal a G34 mutant-specific signature shared with bone tumors. Sci Rep 10 (1):16162. https://doi.org/10.1038/s41598-020-73116-x
10. Shirane K, Miura F, Ito T, Lorincz MC (2020) NSD1-deposited H3K36me2 directs de novo methylation in the mouse male germline and counteracts polycomb-associated silencing. Nat Genet 52(10):1088–1098. https://doi.org/10.1038/s41588-020-0689-z
11. Tanaka S, Ise W, Inoue T, Ito A, Ono C, Shima Y, Sakakibara S, Nakayama M, Fujii K, Miura I, Sharif J, Koseki H, Koni PA, Raman I, Li QZ, Kubo M, Fujiki K, Nakato R, Shirahige K, Araki H, Miura F, Ito T, Kawakami E, Baba Y, Kurosaki T (2020) Tet2 and Tet3 in B cells are required to repress CD86 and prevent autoimmunity. Nat Immunol 21(8):950–961. https://doi.org/10.1038/s41590-020-0700-y
12. Yoshida K, Maekawa T, Ly NH, Fujita SI, Muratani M, Ando M, Katou Y, Araki H, Miura F, Shirahige K, Okada M, Ito T, Chatton B, Ishii S (2020) ATF7-dependent epigenetic changes are required for the intergenerational effect of a paternal low-protein diet. Mol Cell 78 (3):445-458:e446. https://doi.org/10.1016/j.molcel.2020.02.028
13. Kuribayashi W, Oshima M, Itokawa N, Koide S, Nakajima-Takagi Y, Yamashita M, Yamazaki S, Rahmutulla B, Miura F, Ito T, Kaneda A, Iwama A (2021) Limited rejuvenation of aged hematopoietic stem cells in young bone marrow niche. J Exp Med 218(3). https://doi.org/10.1084/jem.20192283
14. Suzuki M, Liao W, Wos F, Johnston AD, DeGrazia J, Ishii J, Bloom T, Zody MC, Germer S, Greally JM (2018) Whole-genome bisulfite sequencing with improved accuracy and cost. Genome Res 28(9):1364–1371. https://doi.org/10.1101/gr.232587.117

Chapter 5

Perturbation of Gene Regulation by Genome Editing

Nan Cher Yeo and George M. Church

Abstract

The RNA-guided endonuclease Cas9 can be converted into a programmable transcriptional repressor. Here we describe a set of protocols for using the catalytically inactive dead Cas9 (dCas9)-based tools, including the bipartite super repressor consisting of the KRAB and MeCP2 domains, to achieve efficient and scalable gene silencing in mammalian cells.

Key words CRISPR-Cas transcriptional repressor, dCas9-KRAB-MeCP2, Single and multiplex gene silencing

1 Introduction

The ability to selectively regulate gene expression is critical for understanding gene function and for manipulating cellular identity. RNA interference (RNAi) is a useful method for targeted gene knockdown and has been widely used for large-scale library screens. RNAi, however, has several limitations—in particular, RNAi-based knockdown suffers from broad off-target effects along with incomplete knockdown [1–3]. Custom DNA-binding proteins, such as zinc finger proteins or transcription activator-like effectors (TALEs), fused to transcriptional repressor domains allow for selective gene suppression, but are not scalable due to the fact that each desired target gene necessitates the creation of a new protein [4–6]. The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9 system, which confers adaptive immunity within bacteria and archaea, has been rapidly adopted for genome engineering in a wide range of cells and model systems [7–12]. Cas9 is an endonuclease that can be directed by short "single guide RNA (sgRNA)" molecules to specific DNA sequence, provided that a protospacer-adjacent motif (PAM) is proximal to the target. Because the target locus is dictated by the sgRNA sequence, this

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_5,

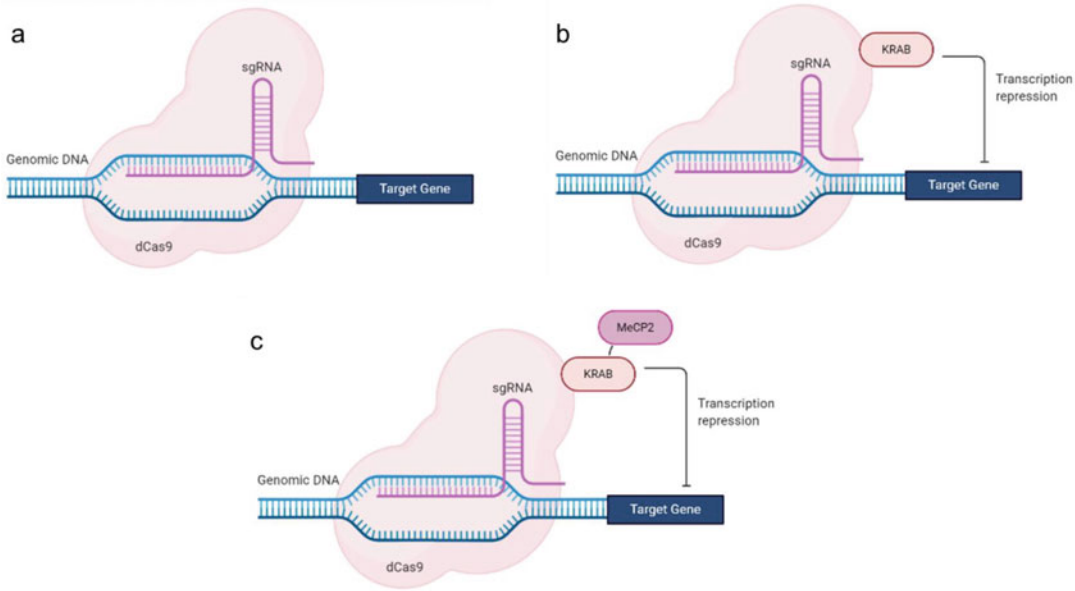


Fig. 1 First and second generation of CRISPR dCas9-based repressors. (**a**) A nucleolytically inactive dCas9 which binds but does not cut the target DNA can act as a transcriptional repressor by sterically blocking passage of transcriptional initiation or elongation in bacteria system. (**b**) Fusing a strong repression domain KRAB to dCas9 provides effective transcriptional repression in mammalian cells, likely by its ability to recruit additional repression protein complex and by blocking transcriptional initiation. (**c**) Fusing several transcriptional regulators such as KRAB and MeCP2 in tandem to dCas9 further enhance the efficiency in gene repression. (Figure was created with BioRender.com)

system is much simpler and more flexible to use. Therefore, it has quickly become a method of choice for diverse genome targeting purposes.

Beyond gene editing, CRISPR-Cas9 system has been repurposed for programmable, targeted gene regulation by developing a catalytically inactive dead Cas9 (dCas9), which remains competent for DNA binding but lacks endonuclease activity [13]. The Cas9 enzyme has two catalytic domains (HNH and RuvC) that mediate DNA cleavage, resulting in DNA double-strand breaks at the target site proximal to PAM. A single point mutation within each of the catalytic domains, namely, D10A and H840A for *Streptococcus pyogenes* Cas9 (SpCas9), causes complete loss of DNA cleavage activity. The resultant dCas9 complexed with a sgRNA molecule can bind tightly to the target locus and sterically repress transcription by blocking either transcriptional initiation or elongation in prokaryotic cells (*see* Fig. 1a). This approach, however, is not effective in mammalian cells. Notably, further improvement in transcriptional inhibition can be achieved by addition of a strong transcriptional repressor, such as the Krüppel-associated box (KRAB) domain, to dCas9, likely by its ability to modify chromatin environment of the target locus [14, 15] (*see* Fig. 1b). Although the

fusion of dCas9 and KRAB can repress genes in mammalian system, its efficiency varies between target sites, and only a small fraction of sgRNAs was effective when paired with the fusion proteins [16, 17]. Consequently, a more robust and efficient tool is highly desirable.

Previous work has shown that by fusing several transcriptional regulators to dCas9 in tandem, a synergistic increase in regulation can be achieved [18–20]. We and others have exploited this strategy to develop the second generation of dCas9 tools for a wide range of regulatory manipulation, including targeted transcriptional repression [16, 17]. In one of our recent works, based on rationale-guided design, we systematically screened and tested more than 80 transcriptional repressors and subsequently engineered a highly effective dCas9-based bipartite repressor consisting the KRAB and methyl CpG binding protein 2 (MeCP2) domains [17] (*see* Fig. 1c). Through detailed characterization, we demonstrated that the new repressor is superior to previous CRISPR repressor platform in various contexts, including single and multiplexed gene targeting and large-scale genetic and epistasis screens. This chapter describes a set of protocols for using dCas9-based tools including the bipartite dCas9-KRAB-MeCP2 super repressor for targeted gene suppression in mammalian cells.

2 Materials

2.1 Molecular Cloning

1. CRISPR repressor and sgRNA expression plasmids: pcDNA3.3_TOPO-dCas9 (Addgene plasmid #47316), dCas9-KRAB (Addgene plasmid #110820), dCas9-KRAB-MeCP2 expression plasmids (Addgene plasmid #110821), and pSB700 gRNA cloning vectors (Addgene plasmid #64046).
2. gRNA oligonucleotides for target sequence (*see* Subheading 3.1 for the discussion on target selection and Subheading 3.2.1 for the design of gRNA oligos).
3. Sequencing primers for validating gRNA oligo inserts (5′-GA GGGCCTATTTCCCATGATTCC-3′) (*see* Subheading 3.2.2 for cloning of gRNA oligos).
4. TE buffer: 10 mM Tris–Cl, pH 7.5, and 1 mM EDTA.
5. BsmBI-v2 type-II restriction enzyme.
6. Agarose and gel electrophoresis apparatus.
7. QIAquick Gel Extraction Kit.
8. T4 DNA Ligase.
9. Chemically competent *E. coli* cells (such as NEB® 5 alpha or NEB® Stable).

10. Luria-Bertani (LB) broth and agar bacterial growth medium.
11. Ampicillin antibiotics.
12. QIAprep Spin Miniprep Kit.

2.2 Tissue Culture, Transfection, and RNA Extraction

1. Cell line: For validation, human embryonic kidney HEK293T cell line is recommended. For working with other cell lines, *see* **Note 1** for additional discussions.
2. DMEM with 10% heat-inactivated FBS.
3. Lipofectamine 2000.
4. Opti-MEM.
5. Puromycin antibiotics (3 μg/mL).
6. Trypsin-EDTA (0.25%).
7. DPBS (pH: 7.0–7.3).
8. RNeasy Plus Mini Kit.

2.3 Analysis of Targeted Gene Expression

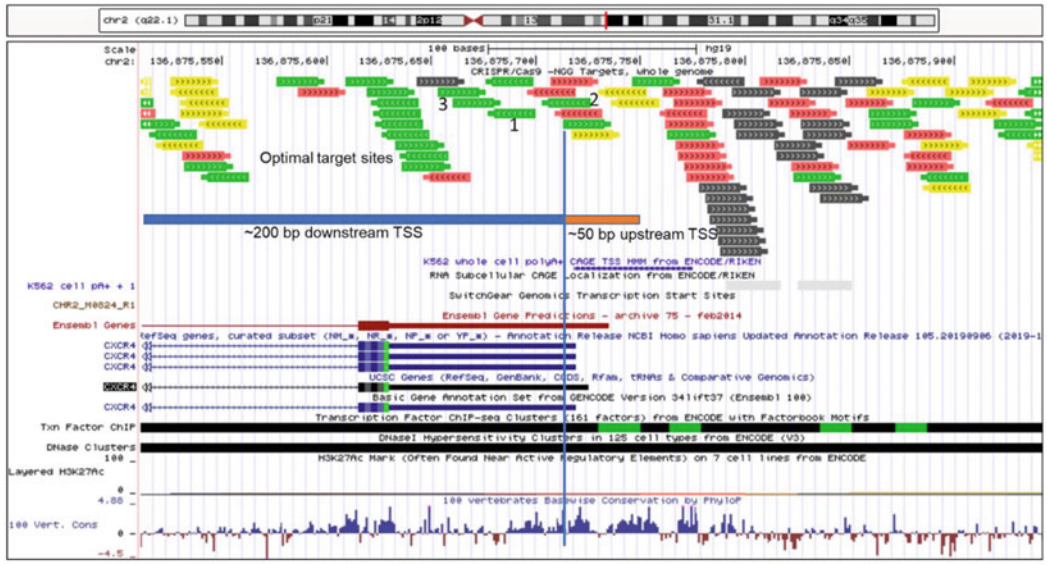
1. Commercial cDNA synthesis kit.
2. Commercial SYBR Green quantitative PCR (qPCR) kit, such as KAPA SYBR Fast 2× qPCR master mix.
3. DNA oligonucleotides for qPCR analysis of target genes.

3 Methods

3.1 Target Selection

To modulate gene expression, a dCas9 regulator is directed within or near the promoter of a target gene. Thus, the target window is not as broad as for gene knockout via Cas9 cutting. For transcriptional repression, the optimal gRNA-targeting window is approximately 50 bp upstream or 200 bp downstream of the transcription start site (TSS) of target genes [15, 17]. Because the targeting location greatly influences the efficacy of gene knockdown, it is critical to have precise information on the location of TSS. Previous studies have shown that cap analysis gene expression sequencing (CAGE-seq) which directly captures the mRNA 5′ cap provides the most accurate TSS mapping [21, 22]. We recommend using CAGE-seq together with other regulatory information such as histone marks and transcription factor chromatin immunoprecipitation (TF ChIP) to identify TSS and active promoters (*see* Fig. 2a). Currently there is no robust tools available for selecting gRNA sequences for targeted gene regulation. Like Cas9, target sites for all dCas9 tools must be followed by the sequence of PAM ("NGG" trinucleotides at 3′ end for SpCas9 and dCas9). DNA sequences targetable by SpCas9 or dCas9 can be identified in the USCS Genome Browser. Each CRISPR target site is annotated with predicted specificity (off-target effects) and predicted efficiency (on-target effects) by various algorithms through the tool CRISPOR to aid target selection (*see* Fig. 2a) [23]. We recommend using

a



b

Guide Sequence	CAGGTAGCAAAGTGACGCCG
Protospacer Adjacent Motif (PAM)	AGG
MIT Guide Specificity Score	94
Efficiency: Doench et al. 2016 Score	86% (62)
Efficiency: Moreno-Mateos (In-vitro) Score	69% (54)
Efficiency: Doench et al. 2014 Score	15
Bae et al. Out-of-Frame Score	67

Fig. 2 Integrated gene annotations in the USCS Genome Browser (genome.ucsc.edu) hg19 and CRISPR target selection. (**a**) A screenshot of the USCS Genome Browser showing a genomic region on chromosome 2 (chr2: 136,875,512-136,875,943) containing part of the CXCR4 gene in human genome (hg19). In the browser, multiple tracks, including CRISPR/Cas9 targetable sites, predicted TSS by CAGE and SwitchGear, and gene and regulatory annotations, are displayed. Based on the predicted TSS and gene annotations, an optimal targeting window (−200 bp to +50 bp proximal to TSS) was marked (blue and orange). Each CRISPR target site is annotated with predicted specificity and efficiency scores provided by the CRISPOR track. Shades of gray stand for sites that are hard to target specifically. Targets that are specific in the genome with different predicted efficiencies are shown in different colors: green (highest efficiency) > yellow > red > blue (lowest efficiency). Examples of three good target sites are marked 1–3. (**b**) The guide sequence can be viewed or retrieved by selecting the desired target site. Shown is an example of the genomic information and specificity/efficiency score associated with the selected target sequence 1

the integrated tools in the USCS Genome Browser (genome.ucsc.edu) to aid the selection of gRNA targets for dCas9-based repressors, including dCas9-KRAB-MeCP2, noting that both location and sequence are of approximately equal importance in design of gRNA targets for gene repression. Finally, while in silico prediction can minimize off-target activities through careful design of gRNAs, for any gene of interest, multiple gRNAs of different sequences should be tested to ensure that the observed phenotype is indeed due to an on-target effect.

3.2 gRNA Design and Construction

3.2.1 Design of gRNA Oligos

1. Select gRNAs using the "CRISPR Targets" track in UCSC genome browser as mentioned in Subheading 3.1 (*see* Fig. 2a). Consider both the predicted specificity (off-target effects) and efficiency (on-target effects) score when choosing targets. In general, optimal gRNAs are those with the greatest on-target efficiency and the least off-target activity.
2. Select the desired genomic target or guide sequences (without including "NGG" PAM) (*see* Fig. 2b). Create a reverse complement (RC) of the guide sequences.
3. For cloning gRNA oligos into the pSB700 backbone, modifications of the guide sequences are required. To do that, append "CACCG" to the 5′ end of the guide sequence. Append "AAAC" to the 5′ end of the RC guide sequence, and append an additional "C" to the 3′ end of the RC sequences (*see* **Note 2** for additional discussion on modifications of guide sequence). As an example, the final oligos for guide sequence " CAGGTAGCAAAGTGACGCCG " and its RC guide sequence (underlined) should be:

   Final forward oligo = **CACCG**CAGGTAGCAAAGTGACG CCG

   Final reverse oligo = **AAAC**CGGCGTCACTTTGCTACCTG**C**
4. Order the final forward and reverse oligonucleotides indicated above.
5. Re-suspend lyophilized forward and reverse oligonucleotides to a final concentration of 100 μM in 1× TE buffer.

3.2.2 Cloning of gRNA Oligos into pSB700 Vectors

6. To anneal gRNA oligos, aliquot 1:1 the forward and reverse oligonucleotides (e.g., 10 μL each) into PCR tubes, vortex, and spin down the oligo mixtures at 100 × g for 15 s. Heat the mixed oligonucleotides to 94 °C for 2 min, and gradually cool to room temperature to facilitate their annealing. Dilute the annealed oligos 20-fold in nuclease-free water for cloning. Store the annealed oligos at −20 °C if they will not be used immediately.
7. Digest the selected pSB700 guide cloning vector with BsmBI enzyme for 1 h at 55 °C by mixing the following: 1–5 μg pSB700 guide cloning vector, 4 μL NE buffer 3.1, 1 μL BsmBI, up to 40 μL distilled water.
8. Run the digestion products on 1% agarose gel. Purify the digested pSB700 (~9 kb) using QIAquick gel extraction kit, and elute the DNA into 15 μL TE buffer to get concentrated samples.
9. Ligate the annealed gRNA oligos from **step 6** into the pre-digested pSB700 vector from **step 8** by mixing the following: 1 μL annealed (diluted) gRNA oligos from **step 6**, 250–500 ng BsmBI-digested pSB700 vector from **step 8**,

2 μL 10× T4 DNA ligase reaction buffer, 1 μL T4 DNA ligase, up to 20 μL distilled water. Incubate the reaction at 16 °C overnight.

10. Next day, transform 1 μL of the ligation reaction into 11 μL of chemically competent *E. coli* cells (such as NEB® 5-alpha or NEB® Stable) according to the manufacturer's protocol. Note: For lentiviral plasmids such as pSB700 vectors, NEB stable cells will provide more consistent plasmid yields.
11. Plate 80 μL of the transformed cells on an ampicillin selection LB plate. Incubate the plated cells overnight at 37 °C for NEB 5-alpha or at 30 °C for NEB Stable.
12. Next day, inspect the LB plates for colony formation. Approximately 50–100 colonies are typically found on the cloning plates and no or very few colonies on the no-insert negative control plate.
13. Pick two to three colonies to analyze for correct insertion of the gRNA oligos via direct colony sequencing. A primer 5′-GAGGGCCTATTTCCCATGATTCC-3′, which targets the U6 promoter upstream of the gRNA cloning site, can be used to sequence the oligo inserts. Note: Several sequencing providers (such as Genewiz) offer services that perform Sanger sequencing directly from bacterial colonies, which greatly reduces the time and costs by eliminating the need to purify plasmids prior to sequencing.
14. After verification of gRNA oligo inserts, inoculate cells from a colony that contains the correct plasmid into ampicillin selection LB liquid, and culture cells overnight. Next day, isolate the plasmid DNA from the cultured cells using QIAprep Spin Miniprep Kit.

3.3 Cell Culture and Transfection

Activity of the dCas9-based transcriptional repressors has been validated in a variety of mammalian cell lines. The protocol below is for HEK239T cells and for transient transfection of dCas9 repressors and gRNA expression plasmids. This protocol is useful when the desired endpoint can be reached via transient expression of dCas9 repressors and the gRNA. (*See* **Notes 1** and **3** for discussions on working with other cell lines and for generating cell lines stably expressing repressor/gRNA).

1. HEK293T cells are maintained in DMEM supplemented with 10% heat-inactivated fetal bovine serum and passaged before reaching 80% confluency. Cells are maintained in an incubator set at 37 °C supplemented with 5% CO_2.
2. Approximately 50,000 cells are seeded per well in 24-well plates, and the next day they are transfected using Lipofectamine 2000 along with the repressor and gRNA expression

plasmids according to the manufacturer's protocol. For each well of cells, 200 ng of dCas9-based repressor, 50 ng of gRNA expression plasmids, and 50 ng of puromycin resistance plasmids are transfected. For multiplex gene repression, 10 ng of each gRNA plasmid are transfected per well. Note: A non-targeting gRNA negative control should always be included in each experiment for comparing the level of targeted gene knockdown with experimental groups.

3. Cells are treated with 3 μg/mL puromycin at 24 h post-transfection to enrich for transfected cells.
4. 72 h after transfection, cells are collected for RNA extraction using RNeasy Plus mini kit. Proceed to analysis of targeted gene knockdown as described in Subheading 3.4.

3.4 Analysis of Targeted Gene Knockdown by Quantitative Real-Time PCR (qPCR)

The level of gene knockdown induced by dCas9 repressor/gRNA can be quickly verified by qPCR analyses (*see* **Note 4** for additional discussions).

1. For each gene-targeting sample or negative control, 500 ng of total RNA is used to make cDNA using a commercial cDNA synthesis kit.
2. Endogenous gene expression is analyzed by qPCR using gene-specific primers (*see* **Note 5** for additional discussions on primer design) and 2× SYBR Green qPCR master mix according to the manufacturer's protocols.
3. Expression of each target transcript is normalized to that of the housekeeping genes (e.g., ACTB), and relative gene expression can be calculated via the $2^{-\Delta\Delta Ct}$ method compared to the negative control groups [24].

4 Notes

1. Experimental conditions may need to be optimized for each cell line. For other cell lines, an initial comparison of different transfection reagents (e.g., Lipofectamine 3000, FuGENE HD, and nucleofection) is recommended.
2. The gRNA oligos are compatible for cloning into a BsmBI-digested pSB700 at a site downstream of a U6 promoter driving the transcription of the gRNA. The "CACC" sequence ensures that the oligo is compatible with the overhangs of the BsmBI-digested pSB700 vector. The "G" is a requirement of Polymerase III promoters and ensures the efficient initiation of the transcription of the gRNA. The "AAAC" sequence ensures that the oligo is compatible for cloning into the BsmBI-digested pSB700 vector. The additional "C" on the 3′ end is needed to anneal the sequence to the initiating "G" added to the forward oligo.

3. Studies that require stable expression or integration of dCas9 repressor/gRNA into the genome, such as studies performed in vivo or for genome-wide phenotypic screens, piggyBac transposon, or lentivirus-mediated delivery method, can be used [17].
4. Whole-transcriptome RNA sequencing can be used to assess on-target and potential off-target efficiency of CRISPR repressors [17].
5. The qPCR primers can be designed using free webtool Primer3 (https://bioinfo.ut.ee/primer3-0.4.0/). The primers should be tested initially to ensure that the PCR product is specific to the target gene transcripts, in order to ensure reliable quantification of the target gene expression.

Acknowledgments

This work was supported by the NIH grants RM1 HG008525 and P50 HG005550 to G.M.C. and the University of Alabama-Birmingham, School of Medicine and Precision Medicine Institute, startup fund to N.C.Y.

References

1. Sigoillot FD, Lyman S, Huckins JF, Adamson B, Chung E, Quattrochi B, King RW (2012) A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. Nat Methods 9:363–366
2. Krueger U, Bergauer T, Kaufmann B, Wolter I, Pilk S, Heider-Fabian M, Kirch S, Artz-Oppitz-C, Isselhorst M, Konrad J (2007) Insights into effective RNAi gained from large-scale siRNA validation screening. Oligonucleotides 17: 237–250
3. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS (2003) Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol 21:635–637
4. Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31: 397–405
5. Joung JK, Sander JD (2013) TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol 14:49–55
6. Margolin JF, Friedman JR, Meyer WK, Vissing H, Thiesen HJ, Rauscher FJ (1994) Krüppel-associated boxes are potent transcriptional repression domains. Proc Natl Acad Sci U S A 91:4509–4513
7. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816–821
8. Cong L, Ran FA, Cox D et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823
9. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. Science 339:823–826
10. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM (2013) Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res 41:4336–4343
11. Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh J-RJ, Joung JK (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol 31:227–229
12. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153:910–918

13. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152:1173–1183
14. Gilbert LA, Larson MH, Morsut L et al (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154: 442–451
15. Gilbert LA, Horlbeck MA, Adamson B et al (2014) Genome-scale CRISPR-mediated control of gene repression and activation. Cell 159: 647–661
16. La Russa MF, Qi LS (2015) The new state of the art: Cas9 for gene activation and repression. Mol Cell Biol 35:3800–3809
17. Yeo NC, Chavez A, Lance-Byrne A et al (2018) An enhanced CRISPR repressor for targeted mammalian gene regulation. Nat Methods 15: 611–616
18. Chavez A, Scheiman J, Vora S et al (2015) Highly efficient Cas9-mediated transcriptional programming. Nat Methods 12:326–328
19. Tanenbaum ME, Gilbert LA, Qi LS, Weissman JS, Vale RD (2014) A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell 159:635–646
20. Zalatan JG, Lee ME, Almeida R et al (2015) Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. Cell 160:339–350
21. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H et al (2014) A promoter-level mammalian expression atlas. Nature 507:462–470
22. Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7:542–561
23. Concordet J-P, Haeussler M (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res 46:W242–W245
24. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C (T)) method. Methods San Diego Calif 25: 402–408

Chapter 6

Analysis of Neutrophil Morphology and Function Under Genetic Perturbation of Transcription Factors In Vitro

Julia Salafranca, Zhichao Ai, Lihui Wang, Irina A. Udalova, and Erinke van Grinsven

Abstract

Hoxb8 cells are immortalized myeloid progenitors that maintain their multipotent potential and can be differentiated into neutrophils. Genetic modification of Hoxb8 cells can be used as a model system for the functional analysis of regulators of neutrophil maturation and effector functions, such as transcription factors. Here we describe the generation of transcription factor (TF) knockout Hoxb8 cell lines in vitro with the lentivirus (lenti)CRISPR-Cas 9 technique. After their differentiation into neutrophils, the study of their maturation profile, morphology, and effector functions, including NETosis, phagocytosis, and ROS production, is described.

Key words Functional analysis, Hoxb8 cells, LentiCRISPR/Cas9, Maturation, Morphology, NETosis, Neutrophil, Phagocytosis, ROS production, Transcription factor

1 Introduction

Neutrophils are important effector cells in innate immunity, possessing a wide range of effector functions, including reactive oxygen species (ROS) production, phagocytosis, and NETosis. These properties enable neutrophils to rapidly respond to stimulation and orchestrate protective immunity. However, excessive infiltration and activation of neutrophils at the inflammatory sites can cause tissue damage, leading to intense local and systemic inflammation [1]. The transcriptional regulation of neutrophil activation and function has only recently received a due attention and begun to be explored [2–4].

HoxB8 provides a powerful model system for in vitro production of neutrophils and generation of stable knockout lines using the CRISPR/Cas9-mediated system. HoxB8 murine myeloid

The authors "Julia Salafranca" and "Zhichao Ai" are equally contributed to this chapter.

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_6,

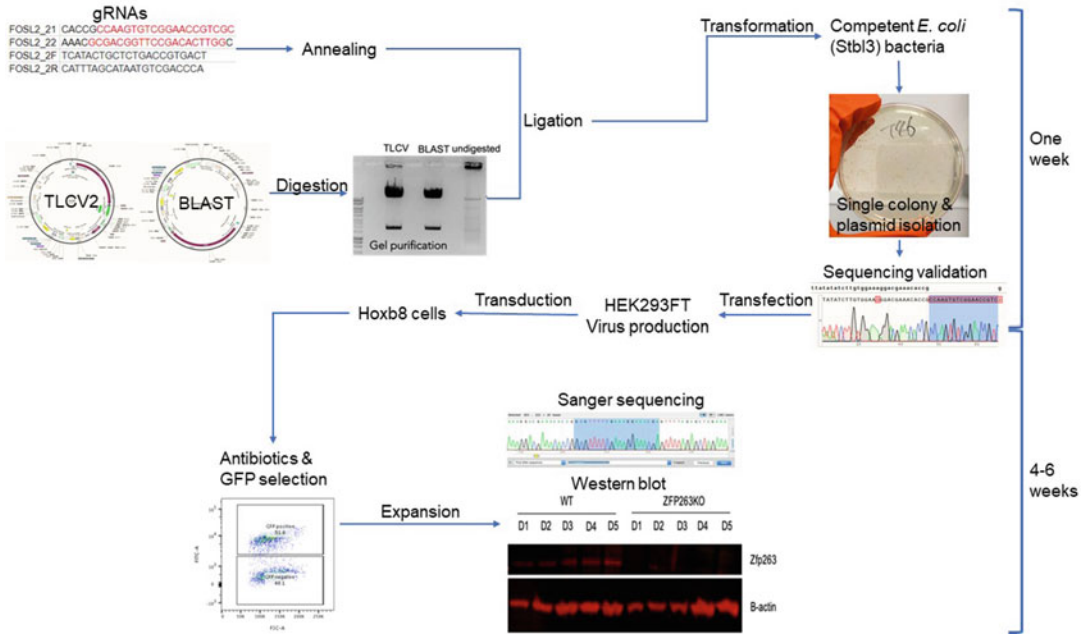


Fig. 1 In vitro generation of Hoxb8 knockout cell lines using the lentiCRISPR/Cas9 technique. If the TF knockout is successful, Hoxb8-derived neutrophil maturity is assessed by studying morphology through imaging cytospins and surface marker expression by flow cytometry. To validate the functional role of the selected TFs in neutrophil effector functions, phagocytosis, generation of ROS, and the formation of NETs are examined. We have extensively used this approach to determine the transcription factor networks that shape neutrophil responses [2]

progenitors are conditionally immortalized by estrogen-driven expression of homeobox oncoprotein HoxB8, which enables the arrest of myeloid differentiation resulting in an infinite myeloid progenitor expansion [5]. In the presence of estrogen, HoxB8 progenitors are similar to granulocyte-monocyte progenitor (GMP) cells, capable of self-renewal by cell division. Upon estrogen deprivation and in the presence of G-CSF, these HoxB8 progenitors progressively differentiate into mature neutrophils with a full range of neutrophil effector functions. Terminally differentiated HoxB8 neutrophils are morphologically indistinguishable from bone marrow mature neutrophils [5]. This makes HoxB8 neutrophils an ideal model to study myeloid cell differentiation and investigate cellular functions of neutrophils under normal or inflammatory conditions.

In vivo generation of genetically modified neutrophils is expensive and time-consuming. We use the lentiCRISPR/Cas9 technique to generate TF knockout Hoxb8 cell lines (Fig. 1), providing a platform that overcomes the inability to genetically modify primary neutrophils ex vivo.

2 Materials

2.1 Common Reagents and Equipment

1. Sterile, nuclease-free water (ddH_2O).
2. 1.5 mL microcentrifuge tubes.
3. 37 °C incubator.
4. Microcentrifuge.
5. Centrifuge.
6. Dulbecco's phosphate-buffered saline (DPBS).
7. Hoxb8 progenitor medium: add 10% fetal bovine serum (FBS), 1% penicillin/streptomycin (P/S), 30 μM β-mercaptoethanol, and 4% stem cell factor (SCF) containing supernatant to RPMI 1640 medium. Add 10 μM estradiol to the medium directly before use.
   (a) SCF containing supernatant is obtained by collecting the supernatant of CHO cells cultured in 10% FBS 1% P/S and 30 μM β-mercaptoethanol and filtering it through a 0.2 μm filter.
8. 10 cm Petri dish.
9. 96-well plate.
10. Phorbol 12-myristate 13-acetate (PMA) (Sigma-Aldrich).
11. BD Cytofix (BD Biosciences) or equivalent fixation buffer.
12. FACS buffer: add 0.1% bovine serum albumin (BSA), 0.01% sodium azide (NaN_3), and 1 mM EDTA to DPBS.
13. Flow cytometer.
14. PBST: add 0.05% Tween-20 to DPBS.

2.2 Cloning Target Oligonucleotides into the Lentiviral Vector

1. LentiCRISPR plasmid: either TLCV2 (Addgene) with puromycin selection cassette and doxycycline-inducible GFP expression or BLAST (Addgene) with blasticidin selection cassette.
2. 10× T4 DNA Ligase Reaction Buffer (NEB).
3. 0.1 M DTT (Invitrogen).
4. Digestion BsmBI enzyme, supplied with 10× NEBuffer™ r3.1 (NEB).
5. Temperature-controlled mixer.
6. 1% agarose gel: Prepare by mixing 2 g of UltraPure agarose (Invitrogen) in 150 mL MilliQ water. Microwave until the solution is clear (2–3 min). Add 8 mL 25× TAE buffer (12.1% w/v Tris base, 2.86% glacial acetic acid, 5% 0.5 M EDTA, top up to 1 L with MilliQ water). Add EtBr.
7. Casting tray, gel chamber, and power pack to solidify and run the gel.

8. Quick-load Purple 1 kB DNA ladder (NEB) or equivalent.
9. 5× DNA Loading Buffer Blue (BioLine).
10. QIAquick PCR purification kit (Qiagen).
11. Designed oligonucleotides targeting the TF of interest (guide (g)RNA), reconstituted in ddH_2O at 100 μM.
12. Thermal cycler.
13. 0.2 mL PCR tubes.
14. T4 DNA Ligase (NEB).
15. One Shot™ Stbl3™ chemically competent *E. coli* (ThermoFisher), which includes the S.O.C. medium.
16. Ampicillin culture plates: dilute LB agar to 37 g/L in MilliQ water. Mix and autoclave. Add ampicillin at 0.1 μg/mL (Sigma-Aldrich). Plate 10 mL per 10 cm Petri dish. Store plates for a maximum time of 1 month before use.
17. Water bath that can reach 42 °C.
18. Shaking incubator.
19. L-shape spreader or glass plating beads.
20. LB broth: Prepare by dissolving LB broth powder at 25 g/L (#L3522-250G, Merck) in MilliQ water. Mix until clear and autoclave. Add ampicillin at 0.1 μg/mL (Sigma-Aldrich).
21. QIAprep spin miniprep kit (Qiagen).
22. NanoDrop One Microvolume UV-vis spectrophotometer (ThermoScientific).
23. U6pro primer (5′-GAGGGCCTATTTCCCATGATT-3′).

2.3 Viral Production and Transduction of Hoxb8 Cells

1. Packaging vectors pMD2.G (Addgene) and pCMV-dR8.91 (Addgene).
2. HEK293T cells (InvivoGen).
3. HEK cells media: 10% FBS DMEM. Culture without P/S unless indicated.
4. Trypsin.
5. Opti-MEM (Gibco).
6. Lipofectamine (Invitrogen).
7. 0.45 μm filters.
8. Hoxb8 cells, kindly provided by B. Walzog (LMU Biomedizinsches Centrum).
9. Polybrene (Merck).
10. Puromycin (ThermoFisher) for TLCV2-based knockouts.
11. Blasticidin (InvivoGen) for BLAST-based knockouts.

2.4 Selection of Cells Successfully Transduced with TLCV2 by GFP Positivity

1. Doxycycline (Merck).
2. Cell sorting medium: 2% FBS DPBS.

2.5 Deriving Mature Neutrophils from Transduced Hoxb8 Progenitors

1. Neutrophil differentiation washing medium: 1% FBS, 1% P/S DPBS.
2. Neutrophil differentiation medium: Add 10% FBS, 1% P/S, 30 μM β-mercaptoethanol, and 4% SCF containing supernatant to RPMI 1640 medium. Add 20 ng/mL G-CSF directly before use.

2.6 Knockout Validation at Protein and DNA Level

1. DNeasy Blood and Tissue Kit (Qiagen).
2. Designed primers for the TFs.
3. Cell lifter.
4. 1% Tx-100 lysis buffer: 1% TX-100, 10% glycerol, 1 mM EDTA, 150 mM NaCl, 50 mM Tris (pH 7.8) in distilled water.
5. Roche protease inhibitors (Sigma-Aldrich). Use one tablet per 10 mL lysis buffer.
6. Qubit Protein Assay Kit (ThermoFisher).
7. Qubit 2.0 Fluorometer (ThermoFisher).
8. 4 × Laemmli loading buffer (Bio-Rad).
9. Full range rainbow molecular weight marker (GE Healthcare).
10. Precast NUPAGE 4–12% Bis-Tris gel (Invitrogen).
11. NuPAGE MOPS SDS running buffer (ThermoFisher); dilute stock 1:20 in distilled water.
12. Heat block.
13. 0.2 μm polyvinylidene difluoride (PVDF) membrane (GE Healthcare).
14. 100% methanol.
15. Filter papers.
16. 10× Transfer buffer: 3% w/v Tris base, 14.4% w/v glycine in distilled water.
17. 1× Transfer Buffer: 10% 10× Transfer buffer, 15% methanol, and 75% distilled water.
18. Blocking buffer/secondary antibody buffer: 5% w/v milk powder in PBST. Shake well until diluted.
19. Primary antibody buffer: 2% BSA in PBST.
20. ECL chemiluminescent substrate solution (GE Healthcare).
21. X-ray film (Super RX; FujiFilm).
22. AGFA Cruis-60 automatic film processor (AGFA-Gaevert).
23. ReBlot Plus Strong Antibody Stripping Solution (Merck).

2.7 Morphology Assessment by Cytospin

1. Slide clip, filter, and cytofunnel.
2. Microscopy slides.
3. Cytospin centrifuge.
4. Epredia™ Shandon™ Kwik-Diff™ Stains (Fisher Scientific).
5. Brightfield microscope.

2.8 Maturation Assessment by Flow Cytometry

1. eBioscience™ Foxp3/Transcription Factor Staining Buffer Set (ThermoFisher).
2. Fixable viability dye.
3. Antibodies and Fc block (Table 1).

2.9 Neutrophil Effector Functions Assays

1. Dihydrorhodamine 123 (DHR) (Invitrogen).
2. *E. coli* BioParticles conjugate (Invitrogen). Reconstitute in 1 mL DPBS.
3. Poly-L-lysine.

Table 1
List of antibodies used for flow cytometry and Western blot

Marker	Fluorochrome	Dilution	Provider
Fc block		1 in 100	BD Biosciences
Far red fixable viability dye	APC-Cy7	1 in 1000	Life Technologies
CD11b	BV510	1 in 200	BioLegend
cKit	BV421	1 in 200	BioLegend
Ly6C	BV-785	1 in 200	BioLegend
Ly6G	BV-711	1 in 200	BioLegend
CXCR2	PE	1 in 200	BioLegend
CXCR4	APC	1 in 200	Life Technologies
CD101	PE-Cy7	1 in 200	Life Technologies
CD3	Percp-Cy5.5	1 in 200	BioLegend
CD19	Percp-Cy5.5	1 in 200	BioLegend
TCRαβ	Percp-Cy5.5	1 in 200	BioLegend
NK1.1	Percp-Cy5.5	1 in 200	BioLegend
Ter119	Percp-Cy5.5	1 in 200	BioLegend
CD11c	Percp-Cy5.5	1 in 200	BioLegend
Siglec F	Percp-Cy5.5	1 in 200	BioLegend
CD115	Percp-Cy5.5	1 in 200	BioLegend
Citrullinated histone 3		1 in 200	Abcam
MPO		1 in 200	Hycult

4. Nunc Lab-Tek II 8 well chamber slide (ThermoFisher).
5. Ionomycin (Merck).
6. 4% paraformaldehyde in DPBS.
7. Antibody blocking buffer: 2% BSA in PBST.
8. Antibody binding buffer: 0.1% BSA in PBST.
9. Secondary antibodies goat anti-rabbit conjugated and goat anti-mouse conjugated.
10. Nuclear staining dye, like DAPI or Hoechst.
11. ProLong™ Gold Antifade Mountant with DAPI (Invitrogen).
12. Sample slides and cover slides.
13. Nail polish to seal the slide.

3 Methods

3.1 Cell Culture

1. HEK293T cells are cultured in a 10 cm plate with 10 mL 10% FBS DMEM media (*see* **Note 1**).
2. Hoxb8 progenitor cells are cultured in 75 cm^2 flasks with 50 mL Hoxb8 progenitor medium. Passage Hoxb8 cells 1 in 10 to maintain the culture when they reach one million cells/mL (*see* **Notes 2** and **3**).

3.2 gRNA Design

For gRNA design, use UniProt web tool to check the protein sequence and isoforms to target common exons encoding them. Use CHOPHOP web tool to find candidate gRNAs based on the target site sequence. These 20 bp oligonucleotides need to be flanked on the 3′ end by a 3 bp NGG (protospacer adjacent motif) PAM sequence (*see* **Note 4**). The specificity of designed gRNA is validated with the Benchling web tool, to ensure that the targeting sequences are within the protein-encoding region and have no major off-target site. The designed gRNA sequences need to be downstream of CACCG, and then the complementary oligonucleotide strand is between AAAC and C, so the gRNA sequences can be conjugated with the Cas9 complex when expressed together.

Primers approximately 300 bp from the CRISPR PAM sequence to amplify the target sequence by PCR are generated by CHOPCHOP.

3.3 Cloning Target Oligonucleotides into the Lentiviral Vector

1. Digest the lentiviral vector on ice by mixing 5 μg of the lenti-CRISPR plasmid, 4 μL of 10× NEBuffer™ r3.1, and 0.4 μL of DTT, and top up to 37 μL with ddH2O. Subsequently, add 3 μL of BsmBI enzyme and spin down to mix. Incubate at 37 °C for 3 h shaking at 600 rpm in a pre-heated thermal mixer (*see* **Note 5**).

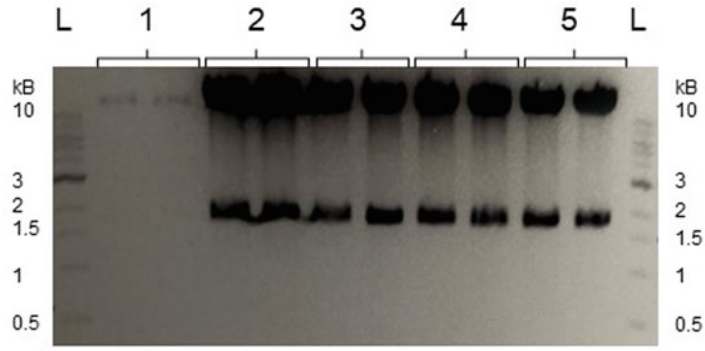


Fig. 2 LentiCRISPR plasmid digestion by BsmBI enzyme. Gel electrophoresis shows that TLCV2 was successfully digested by BsmBI in samples 2, 3, 4, and 5. 1 kB ladder (L) used

2. Dilute 5 μL of sample in 15 μL of DNA loading buffer. Confirm lentiCRISPR plasmid digestion by gel electrophoresis (1% agarose gel) at 120 V for 1 h. Two thick bands of 10 kB and 2 kB indicate successful plasmid digestion (Fig. 2) (*see* **Note 6**).
3. Purify the digested lentiCRISPR plasmid using the QIAquick PCR Purification kit according to the manufacturer's instructions (on ice).
4. Measure the DNA concentration using NanoDrop.
5. Spin down the lyophilized oligonucleotides, and reconstitute them in ddH_2O at a 100 μM concentration. For every TF, use 1 μL of each oligonucleotide, 1 μL of 10× T4 DNA Ligase Reaction Buffer, and 7 μL of ddH_2O in a PCR tube (*see* **Note 7**). Anneal the oligonucleotides by heating at 95 °C for 5 min and ramping down to 25 °C (0.1 °C/s decrease) in a thermal cycler. Dilute the annealed oligonucleotides 1:200 in ddH_2O on ice.
6. For ligation of the annealed oligonucleotides with the digested vector, add in a 0.2 mL PCR tube (on ice), 50 ng of BsmBI digested plasmid, 1 μL of the diluted annealed oligonucleotides, and 1 μL of 10× T4 DNA Ligase Reaction Buffer, and top up to 10 μL with ddH_2O (*see* **Note 7**). Subsequently, add 1 μL of T4 Ligase. Without adding oligonucleotides, use the digested plasmid without T4 Ligase as a negative control and the digested plasmid with T4 Ligase as a positive control. Incubate for a minimum of 4 h at 16 °C in a thermal cycler.
7. For bacteria transformation, heat up water bath to 42 °C, S.O.C. medium to RT, and warm up the ampicillin culture plates to 37 °C. Thaw the vial with the StbI3 bacteria on ice (50 μL/vial). Add 2 μL of the ligation mix (from **step 6**) to 10 μL of bacteria without pipetting up and down. Incubate on ice for 30 min. Subsequently, heat-shock for 20 s at 42 °C in

the water bath, and immediately place the tube on ice for 2 min. Add 250 μL of S.O.C. medium at RT to each vial, and shake horizontally at 37 °C for 1 h in a shaking incubator—leave the lids of the microcentrifuge tubes open, and close the rack with a lid (*see* **Note 8**). Add 100 μL of the transformation mix to ampicillin culture plates, and spread evenly (*see* **Note 9**). Leave in the 37 °C incubator for a minimum of 16 h (*see* **Note 10**).

8. Pick a single colony from the plate, and add it to a Falcon tube containing 20 mL LB broth (*see* **Notes 11** and **12**). Incubate overnight at 37 °C shaking horizontally in a shaking incubator, with the tube lid semi-open and secured with tape. Use the QIAprep spin miniprep kit to purify the plasmid according to the manufacturer's instructions.
9. Measure the DNA concentration by NanoDrop (*see* **Note 13**).
10. Send the samples for sequencing using the U6pro primer to validate the insertion of the gRNA into the plasmid.

3.4 Viral Production and Transduction of Hoxb8 Cells

Cell preparation:

- 5 days in advance, thaw HEK293T cells.
- Culture Hoxb8 cells for 4–5 days before transduction.

Day 1:

1. Streak out bacterial cultures of packaging plasmids pCMV-dR8.2 and pCMV-VSV-G on ampicillin plates, and incubate at 37 °C overnight.
2. Split HEK293T cells into as many plates as you have gRNA and vector combinations. Seed two million HEK293T cells in a 10 cm dish in 10 mL of 10% FBS DMEM.

Day 2:

1. To purify the packaging plasmids, repeat **step 8** of Subheading 3.3.
2. Replace the HEK293T cell medium with 12 mL 10% FBS DMEM per dish.
3. Prepare transfection mix containing 200 μL Opti-MEM, 8 μg of the plasmid containing the gRNA, 4 μg of pCMV-dR8.2, and 4 μg pCMV-VSV-G packaging plasmids (*see* **Note 7**). Subsequently, add 20 μL of lipofectamine. Incubate for 30 min at RT.
4. Add the transfection mix dropwise to the HEK293T cells. Use a non-targeting vector (such as the ligated TLCV2 or BLAST plasmid) to control for any indirect effect of the lentivirus infection on the cells. Incubate cells at 37 °C for 18–20 h.

Day 3:

1. Replace half of the HEK293T medium with 10% FBS 1% P/S DMEM.

Day 4:

1. To collect the virus, aspirate the medium and filter (0.45 μm filters). When 10 mL of fresh 10% FBS 1% P/S DMEM medium is added to the cells, another viral collection can be done in 24 h. Do two to three viral collections to have enough virus-containing media if transduction needs to be repeated. Store collected viral media at −80 °C.
2. To prepare for lentivirus transduction, count and resuspend the Hoxb8 cells at 500000 cells/mL. Add polybrene to a final concentration of 6 μg/mL to increase cell membrane permeability, and aliquot 250,000 cells into each well of a 6- or 12-well plate. Add 4 mL of virus-containing media (with pre-added 6 μg/mL polybrene, 4% SCF, 30 μM β-mercaptoethanol, 10 μM estradiol, 10% FBS) to each well (*see* **Note 7**). Spin for 90 min at RT 1500 × *g* (acceleration and break at 1). Add 2 mL of fresh medium (*see* **Note 14**).

Day 5:

1. After 24 h exchange the medium by aspirating 3.5 mL of culturing medium from each well (*see* **Note 15**) and adding 3 mL of fresh medium to dilute the polybrene to a non-toxic concentration (~2 μg/mL). Check the cells daily until antibiotic selection.

Day 8:

1. 72 h after transduction, start antibiotic selection. Use puromycin for TLCV2 clones and blasticidin for BLAST clones. Dilute puromycin and blasticidin in Hoxb8 medium to 8 μg/mL and 6 μg/mL, respectively. Aspirate 3 mL of medium from the top of the well (*see* **Note 15**), and add 3 mL of antibiotic-containing medium to each well (final concentration 4 μg/mL or 3 μg/mL, respectively). Add puromycin to Hoxb8 cells that have not undergone transduction as a negative control, and culture them without puromycin as a positive control. Cells should be fully selected in 48–72 h.

Day 10/11:

1. Transfer the selected cells into a 10 cm plate with 20 mL of medium (with 10 μM estradiol and 8 μg/mL puromycin or 6 μg/mL blasticidin), and let the cells grow until the number reaches ten million (*see* **Note 16**).

3.5 Selection of Cells Successfully Transduced with TLCV2 by GFP Positivity

NB: this step is skipped for BLAST clones.

1. 24 h before FACS sorting, add 1 μg/mL doxycycline to the transduced Hoxb8 cells, and use non-transduced cells as a control (*see* **Note 17**).
2. To prepare for FACS sorting, wash the cells with DPBS, resuspend in 1 mL of 2% FBS DPBS at ten million cells/mL, and filter. Prepare tubes with 1 mL of Hoxb8 medium (with 10 μM estradiol and 8 μg/mL puromycin) to collect the sorted cells.
3. Sort GFP+ cells. Replate the sorted cells in a 6-well plate with 6 mL of medium, and when they reach one million cells/mL, transfer them to a 10 cm plate with 20 mL of media for further expansion (until ten million cells) (*see* **Note 18**).

3.6 Deriving Mature Neutrophils from Transduced Hoxb8 Progenitors

1. Spin down 2.5×10^6 transduced Hoxb8 progenitors.
2. Wash with ample 1% FBS 1% P/S DPBS to wash away the residual estradiol.
3. Resuspend cells in 13 mL of differentiation medium in a 10 cm plate.
4. Culture for 4–5 days for full neutrophil differentiation (*see* **Note 19**).

3.7 Knockout Validation at Protein and DNA Level

Western blot is used to validate the absence of the protein of interest in the knockout Hoxb8 cells.

3.7.1 Western Blot

1. If it is necessary to induce the expression of the protein of interest, differentiate the Hoxb8 TF knockout progenitor cells into mature neutrophils.
2. Harvest the cells by gently scraping the culture plate with a cell lifter, and wash 1× with PBS (centrifuge at 510*g* for 5 min).
3. Lyse the cell pellets in cold 1% Tx-100 lysis buffer supplemented with Roche protease inhibitors. Incubate on ice for 30 min, and centrifuge at 16,278*g* for 10 min at 4 °C to remove cellular debris. Collect supernatant (*see* **Note 20**).
4. Quantify protein concentration in the lysate using the Qubit Protein Assay Kit according to the manufacturer's instructions, and measure in a Qubit 2.0 Fluorometer.
5. Normalize protein sample concentration with ddH_2O, and add 4× Laemmli loading buffer (*see* **Note 21**).
6. Heat 10–20 μg of lysates in a heat block at 100 °C for 5 min, and spin to return condensation to the sample.
7. Load 15 μL of samples along with a molecular weight marker on a precast NUPAGE 4–12% Bis-Tris gel in NuPAGE MOPS SDS running buffer for 1 h at 160 V or until the blue dye gets to the bottom of the gel.

8. Before gel transfer to a PVDF membrane by wet Western blotting, briefly activate PVDF membranes in 100% methanol. Soak four sheets of filter paper with transfer buffer. Place the gel on top of two soaked filter papers, followed by the PVDF membrane and two soaked filter papers to make a Western blot sandwich. Place the assembled sandwich between two transfer buffer-soaked sponges in a transfer cassette. Remove air bubbles by carefully pressing the blot sandwich. Place the cassette in a transfer tank filled with pre-cooled transfer buffer at 72 V for 2 h, 4 °C.
9. Following protein transfer, incubate the PVDF membrane in blocking buffer for 1 h at RT with gentle shaking. Remove the blocking buffer, and incubate the membrane with primary antibodies diluted in antibody binding buffer overnight with gentle shaking at 4 °C (*see* **Note 22**). Use β-actin or GAPDH as a control. The next day, wash the membrane 3× with PBST with gentle shaking for 10 min between each wash step. Incubate the membrane with HRP-conjugated secondary antibody diluted in blocking solution for 1 h with gentle shaking at RT. Repeat the three washes. Incubate the membrane with ECL according to the manufacturer's protocol. Develop X-ray film using an AGFA Cruis-60 automatic film processor to visualize the protein bands (*see* **Note 23**).

3.7.2 Sanger Sequencing

To validate CRISPR-Cas9-mediated genomic editing, the fragment of target genome is isolated and amplified for validation at DNA level by Sanger sequencing.

1. Wash WT and knockout undifferentiated Hoxb8 cells with DPBS.
2. Isolate genomic DNA from one million cells using the DNeasy Blood and Tissue Kit following the manufacturer's instructions.
3. Measure the DNA concentration by NanoDrop.
4. Normalize concentration to 100 ng/μL with ddH_2O.
5. Amplify the fragment of interest in a PCR reaction, using the designed primers.
6. Purify the PCR amplification product using the QIAquick PCR Purification kit according to the manufacturer's instructions.
7. Measure the DNA concentration by NanoDrop (*see* **Note 13**).
8. Send the samples for sequencing using the U6pro primer.

3.8 Morphology Assessment by Cytospin

1. Pre-wash the differentiated neutrophils with DPBS twice to remove cellular debris (centrifuge at 510*g* 5 min RT and discard supernatant).
2. Resuspend at 5×10^5 cells/mL.
3. Assemble the slide clip, slide, filter card, and cytofunnel, and aliquot 100 μL of cells into the funnel. Spin at 400*g* for 5 min.
4. Air-dry the slide before staining.
5. Use Epredia™ Shandon™ Kwik-Diff™ Stains to stain the slide according to manufacturer's instructions.
6. Air-dry.
7. Use a brightfield microscope to assess the neutrophil morphology. Morphological quantification is based on nuclear size and shape, staining density of chromatin, and presence of granules to define metamyelocyte, banded neutrophils, and segmented neutrophils.

3.9 Assessing Neutrophil Maturation by Flow Cytometry

1. Wash the differentiated neutrophils with DPBS (510*g* 5 min).
2. Resuspend in FACS buffer and aliquot 100,000 cells/well in a 96-well plate.
3. Spin 2 min at 740*g* RT and discard supernatant.
4. Add 20 μL/well of Fc block (10 μg/mL in FACS buffer), and incubate 20 min on ice.
5. Without washing, add 25 μL/well of the antibody mix (Table 2). Perform the viability stain according to manufacturer's instructions. Incubate 30 min at 4 °C in the dark.

Table 2
Extracellular markers antibody panel

Marker
Far red fixable viability dye
CD11b
cKit
Ly6C
Ly6G
CXCR2
CXCR4
CD101
Lineage (CD3, CD19, TCRαβ, NK1.1, Ter119, CD11c, Siglec F, and CD115)

3.9.1 Intranuclear Staining for Flow Cytometry

NB: if desired the TF knockout can also be confirmed by studying the expression of the TF of interest in flow cytometry through intranuclear staining with a fluorochrome-conjugated antibody.

6. Add 150 μL of FACS buffer, centrifuge 2 min at 740*g* 4 °C, and discard supernatant (×2).
7. Resuspend in 100 μL of Foxp3 fixation/permeabilization working solution, and incubate for 30–60 min at 4 °C in the dark.
8. Centrifuge samples at 740*g* for 2 min at RT. Discard the supernatant.
9. Resuspend in 150 μL 1× Permeabilization Buffer, and centrifuge at 740*g* for 2 min at RT. Discard the supernatant. Repeat wash.
10. Add the intracellular antibody specific for your TF of interest diluted in 1× Permeabilization Buffer. Incubate for 30 min RT in the dark.
11. Wash twice with 150 μL 1× Permeabilization Buffer.
12. Resuspend in 200 μL of cold FACS buffer. Keep the cells in the dark at 4 °C until the analysis.
13. Analyze the cells by flow cytometry. Pre-neutrophils are defined as Lineage- CD11b + Ly6Cint CXCR4+ cKit+ CXCR2-, immature neutrophils are characterized as Lineage- CD11b + Ly6Cint CXCR4- cKit- CXCR2+ Ly6G+ CD101-, and mature neutrophils are defined as Lineage- CD11b+ Ly6Cint CXCR4- cKit- CXCR2+ Ly6Ghigh CD101+.

3.10 Neutrophil Effector Functions Assays

3.10.1 Reactive Oxygen Species (ROS) Production

1. Plate 100 μL/well of Hoxb8 neutrophils at a 20 million cells/mL concentration (2×10^6 cells/well) in duplicate in two 96-well plates. Day 5 fully differentiated Hoxb8 neutrophils can be expected to be fully competent at ROS production, whereas day 2 immature Hoxb8 neutrophils can be used as controls producing no or low levels of ROS.
2. Add 2.5 μg/mL DHR to detect ROS production and 30 μg/mL PMA to stimulate ROS production. Reduce exposure to light as much as possible.
3. Incubate for 20 min. Incubate one plate at 37 °C for ROS production to occur, and incubate the other at 4 °C as a control. Stop the reaction by putting the plates on ice.
4. Wash the cells with 100 μL/well DPBS; spin the plates at 740*g* for 2 min at 4 °C. Discard the supernatant.
5. Fix the cells with 80 μL/well cytofix for 20 min at 4 °C.

6. Wash with 150 μL/well FACS buffer, spin (740*g* 2 min 4 °C), and discard supernatant.
7. Resuspend cells in 250 μL/well FACS buffer, and analyze by flow cytometry at appropriate excitation and emission setup for DHR.

3.10.2 Phagocytosis

1. Repeat **step 1** from Subheading 3.10.1.
2. Add 10 μL/well *E. coli* BioParticles conjugate.
3. Incubate for 20 min (at 37 °C or 4 °C). Incubate one plate at 37 °C for phagocytosis to occur, and incubate the other at 4 °C as a control for surface adhesion of BioParticles. Stop the reaction by putting the plates on ice.
4. Repeat steps 4–9 from Subheading 3.7.1, and measure at appropriate excitation and emission setup to detect BioParticles fluorescence.

3.10.3 NETosis

1. Coat an 8-well chamber slide with 100 mL/chamber poly-L-lysine at 37 °C for 30 min. Remove poly-L-lysin and wash with DPBS.
2. Seed 200 μL/chamber of Hoxb8 neutrophils at 2×10^6 cells/mL into the pre-coated poly-L-lysin chambers. Day 5 fully differentiated Hoxb8 neutrophils can be expected to be fully competent at NETosis, whereas day 2 immature Hoxb8 neutrophils can be used as controls producing no or low levels of NETosis.
3. Induce NETs by stimulating the neutrophils with 10 μM PMA and 10 μM ionomycin. Add the same μL of buffer to the control samples (no PMA and ionomycin stimulation). Incubate overnight at 37 °C.
4. Gently remove the solution from the chambers and wash with DPBS (*see* **Note 24**).
5. Fix the cells with 100 μL/chamber 4% paraformaldehyde in DPBS for 30 min at RT. Subsequently, remove the 4% paraformaldehyde and add 300 μL of DPBS. Slides can be kept at 4 °C for up to 2 weeks before imaging.
6. Aspirate the DPBS and incubate the cells with 200 μL antibody blocking buffer for 20 min.
7. Rinse the cells with PBST, and add 200 μL/chamber of the primary antibodies rabbit anti-citrullinated histone 3 and mouse anti-mouse MPO at 5 μg/mL and 0.5 μg/mL, respectively, diluted in antibody binding buffer (0.1% BSA in PBST) (*see* **Note 25**). Incubate for 2 h at RT or overnight at 4 °C.
8. Rinse three times with PBST for 5 min.

9. Add 200 μL/chamber of the secondary antibodies (goat anti-rabbit conjugated and goat anti-mouse conjugated) diluted in antibody binding buffer.
10. Rinse three times with PBST for 5 min.
11. Nuclear staining can be performed with or without the chamber, depending on the microscope.
    (a) With the chamber, add 100 μL/chamber of nuclear stain and incubate for 5–10 min. Image directly.
    (b) The alternative is to remove the chamber and seal by adding a drop of DAPI reconstituted mounting medium (antifade gold), adding a cover slide, and sealing with nail polish. The slide can be kept at 4 °C for several days before imaging.
12. Image on fluorescence microscope with appropriate excitation/emission setup for fluorochromes used.

4 Notes

1. To passage HEK293T cells, aspirate the media carefully, and rinse the cells with 4 mL DPBS. Add 3 mL of trypsin and incubate at 37 °C for 4–5 min until the cells look loose under the microscope. Add 6 mL of media (10% FBS DMEM) to stop the digestion. Collect all the solution into a tube and centrifuge at 510*g* RT for 5 min. Discard the supernatant, and resuspend the cells in a new plate at the appropriate concentration. To maintain the cells, use a 1 in 10 passage.
2. Do not use Hoxb8 cells over 10 passages. After 5 passages, Hoxb8 cells are suboptimal.
3. Large number formation of grape-like cell clusters is a good sign of robust Hoxb8 cell viability and proliferation. However, cluster formation complicates cell counting if using an automated cell counter. For reliable counts a hemocytometer is recommended.
4. Choose 2–3 gRNAs with high specificity, intermediate-high efficiency, and a maximum of 2/3 mismatches. Exons near the N-terminus are preferentially targeted to increase the likelihood that non-functional protein is produced.
5. While the plasmid is being digested, start annealing the oligonucleotides (**step 5**).
6. Gel electrophoresis to check plasmid digestion can be carried out the next day. Keep the digestion mix at 4 °C. However, checking for plasmid digestion before **step 3** is recommended to reduce the waste of resources.

7. Prepare a mastermix with the common components for all samples.
8. Optimize to incubate several microcentrifuge tubes shaking without contamination.
9. The transformation mix can be spread with an L-shape spreader or by gently swirling 6–10 glass balls around the plate, which are removed before incubating.
10. The transformation mix can be stored up to 48 h at 4 °C.
11. The single bacterial colonies can also be added to a 12-well plate (4 mL of LB broth/well) or a 6-well plate (6 mL of LB broth/well). Only 2 mL is needed for the purification. However, attention must be paid to evaporation when using small LB broth volumes.
12. Pick a minimum of 2/3 colonies per plate, to maximize the chance of picking the clone with the successful integration of the gRNA into the plasmid.
13. DNA concentration should be at least 100 ng/μL to proceed with the protocol.
14. If using a 12-well plate, take 1 mL of media out before adding the 2 mL of fresh media.
15. Aspirate the media from the top of the well so that cells remain undisturbed at the bottom.
16. Cells will take approximately 7–11 days to reach ten million cells.
17. GFP expression can be quickly checked by flow cytometry before sorting the cells.
18. Often cells need to be resorted after expansion. Check GFP expression by flow cytometry.
19. By day 5 of differentiation, the total cell number should be the seeded cells x10, and a 40% survival rate is expected.
20. Supernatants are stored at −80 °C for long-term storage.
21. Samples can be stored at −20 °C for short-term storage.
22. The primary antibody dilution can be stored at 4 °C and reused.
23. If multiple protein detections of similar size are required on the same membrane, antibodies are stripped from the membrane using ReBlot Plus Strong Antibody Stripping Solution as per the manufacturer's instructions.
24. After the cells have been fixed, add all solutions gently to the borders of the chamber to keep cells from detaching.
25. Primary antibodies dilution can be kept and reused for up to 6 months.

Acknowledgments

We thank Profs Hans Haecker (University of Utah, USA) and Barbara Walzog (Ludwig Maximilian University of Munich, Germany) for sharing HoxB8 cells and culture protocols with us. This work was supported by the Research into Inflammatory Arthritis Centre Versus Arthritis UK, based in the Universities of Oxford, Glasgow, Birmingham, and Newcastle (22072) and the Kennedy Trust Prize studentship (PhD studentships to J.S.); Oxford-BMS program (fellowship to L.W.); the Chinese Science Council (PhD studentship to Z.A.); and the Wellcome Trust (Investigator Award 209422/Z/17/Z to I.A.U., EvG).

References

1. Beyrau M, Bodkin JV, Nourshargh S (2012) Neutrophil heterogeneity in health and disease: a revitalized avenue in inflammation and immunity. Open Biol 2(11):120134. https://doi.org/10.1098/rsob.120134
2. Khoyratty TE, Ai Z, Ballesteros I, Eames HL, Mathie S, Martín-Salamanca S, Wang L, Hemmings A, Willemsen N, von Werz V, Zehrer A, Walzog B, van Grinsven E, Hidalgo A, Udalova IA (2021) Distinct transcription factor networks control neutrophil-driven inflammation. Nat Immunol 22(9): 1093–1106. https://doi.org/10.1038/s41590-021-00968-4
3. Ballesteros I, Rubio-Ponce A, Genua M, Lusito E, Kwok I, Fernández-Calvo G, Khoyratty TE, van Grinsven E, González-Hernández S, Nicolás-Ávila J, Vicanolo T, Maccataio A, Benguría A, Li JL, Adrover JM, Aroca-Crevillen A, Quintana JA, Martín-Salamanca S, Mayo F, Ascher S, Barbiera G, Soehnlein O, Gunzer M, Ginhoux F, Sánchez-Cabo F, Nistal-Villán E, Schulz C, Dopazo A, Reinhardt C, Udalova IA, Ng LG, Ostuni R, Hidalgo A (2020) Co-option of neutrophil fates by tissue environments. Cell 183(5): 1282–1297.e1218. https://doi.org/10.1016/j.cell.2020.10.003
4. Xie X, Shi Q, Wu P, Zhang X, Kambara H, Su J, Yu H, Park SY, Guo R, Ren Q, Zhang S, Xu Y, Silberstein LE, Cheng T, Ma F, Li C, Luo HR (2020) Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. Nat Immunol 21(9):1119–1133. https://doi.org/10.1038/s41590-020-0736-z
5. Wang GG, Calvo KR, Pasillas MP, Sykes DB, Häcker H, Kamps MP (2006) Quantitative production of macrophages or neutrophils ex vivo using conditional Hoxb8. Nat Methods 3(4): 287–293. https://doi.org/10.1038/nmeth865

Chapter 7

Detection of *TP53* Mutation in Acute Myeloid Leukemia by RT-PCR-Based Sanger Sequencing

Emily R. Novak, Anagha Deshpande, Darren Finlay, James R. Mason, Aniruddha J. Deshpande, Peter D. Adams, and Sha Li

Abstract

The *TP53* gene is known to be one of the most frequently mutated genes in various human cancers. In de novo acute myeloid leukemia (AML), *TP53* has been found to be mutated in ~10% of patients. Although the frequency of *TP53* mutations in AML is substantially lower compared to other human cancers, *TP53* mutations in AML are associated with poor response to chemotherapy and poor outcomes. Therefore, assessment of *TP53* status is critical in clinical routines and research studies. In this chapter, we described the use of conventional RT-PCR for rapid detection of *TP53* mutations by Sanger sequencing. We use AML cells as an example but provide sufficient details for usage in other cell types.

Key words *TP53*, Mutation, RT-PCR, AML, Sanger sequencing

1 Introduction

Mutations in the *TP53* gene are one of the most commonly acquired mutations in cancers. The p53 protein, encoded by the *TP53* gene, performs multifaceted functions in apoptosis, DNA damage responses, autophagy, and cellular metabolism [1]. In hematologic malignancies, such as acute myeloid leukemia (AML), *TP53* is infrequently mutated in ~10% among newly diagnosed cases [2]. However, alterations or loss of *TP53* has been associated with a poor prognosis in AML [3, 4]. Consequently, analysis of *TP53* mutations has been incorporated into routine clinical practices to facilitate effective therapeutic decision-making.

Here, we performed RT-PCRs to amplify the full length of the p53 CDS in 13 samples, including 3 primary human patients' samples, 2 patient-derived xenograft (PDX) samples, as well as 8 well-characterized AML cell lines as positive controls for the methodology. All of these CDS amplicons were sequenced directly by Sanger sequencing. In the case of three primary human patients'

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_7,

samples and two patient-derived xenograft (PDX) samples in which no mutations had so far been identified, cell viability assay in the presence of MDM2 (negative regulator of p53) inhibitor HDM201 was carried out to ascertain whether they had functional wild-type p53 [5]. Given the low costs in sequencing and the short sample handling time (around 1.5 h) described in this chapter, make it practical for determining *TP53* status in different cell types in regular research labs, followed by functional validation.

2 Materials

2.1 Cell Culture

1. Equipment and supplies: T25 non-treated flasks, 6-well non--treated plates, 384-well non-treated plates, 15 mL sterile conical tubes, sterile tips, pipettes, microscope, hemacytometer, benchtop centrifuge, laboratory CO_2 water-jacketed incubators, and biosafety cabinet.
2. AML cell lines (OCI-AML3, MOLM13, MV411, Kasumi-1, KG-1a, THP1, HL-60, and U937) were obtained from ATCC (Manassas, VA, USA) and DSMZ (German Collection of Microorganisms and Cell Cultures). Cells were cultured in RPMI-1640 medium with 10% fetal bovine serum (FBS) and 1× Pen/Strep.
3. Primary AML samples from patients (named Patient #1, #2, and #3 in this study) were obtained from Carol Burian and Dr. James Mason (Scripps MD Anderson Center, La Jolla, CA) under approved Institutional Review Board protocol 13-6180. Peripheral blood mononuclear cells were isolated by Ficoll–Paque PLUS (GE Healthcare) centrifugation according to the manufacturer's instructions, and red blood cells were lysed using RBC lysis buffer (Alfa Aesar). Final peripheral blood mononuclear cells were resuspended in Bambanker serum-free freezing medium (Wako Pure Chemical Industries, Ltd) and stored frozen.
4. Patient-derived xenograft (PDX) samples (named PDX #1 and #2 in this study) were obtained from the Jeremias Lab (Munich, Germany) and were cultured in IMDM medium with 20% BIT (Stemcell Technologies), human cytokines (SCF, IL-3, IL-6, and GM-CSF), and StemRegenin 1 (SR1) and UM171, as described [6].
5. Phosphate-buffered saline.

2.2 RNA Isolation

1. Equipment and supplies: 1.5 mL sterile tubes, filter pipette tips, benchtop centrifuge, and NanoDrop spectrophotometer.
2. TRIzol Reagent (Thermo Fisher Scientific).
3. Direct-zol RNA Miniprep Plus kit (Zymo Research).

2.3 cDNA Synthesis

1. Equipment and supplies: 0.2 mL sterile PCR tubes, filter pipette tips, PCR machine, benchtop mini centrifuge, and ice bucket.
2. cDNA synthesis reaction reagents: dNTPs (10 mM), oligo (dT) primer (100 μM), 5× Reaction Buffer, RiboLock RNase Inhibitor (40 U/μL) (Thermo Fisher Scientific), RevertAid Reverse Transcriptase (200 U/μL) (Thermo Fisher Scientific), and nuclease-free H_2O.

2.4 TP53 *and* GAPDH *PCR*

1. Equipment and supplies: 0.2 mL PCR tubes, sterile pipette tips, PCR machine, and benchtop mini centrifuge.
2. PCR reaction reagents: dNTPs (10 mM), Phusion High-Fidelity DNA polymerase (NEB), 5× Phusion HF Buffer, and primer sets for *TP53* and *GAPDH*.

2.5 Electrophoresis

1. Equipment and supplies: Agarose gel electrophoresis tank and power supply and agarose gel imager.
2. Electrophoresis reagents: Agarose, GeneRuler 1 kb and 100 bp ladders, loading buffer, and TAE.

2.6 TP53 *PCR Product Purification*

1. Equipment and supplies: 1.5 mL sterile tubes, sterile pipette tips, DynaMag-2 magnet (Invitrogen), PCR machine, and benchtop mini centrifuge.
2. AMPure XP beads (Beckman Coulter).
3. 70% ethanol.

2.7 Cell Viability Assays

1. Equipment and supplies: Sterile tubes, sterile pipette tips, Echo 555 Liquid Handler, hemacytometer, and PerkinElmer Envision Microplate Reader.
2. CellTiter-Glo (Promega).
3. HDM201 (Selleckchem) and DMSO.

2.8 Software and Algorithms

1. SnapGene (Version 6.0.2): https://www.snapgene.com.
2. GraphPad Prism (Version 9.3.1): https://www.graphpad.com.

3 Methods

3.1 Design of TP53 PCR Primers and Sequencing Primers

1. Copy human *TP53* mRNA (NM_001126112.3) sequence from NCBI website (https://www.ncbi.nlm.nih.gov/nuccore/NM_001126112.3), and paste into SnapGene for the visualization and annotation.
2. Design full-length *TP53* CDS PCR primers F1 (forward primer) and R1 (reverse primer) (*see* Fig. 1) manually on the SnapGene. The PCR product covers the entire human *TP53* CDS region (exons 2–11). The expected size of the PCR product is 1359 bp.

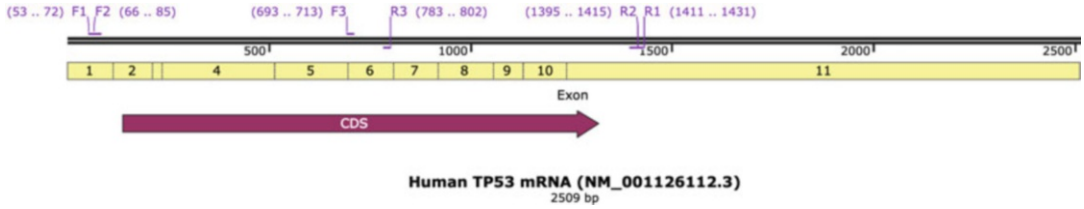


Fig. 1 Human *TP53* mRNA (NM_001126112.3) structure. Primers used in this study are indicated

Table 1
Primers

TP53 CDS PCR primer
F1: 5′- GACACTTTGCGTTCGGGCTG-3′
R1: 5′- CTGACGCACACCTATTGCAAG-3′
TP53 sequencing primer
F2: 5′- CGGGCTGGGAGCGTGCTTTC-3′
F3: 5′- GCGATGGTCTGGCCCCTCCTC-3′
R2: 5′- GCAAGCAAGGGTTCAAAGACC-3′
R3: 5′- CTCATAGGGCACCACCACAC-3′

3. Design *TP53* sequencing primers F2, F3, R2, and R3 (*see* Fig. 1) in both sense and antisense directions to achieve accurate and full coverage (*see* **Note 1**). F1 and R1 can also be used as sequencing primers to replace F2 and R2, respectively.
4. Order the designed oligos (Table 1) from IDT (https://www.idtdna.com/pages). Dissolve and dilute the oligos in ddH_2O and to a final concentration of 100 μM.

3.2 Design of GAPDH PCR Primers

As a positive control for RNA quality and PCR procedure, we included *GAPDH* gene. *GAPDH* (glyceraldehyde-3-phosphate dehydrogenase) (NM_001289746.2) is one of the most commonly used housekeeping genes used in comparisons of gene expression data. The following primers were used for a small amplicon of human *GAPDH*: F: 5′- GAGAGACCCTCACTGCTG-3′ and R: 5′- GATGGTACATGACAAGGTGC-3′. The expected size of the PCR product is 135 bp.

3.3 RNA Isolation

1. Pellet the cells (5×10^6 cells) by centrifugation ($400 \times g$ for 3 min), and discard the supernatant. To reduce RNA degradation, proceed with RNA isolation immediately, or quick-freeze samples immediately, and store at −80 °C or in liquid nitrogen until RNA isolation.

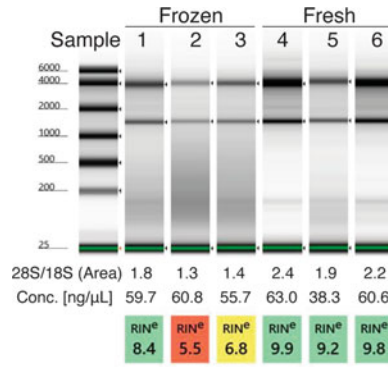


Fig. 2 Six samples (samples #1–3 are frozen samples and #4–6 are freshly harvested samples) used in this study with different levels of RNA degradation were analyzed using the Agilent 4200 TapeStation Instrument and the Agilent RNA ScreenTape assay. The determined RIN values, 28S/18S (Area), and concentration [ng/μL] are shown under the gel image

2. Add 600 μL of TRIzol Reagent (Thermo Fisher Scientific) to the pellet. Pipet the lysate up and down several times to homogenize. Samples in TRIzol can be stored at 4 °C overnight or at −20 °C for up to a year.
3. Total RNA was isolated using Direct-zol RNA Miniprep Plus kit (Zymo Research). The extraction procedure is executed exactly as described by the manufacturer's manual. Though the on-column DNase I treatment is optional, we do recommend including it.
4. Determine the purity and concentration of RNA using a NanoDrop spectrophotometer. A A260/A280 ratio of ~2 is considered pure.
5. (Optional) Determine the RNA integrity and purity using Agilent TapeStation (see **Note 2**). Six samples in this study were analyzed, and an image produced by TapeStation is shown in Fig. 2.

3.4 cDNA Production

Thaw and keep all reagents on ice.

1. Add the following reagents (Table 2) into a 200 μL sterile, nuclease-free PCR tube in the indicated order. A volume of 500–1000 ng of total RNA for each sample was reverse transcribed to cDNA.
2. Denature RNA for 5 min at 65 °C and hold at 4 °C indefinitely in the thermal cycler. Centrifuge briefly and put back on ice.
3. Add the following components (Table 3) into the same tube in the indicated order.

Table 2
Reagents

Components	Volume
Total RNA (500–1000 ng)	Variable
Oligo(dT) primer (100 μM)	1.5 μL
Nuclease-free H_2O	To 12.5 μL
Total volume	14 μL

Table 3
Reagents

Components	Volume
5× reaction buffer	4 μL
RiboLock RNase inhibitor (40 U/μL)	0.25 μL
dNTPs (10 mM)	1.5 μL
RevertAid reverse transcriptase (200 U/μL)	0.25 μL
Total volume	6 μL

4. Incubate the 20 μL cDNA synthesis reaction in the thermal cycler. Set up the following steps: 10 min at 25 °C, 60 min at 42 °C, 10 min at 72 °C, and hold indefinitely at 4 °C. The reaction product can be directly used in PCR application described in Subheadings 3.5 and 3.6 or stored at −20 °C.

3.5 *Human* TP53 *PCR*

Thaw and keep all reagents on ice.

1. Add the following components (Table 4) into a 200 μL sterile, nuclease-free PCR tube in the indicated order.
2. Perform PCR in a thermal cycler with a heated lid (Table 5).
3. Load 5 μL of the PCR product on 2% agarose gel. A single distinct *TP53* band (1359 bp) is observed after ethidium bromide staining. The image is shown in Fig. 3.

3.6 *Human* GAPDH *PCR*

1. Prepare the same PCR reaction as shown in Subheading 3.5, **step 1**, except replacing *TP53* forward and reverse primers with *GAPDH* forward and reverse primers, respectively (indicated by "a" in Table 4).
2. Modify the thermal cycler program shown in Subheading 3.5, **step 2**, with these changes (also indicated by "a" in Table 5): (1) annealing at 61 °C; (2) extension for 10 s.
3. Load 5 μL of the PCR product on 2% agarose gel. A single distinct *GAPDH* band (135 bp) is observed after ethidium bromide staining. The image is shown in Fig. 3.

Table 4
Reagents

Components	Volume
cDNA	2 μL (*see* **Note 3**)
5× phusion HF PCR buffer	5 μL
dNTP mix (10 mM)	1 μL
TP53 forward primer F1 (10 μM)[a]	2.5 μL
TP53 reverse primer R1 (10 μM)[a]	2.5 μL
Phusion DNA polymerase (2 U/μL)	0.5 μL
Nuclease-free H_2O	36.5 μL
Total volume	50 μL

[a]Replace with *GAPDH* primers for *GAPDH* PCR

Table 5
PCR steps

Steps	Temperature	Time	Cycles
Initial denaturation	98 °C	30 s	1
Denaturation Annealing Extension	98 °C 64°C[a] 72 °C	10 s 30 s 45 s[a]	30
Final extension	72 °C	10 min	1
Hold	4 °C	∞	

[a]Modify for *GAPDH* PCR

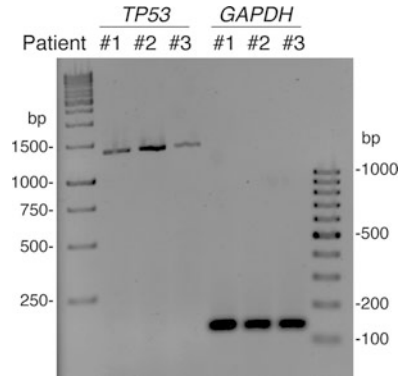


Fig. 3 *TP53* and *GAPDH* PCR result from three primary human patients' samples. The expected size of *TP53 and GAPDH* is 1359 bp and 135 bp, respectively

3.7 Human TP53 PCR Purification

1. *TP53* PCR products were cleaned using AMPure beads (Beckman Coulter) following the manufacturer's instruction.
2. Determine the purity and concentration of *TP53* amplicon using a NanoDrop spectrophotometer.
3. Mix 40 ng (in 10 μL) of the purified PCR product with 5 μM of sequencing primer (in 5 μL) together for one Sanger sequencing. We used commercial sequencing service provider such as Genewiz (https://www.genewiz.com).

3.8 Sanger Sequencing Result Analysis

1. Download trace file (.ab1) of the sequencing result from sequencing provider portal. Import and align Sanger Traces to human *TP53* mRNA sequence (NM_001126112.3) using SnapGene.
2. Carefully view the sequence and trace (chromatogram) to identify mutations and polymorphisms (e.g., codon 72 polymorphism).

3.9 Cell Viability Assays

1. Harvest cells at 1×10^5/mL with cell numbers determined by trypan blue using hemacytometer. Seed 2500 cells (in 25 μL) in a 384-well microplate spotted with HDM201 (MDM2 inhibitor) or vehicle control (DMSO) using an Echo 555 Liquid Handler. Incubate cells for 2 days.
2. Add 10 μL of CellTiter-Glo reagent to each well, and centrifuge at 1000 x g for 5 min to remove any bubbles. Luminescence signals were detected on a PerkinElmer Envision Microplate Reader.
3. Data were analyzed using GraphPad Prism software (*see* Fig. 4). HDM201 promotes antiproliferative activity selectively in cells with wild-type *TP53*.

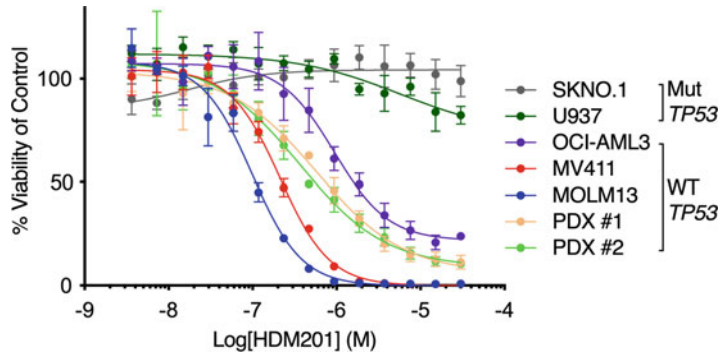


Fig. 4 Viability of indicated cells in the presence of MDM2 inhibitor HDM201 at various concentrations (14-point dose response). The percentages compared with the DMSO vehicle control were curve fitted using nonlinear regression (log [inhibitor] versus normalized response, variable slope) and represent the mean ± SD ($n = 4$). Cells with wild-type (WT) or mutant (Mut) *TP53* are indicated

4 Notes

1. We typically get DNA sequencing read lengths up to ~1000 bases with a reliable length of 700–800 bases for high-quality template. Since sequencing quality may vary, extra sequencing primers can be designed to have full coverage.
2. One of the drawbacks of NanoDrop spectrophotometer is the inability to assess RNA integrity. RNA integrity can be evaluated using an Agilent TapeStation Instrument or similar. The instrument software assigns an RNA Integrity Number (RIN) at a scale from 1 to 10 (with 10 indicating highly intact RNA and 1 indicating strongly degraded RNA) [7]. RNA integrity is greatly influenced by how well the sample was preserved. We found RIN and 28S/18S values from frozen samples are lower (indicating degraded RNA) compared to freshly harvested samples (*see* Fig. 2). However, the sample quality is still sufficient for RT-PCR with comprised cDNA yields. A less costly but time-consuming alternative for assessment of RNA integrity would be by electrophoresis on denaturing agarose gels.
3. If obtaining low yield or no RT-PCR product, the volume of cDNA synthesis reaction mixture should be increased (i.e., 5 μL in Table 4). RNA integrity should also be assessed.

Acknowledgments

Primary AML samples from patients were provided by SBP's Tumor Analysis Shared Resources through a collaboration with Carol Burian and Dr. James Mason of Scripps MD Anderson (La Jolla, CA). Patient-derived xenograft (PDX) samples were obtained from the Jeremias Lab (Munich, Germany). We thank Hillarie Austin and Kang Liu at SBP Core Facilities for their technical assistance. SBP's Core Facilities are supported by NCI Cancer Center support grant P30CA030199.

References

1. Kastenhuber ER, Lowe SW (2017) Putting p53 in context. Cell 170(6):1062–1078
2. Cancer Genome Atlas Research Network, et al. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 368(22): 2059–2074
3. Papaemmanuil E et al (2016) Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med 374(23):2209–2221
4. Seifert H et al (2009) The prognostic impact of 17p (p53) deletion in 2272 adults with acute myeloid leukemia. Leukemia 23(4):656–663
5. Jeay S et al (2018) Dose and schedule determine distinct molecular mechanisms underlying the efficacy of the p53-MDM2 inhibitor HDM201. Cancer Res 78(21):6257–6267
6. Pabst C et al (2014) Identification of small molecules that support human leukemia stem cell activity ex vivo. Nat Methods 11(4): 436–442
7. Schroeder A et al (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3

Chapter 8

Assessing the Activity of Transcription Factor FoxO1

Limin Shi, Zhipeng Tao, and Zhiyong Cheng

Abstract

The transcription factor FoxO1 (forkhead box O1) regulates genes that are involved in development, metabolism, cellular innovation, longevity, and stress responses. Assessment of FoxO1 activity is therefore critical to understand the regulatory network of this transcription factor. FoxO1 transactivation activity relies on its ability to bind to the promoters of target genes, which is controlled by posttranslational modifications (e.g., dephosphorylation or phosphorylation) that may promote nuclear translocation or exclusion of FoxO1. In this chapter we describe the protocols for FoxO1 activity assessment using Western blotting analysis of the posttranslational modification of FoxO1 in whole cell lysates and ELISA of DNA binding activity of FoxO1 in nuclear extracts.

Key words Transcription factor, FoxO1, Posttranslational modification, Phosphorylation, DNA binding activity

1 Introduction

The transcription factor FoxO1 (or orthologs) regulates an array of genes that are involved in development, metabolism, cellular innovation, longevity, and stress responses across species [1–3]. To transactivate gene expression, FoxO1 translocates into the nucleus and binds to the promoters of target genes [1, 4, 5]. Posttranslational modifications (e.g., phosphorylation, acetylation, methylation, and ubiquitination) of FoxO1 protein play a central role in FoxO1 translocation between the nucleus and cytoplasm [1]. For instance, insulin elicits the signal cascade through insulin receptor, activating insulin receptor substrates and protein kinase B (PKB/Akt). Akt in turn induces phosphorylation at Thr24, Ser256, and Ser319 of mouse Foxo1 (at Thr32, Ser253, and Ser315 of human FOXO1), which promotes FoxO1 translocation from the nucleus to the cytoplasm and suppresses its transactivation

The authors "Limin Shi" and "Zhipeng Tao" are equally contributed to this chapter.

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_8,

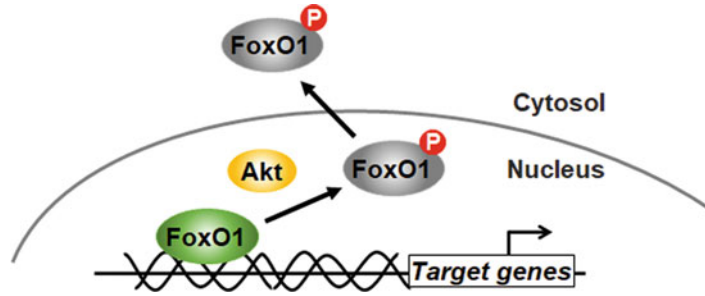


Fig. 1 The schematic view of FoxO1 activity regulated by its DNA (promoter)-binding ability and nuclear translocation or exclusion. Posttranslational modifications control FoxO1 subcellular location and ability to bind to the promoters of target genes. Akt-mediated phosphorylation is known to exclude FoxO1 from the nucleus

activity (*see* Fig. 1) [1, 4]. FoxO1 may be polyubiquitinated after nuclear exclusion and undergo proteasomal degradation [6, 7].

Assessment of FoxO1 activity can be based on its subcellular locations, stability, and transcriptional activity. Determination of the ratios of phosphorylated FoxO1 versus total FoxO1 protein in whole lysates by Western blotting reflects the status of both FoxO1 deactivation, expression, and stability (or degradation) [8]. Alternatively, co-staining of FoxO1 and nucleus using immunohistochemistry or immunocytochemistry can detect the subcellular locations [9]. To detect whether FoxO1 is active and able to bind to a specific gene, gel shift assay (or electrophoretic mobility shift assay, EMSA) can be used after incubating nuclear extracts with radioactive probes (the target sequences) [10, 11]. In addition, luciferase reporter assays have been developed to monitor FoxO1 binding to the specific promoters and transactivation activity [12, 13]. Using FoxO1 elements (i.e., oligonucleotides containing FoxO1 binding motifs) to precoat the plate, it makes ELISA possible to detect active FoxO1 in nuclear extraction with no need of radioactive labeling (for EMSA) or transfection (for luciferase reporter assays) procedures [14–16]. In this chapter we describe the protocols assessing FoxO1 activity in 3 T3 L1 cells using two most accessible methods, Western blotting analysis and ELISA.

2 Materials

2.1 Cell Culture

1. Equipment and supplies: biosafety cabinet, laboratory water-jacketed CO_2 incubator, light microscope, benchtop centrifuge, 10 cm tissue culture dishes, 6-well tissue culture plates, 50 mL sterile conical tubes, sterile tips, and pipettes.
2. Phosphate-buffered saline (PBS).

3. 3 T3 L1 cell line (CL-173) purchased from ATCC (Manassas, VA, USA).
4. Basal medium: Dulbecco's modified Eagle's (DMEM) medium, 10% fetal bovine serum (FBS), and 1× Pen/Strep.

2.2 Adipocyte Differentiation

1. Equipment and supplies: biosafety cabinet, laboratory water-jacketed CO_2 incubator, inverted microscope, 6-well tissue culture plates, sterile tips, pipettors, and pipettes.
2. Differentiation media (DM): DMI – DMEM supplemented with 10% FBS, P/S (1×), IBMX (0.5 mM), dexamethasone (1 μM), insulin (1 μg/mL), and rosiglitazone (2 μM). DMII – DMEM supplemented with 10% FBS, P/S (1×), and insulin (1 μg/mL).
3. Maintenance medium: DMEM media containing 10% FBS, 1× Pen/Strep, and insulin (0.5 μg/mL).

2.3 Western Blotting

1. Equipment and supplies: ChemiDoc Imaging Systems (Bio-Rad), benchtop centrifuge, microplate reader, Bullet Blender® (Next Advance, Inc.), heat block, cell lifters, 50 mL sterile conical tubes, sterile tips, 2 mL sterile tubes, pipettes.
2. Phosphate-buffered saline (PBS).
3. DC protein assay kit (Bio-Rad, Cat No.: 5000111).
4. Bovine serum albumin (BSA) standard.
5. PLC lysis buffer (30 mM HEPES, pH 7.5, 150 mM NaCl, 10% glycerol, 1% Triton X-100, 1.5 mM $MgCl_2$, 1 mM EGTA, 10 mM NaPPi, 100 mM NaF, 1 mM Na_3VO_4) supplemented with protease inhibitor cocktail and 1 mM PMSF (freshly added right before use).
6. Loading buffer (5×): 0.25 M Tris–HCl (pH 6.8), 25% glycerol, 5% SDS, 0.25% bromophenol blue, 0.5 M DTT.
7. Running buffer (1×): 25 mM Tris-Base, 192 mM glycine, 0.1% SDS (pH 8.3).
8. Transfer buffer: 25 mM Tris-Base, 192 mM glycine (pH 8.3), 15–20% methanol.
9. Washing buffer: 50 mM Tris–HCl, 150 mM NaCl, 0.1% Tween 20 (pH 7.4).
10. ECL substrates.
11. Antibodies: FoxO1 (L27) antibody (Cell Signaling Technology, Cat No: 9454); anti-phospho-FOXO1 (pSer256) antibody (Millipore Sigma, Cat No: SAB4300094); GAPDH antibody (ThermoFisher Scientific, Cat No: MA5–15738).

2.4 ELISA

1. Equipment and supplies: microplate reader, benchtop centrifuge, rocking platform, pipettes and pipets, and cell lifters.
2. Phosphate-buffered saline (PBS).
3. Hypotonic buffer (1×): 20 mM HEPES, 5 mM NaF, 10 uM Na_2MoO_4, 0.1 mM EDTA, pH 7.5.
4. Nuclear Extract Kit (Active Motif, Cat No.:40010); TransAM® FKHR (FOXO1) DNA-binding ELISA kit (Active Motif, Cat No.: 46396).
5. Antibodies: α-Tubulin (11H10) antibody (Cell Signaling Technology, Cat No: 2125) and Lamin A/C (4C11) antibody (Cell Signaling Technology, Cat No: 4777).

3 Methods

3.1 Culture and Differentiation of 3 T3 L1 Cells

1. Seed 3 T3 L1 cells in 10 cm dishes, and then subculture in 6-well plates with basal media.
2. Change the media every 2 days until the cells reach confluence (day 0).
3. Change the media one more time, and maintain the cells in basal media till day 2.
4. At the end of day 2, replace basal media with differentiation medium I (DMI).
5. At the end of day 4, replace DMI with differentiation medium II (DMII).
6. At the end of day 6, replace DMII medium with maintenance medium.
7. Change maintenance medium every 2 days until day 12 when the cells were fully differentiated, which was confirmed by oil red O staining (*see* Fig. 2a) [17–19].

3.2 Western Blotting Analysis of FoxO1 Activity

1. At the indicated time points (days 0, 6, and 12) during cell differentiation (*see* Fig. 2a), remove the media, and wash the cells with cold PBS twice.
2. Harvest the cells in 2 mL cold PBS using cell lifters.
3. Centrifuge at 5000 × g, 4 °C for 5 min to pellet the cells, and discard the supernatant.
4. Lyse the cells with a Bullet Blender according to the manufacturer's instruction.
5. Centrifuge at 12,000 × g, 4 °C for 10 min.
6. Transfer the cell lysate to a clean centrifuge tube.
7. Measure protein concentration for each cell lysate on a microplate reader using a DC protein assay kit and BSA as the protein standard for calibration curve.

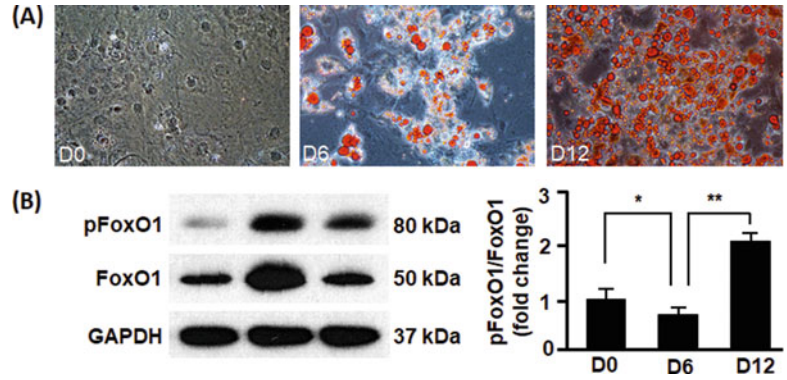


Fig. 2 Measurement of FoxO1 activity during adipocyte differentiation by Western blotting analysis of total and phosphorylated FoxO1 in whole cell lysates. (**a**) Differentiation of 3 T3-L1 cells was induced, and the cells were stained by oil red O as described in Subheading 3.1, which showed a gradual increase in lipid droplet formation and accumulation from day 0 to days 6 and 12. Magnification: 200×. (**b**) Western blotting analysis of total and phosphorylated FoxO1 (at Ser356) in whole cell lysates as described in Subheading 3.2, which suggested a downregulation of inhibitory phosphorylation (pFoxO1/FoxO1) from day 0 to day 6 and then an upregulation of inhibitory phosphorylation from day 6 to day 12

8. Make 0.3–0.5 mL aliquots of cell lysates and store them at −80 °C for later use.
9. Take one aliquot of each lysate, and dilute to a final concentration of 1–2 mg/mL with 5× loading buffer (and water when necessary; *see* **Note 1**).
10. Incubate the samples on heat block (100 °C) for 10 min, and then let sit at room temperature.
11. When the samples cool down to room temperature, briefly centrifuge (1000 × g, 20 sec).
12. Perform Western blotting analysis, using 7.5% gel (*see* **Note 2**) for SDS-PAGE (60 V for 30 min, following by 110 V for 2 h) and wet transfer technique (110 V for 1 h).
13. Use Pierce ECL Western blotting substrate and ChemiDoc Imaging Systems for Western blotting detection of total and phosphorylated (at Ser256) FoxO1 proteins on day 0, day 6, and day 12 (*see* Fig. 2b, left).
14. The band densities were quantified to calculate the ratio of phosphorylated FoxO1 versus total FoxO1 proteins, and fold changes were calculated by taking D0 value as 1 (*see* Fig. 2b, right; *see* **Note 3**).

3.3 ELISA of FoxO1 Activity

1. At the planned time points (e.g., days 0, 6, and 12 during cell differentiation), remove the media and wash the cells with cold PBS twice.
2. Harvest the cells in 2 mL cold PBS using cell lifters.
3. Centrifuge at 5000 × g, 4 °C for 5 min to pellet the cells, and discard the supernatant.
4. Prepare 1× hypotonic buffer and complete lysis buffer according to the manufacturer (Active Motif)'s instruction.
5. Resuspend cells in 500 μL hypotonic buffer (1×) and let sit on ice for 15 min.
6. Add 25 μL detergent and vortex vigorously for 10 sec (*see* **Note 4**).
7. Centrifuge at 14000 × g at 4 °C for 10 min.
8. Transfer supernatant into a pre-chilled 1.7 mL microcentrifuge tube; label it as cytoplasmic fraction (CF; *see* **Note 5**).
9. The pellet is suspended in 50 μL complete lysis buffer by pipetting up and down, let sit on ice for 30 min, and tap the tube every 5 min.
10. Vortex vigorously for 30 sec.
11. Centrifuge at 14,000 at 4 °C for 10 min.
12. Transfer the supernatant into a pre-chilled 1.7 mL microcentrifuge tube, make aliquots, and label them as nuclear fraction (NF; *see* **Note 6**).
13. Measure protein concentrations in NF from **step 12** and CF from **step 8** using a DC protein assay kit (Bio-Rad).
14. Determine the fractionation efficiency by Western blotting analysis of marker proteins in CF and NF (*see* Fig. 3a and **Note 7**).
15. Prepare complete binding buffer, 1× washing buffer, and 1× antibody binding buffer according to the manufacturer (Active Motif)'s instruction.
16. Dilute NF with complete lysis buffer (Active Motif) to make a final protein concentration of 0.5 mg/mL for the following ELISA.
17. Add 40 μL complete binding buffer to each well of the ELISA strips.
18. Add 10 μL sample (5 μg protein; *see* **Note 8**) to each well, or 10 μL of nuclear extract (5 μg, diluted in 10 μL complete lysis buffer) provided by Active Motif as the positive control (PC), or 10 μL complete lysis buffer as the negative control (NC).
19. Seal the plate with an adhesive cover, and incubate on a rocking platform (100 rpm) at room temperature for 1 h.

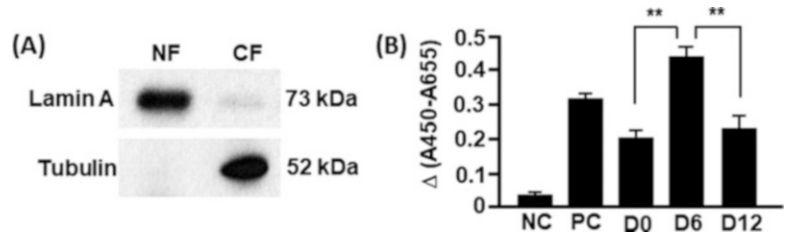


Fig. 3 Measurement of FoxO1 activity during adipocyte differentiation by ELISA. (**a**) Nuclear extraction and verification of fractionation efficiency. Nuclear fraction (NF) and cytoplasmic fraction (CF) were prepared as described in Subheading 3.3, and the fractionation efficiency was tested by probing nuclear protein (lamin **a**) and cytoplasmic protein (tubulin) in NF and CF using Western blotting analysis. (**b**) ELISA analysis of DNA-binding ability of FoxO1 in nuclear fractions isolated from 3 T3-L1 cells at days 0, 6, and 12 during differentiation. Complete lysate buffer and FoxO1 element (containing FoxO1 binding motif) were used as the negative control (NC) and positive control (PC), respectively. The results suggested an upregulation of FoxO1's DNA-binding activity from day 0 to day 6 and then a downregulation of FoxO1's DNA-binding activity from day 6 to day 12

20. Wash each well 3 times with 200 μL 1× washing buffer. To remove residual washing buffer from the wells after each wash, flick the plate over a sink to empty the wells, and then tap the inverted plate on absorbent paper towels 5–7 times.
21. Dilute FKHR (FoxO1) antibody at 1:1000 dilution with 1× antibody binding buffer, and add 100 μL diluted FKHR (FoxO1) antibody to each well (*see* **Note 9**).
22. Seal the plate and incubate at room temperature for 1 h. without agitation.
23. Wash each well 3 times with 200 μL 1× washing buffer. After each wash, flick the plate over a sink to empty the wells, and then tap the inverted plate on absorbent paper towels 5–7 times.
24. Dilute HRP-conjugated antibody at 1:1000 dilution with 1× antibody binding buffer, and then add 100 μL diluted HRP-conjugated antibody to each well (*see* **Note 10**).
25. Seal the plate and incubate at room temperature for 1 h (without agitation). During the incubation, place the developing solution to equilibrate at room temperature.
26. Wash each well 4 times with 200 μL 1× washing buffer. After each wash, flick the plate over a sink to empty the wells, and then tap the inverted plate on absorbent paper towels 5–7 times.
27. Add 100 μL developing solution to each well, cover the plate with foil to protect from direct light, and incubate 5 min at room temperature (*see* **Note 11**).

28. Add 100 μL stop solution, and the blue color turns into yellow.
29. Read absorbance on a microplate reader immediately at 450 nm (A450) with a reference wavelength of 655 nm (A655), and the differences between A450 and A655 were calculated to assess DNA-binding activities of FoxO1 for the negative control (NC), positive control (PC), and samples (*see* Fig. 3b and **Note 12**).

4 Notes

1. For optimal signal detection, the amount of protein loaded in each well should be no less than 20 μg. Fat cells contain lower protein content than other cells such as hepatocytes. When the final protein concentration is lower than 1 mg/mL, the volume of loaded sample in a well should be increased accordingly to make the total amount of protein no less than 20 μg (e.g., 10 μL × 2 mg/mL, 20 μL × 1 mg/mL, and 25 μL × 0.8 mg/mL).
2. Use of 7.5% gel can achieve a better separation than 10% or 12.5% gel.
3. Downregulation of pFoxO1/FoxO1 indicates an increase in FoxO1 activity (e.g., D6), while upregulation of pFoxO1/FoxO1 indicates decrease of FoxO1 activity (e.g., D12), because phosphorylation at Ser256 promotes nuclear exclusion of FoxO1 [1, 4]. These results in Fig. 2 suggest that FoxO1 is activated and then deactivated during the cell differentiation.
4. Depending on cell types, a Dounce homogenizer may be necessary to enhance cell lysis. We found no difference in samples prepared with and without a Dounce homogenizer for 3T3L1 cells.
5. CF can be stored at −80 °C for later analysis of fractionation efficiency by Western blotting analysis of cytosolic protein (e.g., tubulin) and nuclear protein (e.g., lamin A) as described in Subheading 3.2.
6. NF can be stored at −80 °C for later analysis of fractionation efficiency by Western blotting analysis of cytosolic protein (e.g., tubulin) and nuclear protein (e.g., lamin A) as described in Subheading 3.2 and for the measurement of FoxO1's DNA binding activity.
7. An optimal or efficient fractionation will show cytosolic protein tubulin present in CF but not in NF and nuclear protein lamin A in NF but not in CF. Contamination is manifested by the presence of tubulin in NF or lamin A in CF. The Western blotting analysis (*see* Fig. 3a) verifies that the fractionation was efficient.

8. It is suggested that 2–20 μg protein be used; we found that 5 μg protein produced strong enough signal.
9. Antibody dilution can also be performed during the 1 h of **step 18** of Subheading 3.3 or right before flicking the plate over a sink to empty the wells during the last washing in **step 19** of Subheading 3.3.
10. Antibody dilution can also be performed right before flicking the plate over a sink to empty the wells during the last washing in **step 22** of Subheading 3.3.
11. It is suggested that the incubation lasts 2–10 min; we found 5 min is sufficient for chromogen development. To avoid overdevelopment, monitor the blue color development in the sample and positive control wells until it turns medium to dark blue.
12. For best practice, the microplate reader should be turned on 15 min before the measurement. The intensity of Δ(A450-A655) is directly proportional to the abundance of active FoxO1 in the nuclear extract, which is captured by the immobilized DNA molecules (FoxO1 elements) in each well. The ELISA data in Fig. 3b shows that the abundance of activated FoxO1 increases first (day 6 vs day 0) and then decreases (day 12 vs day 6), which is consistent with the results from Western blotting data (*see* Fig. 2). Together, the data verifies that FoxO1 activity undergoes up- and downregulation to accomplish adipocyte differentiation [8].

Acknowledgments

This work was supported in part by the American Heart Association Grant (18TPA34230082 to Z.C.) and the USDA National Institute of Food and Agriculture Grant (1020373 to Z.C).

References

1. Cheng Z (2019) The FoxO-autophagy axis in health and disease. Trends Endocrinol Metab 30(9):658–671. https://doi.org/10.1016/j.tem.2019.07.009
2. Cheng Z (2015) FoxO1: mute for a tuned metabolism? Trends Endocrinol Metab 26(7):402–403. https://doi.org/10.1016/j.tem.2015.06.006
3. Calissi G, Lam EW, Link W (2021) Therapeutic strategies targeting FOXO transcription factors. Nat Rev Drug Discov 20(1):21–38. https://doi.org/10.1038/s41573-020-0088-2
4. Cheng Z, White MF (2011) Targeting Forkhead box O1 from the concept to metabolic diseases: lessons from mouse models. Antioxid Redox Signal 14(4):649–661. https://doi.org/10.1089/ars.2010.3370
5. Van Der Heide LP, Hoekman MF, Smidt MP (2004) The ins and outs of FoxO shuttling: mechanisms of FoxO translocation and transcriptional regulation. Biochem J 380(Pt 2):297–309. https://doi.org/10.1042/BJ20040167
6. Aoki M, Jiang H, Vogt PK (2004) Proteasomal degradation of the FoxO1 transcriptional regulator in cells transformed by the P3k and Akt oncoproteins. Proc Natl Acad Sci U S A 101(37):13613–13617. https://doi.org/10.1073/pnas.0405454101

7. Yamagata K, Daitoku H, Takahashi Y et al (2008) Arginine methylation of FOXO transcription factors inhibits their phosphorylation by Akt. Mol Cell 32(2):221–231. https://doi.org/10.1016/j.molcel.2008.09.013
8. Zou P, Liu L, Zheng L et al (2014) Targeting FoxO1 with AS1842856 suppresses adipogenesis. Cell Cycle 13(23):3759–3767. https://doi.org/10.4161/15384101.2014.965977
9. Wang S, Xia P, Huang G et al (2016) FoxO1-mediated autophagy is required for NK cell development and innate immunity. Nat Commun 7:11023. https://doi.org/10.1038/ncomms11023
10. Park CH, Skarra DV, Rivera AJ, Arriola DJ, Thackray VG (2014) Constitutively active FOXO1 diminishes activin induction of Fshb transcription in immortalized gonadotropes. PLoS One 9(11):e113839. https://doi.org/10.1371/journal.pone.0113839
11. Brent MM, Anand R, Marmorstein R (2008) Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. Structure 16(9):1407–1416. https://doi.org/10.1016/j.str.2008.06.013
12. Langlet F, Haeusler RA, Linden D et al (2017) Selective inhibition of FOXO1 activator/repressor balance modulates hepatic glucose handling. Cell 171(4):824–835. e818. https://doi.org/10.1016/j.cell.2017.09.045
13. Zanella F, Rosado A, Garcia B, Carnero A, Link W (2009) Using multiplexed regulation of luciferase activity and GFP translocation to screen for FOXO modulators. BMC Cell Biol 10:14. https://doi.org/10.1186/1471-2121-10-14
14. Liu L, Tao Z, Zheng LD et al (2016) FoxO1 interacts with transcription factor EB and differentially regulates mitochondrial uncoupling proteins via autophagy in adipocytes. Cell Death Discovery 2:16066. https://doi.org/10.1038/cddiscovery.2016.66
15. Chatterjee S, Daenthanasanmak A, Chakraborty P et al (2018) CD38-NAD(+)Axis regulates immunotherapeutic anti-tumor T cell response. Cell Metab 27(1):85–100. e108. https://doi.org/10.1016/j.cmet.2017.10.006
16. Chakraborty P, Vaena SG, Thyagarajan K et al (2019) Pro-survival lipid Sphingosine-1-phosphate metabolically programs T cells to limit anti-tumor activity. Cell Rep 28(7):1879–1893. e1877. https://doi.org/10.1016/j.celrep.2019.07.044
17. Tao Z, Shi L, Parke J et al (2021) Sirt1 coordinates with ERalpha to regulate autophagy and adiposity. Cell Death Discovery 7(1):53. https://doi.org/10.1038/s41420-021-00438-8
18. Tao Z, Liu L, Zheng LD, Cheng Z (2019) Autophagy in adipocyte differentiation. Methods Mol Biol 1854:45–53. https://doi.org/10.1007/7651_2017_65
19. Tao Z, Zheng LD, Smith C et al (2018) Estradiol signaling mediates gender difference in visceral adiposity via autophagy. Cell Death Dis 9(3):309. https://doi.org/10.1038/s41419-018-0372-9

Chapter 9

Targeting Transcription Factors in Cancer: From "Undruggable" to "Druggable"

Zhipeng Tao and Xu Wu

Abstract

Deregulation of transcription factors is critical to hallmarks of cancer. Genetic mutations, gene fusions, amplifications or deletions, epigenetic alternations, and aberrant post-transcriptional modification of transcription factors are involved in the regulation of various stages of carcinogenesis, including cancer initiation, progression, and metastasis. Thus, targeting the dysfunctional transcription factors may lead to new cancer therapeutic strategies. However, transcription factors are conventionally considered as "undruggable." Here, we summarize the recent progresses in understanding the regulation of transcription factors in cancers and strategies to target transcription factors and co-factors for preclinical and clinical drug development, particularly focusing on c-Myc, YAP/TAZ, and β-catenin due to their significance and interplays in cancer.

Key words Transcription factor, Transcription co-factors, Cancer, Undruggable, Druggable

1 Introduction

Transcription factors are proteins that bind to specific DNA sequences and regulate gene expression [1], which are essential for almost all aspects of cellular functions. Deregulation of gene expression is associated with hallmarks of various types of cancers [2]. Transcription factors themselves are often altered in cancers through genetic mutations, gene amplifications or deletions, epigenetic alternations, and aberrant post-transcriptional modification [3], leading to deregulation of their functions and enhancing tumorigenesis. In the past decades, numerous transcription factors have been revealed as critical regulators of cancer cell proliferation, invasion, the epithelial-mesenchymal transition (EMT), and "stemness" [2–4]. These transcription factors and their co-factors include YAP/TAZ (YAP1/WWTR1) [5, 6], c-Myc [7, 8], β-catenin [9, 10], FOX (Forkhead box) proteins (FOXO1, FOXO3, FOXO4, FOXL2, FOXC1, FOXC2, FOXP3, FOXM1, FOXK2)

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_9,

[11–17], STAT3 [18], the nuclear factor kappa B (NF-κB) [19], RUNX family of transcription factors (RUNX1, RUNX2, RUNX3) [20], YY1 [21], Activator Protein-1 (AP-1) [22], p53 [23, 24], and NF-E2 p45-related factor 2 (Nrf2) [25]. Thus, targeting the deregulation of transcription factors at both transcriptional and post-transcriptional levels would lead to promising therapeutics for cancers. Traditionally, transcription factors are thought as "undruggable," due to the challenge of using small molecules to disrupt protein–DNA or protein–protein interactions or lack of defined ligand binding sites in transcription factors which allows inhibition of their functions (these features are well known in "druggable" targets, such as enzymes or receptors). With the progresses in elucidation of so-called "hotspot" amino acid residues contributing to the majority of the interaction energy, the leap from "undruggable" to "druggable" becomes reality [3, 26, 27]. At the same time, allosteric modulation of protein–protein interactions [28, 29] and gene therapy [30, 31] have provided alternative and additional approaches for curbing cancer.

In this review, we summarized strategies to directly or indirectly target the transcription factors designed with various approaches, such as targeting post-transcriptional regulation (phosphorylation, ubiquitination, acetylation), targeting protein–protein interaction, targeting new allosteric or ligand-binding site, targeting protein degradation, or targeting gene transcription.

2 Transcription Factors Involved in Tumorigenesis

Over the last decades, the functions and mechanisms of transcription factors in the regulation of cancer initiation and progression have been progressively recognized (*see* Fig. 1). For example, in oral squamous cell carcinoma (OSCC), the transcriptional co-activators YAP and TAZ promote the pro-tumorigenic signals, and hyperactive YAP and TAZ contribute to the onset of OSCC through promotion of OSCC cell proliferation, survival, and migration in vitro and tumor growth and metastasis in vivo [32]. Similarly, the regulatory roles of YAP and TAZ in tumorigenesis have been confirmed in the tissues of the liver [33], breast [34], uterine [35], lung [36], etc. Another critical oncoprotein, c-Myc, has been shown to be implicated in stimulating the progression of various cancers, mainly through its ability to promote cancer cell growth and cellular survival mechanisms and maintaining cancer stem cells [37, 38]. β-catenin plays vital roles in the development and tumorigenesis as the key mediator of Wnt signaling pathway [39]. In prostate cancer, activation of the Wnt/β-catenin pathway stimulates prostate cell proliferation, differentiation, and the EMT, which is thought as the contributor for invasive behavior of tumor cells [40]. FOX proteins are a group of multifarious

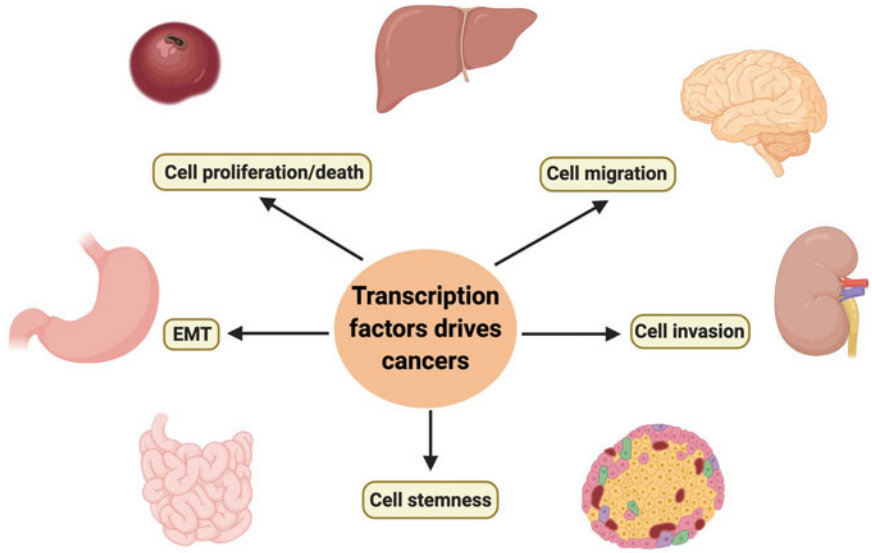


Fig. 1 Transcription factors (TFs) drive cancer. TFs are involved in tumorigenesis via various mechanisms including the regulation of cancer cell proliferation, apoptosis, migration, invasion, and "stemness"

transcription factors implicated in initiation, development, and progression of almost all kinds of cancers [41–43]. In colorectal cancer (CRC), metabolic stress and chemotherapy stimulate the translocation of FOXO3a into the mitochondria to facilitate mitochondrial metabolism and cell survival in tumor cells [43, 44]. In gastrointestinal cancer, FOXM1 was implicated as a critical regulator in the proliferation, migration, and invasion of GI cancer cells, and FOXM1 modulates EMT through its crosstalk with Wnt/β-catenin signaling pathway [45]. RUNX1, RUNX2, and RUNX3 proteins are essential for tissue and organ developmental processes [46]. Disruption of the normal developmental processes has contributed to cancer cell survival, invasion, and EMT [47–50]. RUNXs have been demonstrated to interplay with Wnt/β-catenin signaling pathways. For example, RUNX could either directly modulate β-catenin/TCF-4 transcriptional activity or indirectly target on other Wnt/β-catenin signaling nodes. In a feedback regulation, β-catenin and its transcriptional co-factors could also control RUNX gene expression [51, 52]. The nuclear factor kappa B (NF-κB) consists of a family of transcription factors involved in the regulation of oncogenesis, as well as inflammation and tumor immunity [53]. YAP/TAZ has been targeted by NF-κB through directly transcriptional regulation [54], suggesting that an extensive transcription factor network is involved in tumorigenesis.

Given the critical roles of YAP/TAZ, c-Myc, and β-catenin in cancer development, progression, and metastasis, and their extensive crosstalk with other transcription factors, we will focus on the discussion of (1) the regulatory roles of YAP/TAZ, c-Myc, and

β-catenin in cancer, especially summarizing their aberrant expression and modulation in both transcriptional and post-transcriptional levels in cancers, and (2) the therapeutic strategies developed in the recent years by targeting these transcription factors directly or indirectly.

3 Regulation of Transcriptional Factors in Cancer Development

3.1 c-Myc

The transcription factor c-Myc is a master regulator of cell proliferation, cell growth, cell differentiation, and cell death, by binding to consensus DNA elements (5′-CACGTG-3′) and driving the expression of target genes (*Cyclin D2*, *CDK4*, *p21, p15, CDH2*, and *CEBP*). Regulations of c-myc mRNA and c-Myc protein at transcriptional and post-transcriptional levels are tightly controlled. Deregulation of their levels will have a critical impact on cell proliferation and cell fate. Numerous studies have indicated that the aberrant expression of the c-myc oncogene, due to either transcriptional overexpression (gene amplification, translocation, alterations in upstream signaling pathways) or c-Myc protein stabilization, has been implicated in various cancers, including breast, ovarian, and prostate cancers, leukemia, and lymphoma. Indeed, high c-Myc protein levels are not only able to drive tumor initiation and progression but also essential for tumor maintenance, as sustained c-Myc overexpression is critical to cancer cells and reduction in c-Myc levels leads to growth arrest, apoptosis, and differentiation of cancer cells [55].

3.1.1 Transcriptional Regulation of c-myc

Gene amplification of c-myc is the most common type of c-my deregulation in cancers. C-myc locates in Chromosome *8q24*, a region frequently amplified in cancers (in 18.92% cancers), including leukemia [56], neuroblastoma [57], small cell lung cancer [58], and ovarian, breast, pancreatic, prostate, colorectal, and squamous cell lung cancers [59, 60].

The upstream transcription factors that directly bind with c-myc promoter have been widely studied and reviewed, for example, β-catenin and γ-catenin activate the c-myc promoter at its c-myc's TCF-4 (T-cell factor (4) binding sites, and Wnt signaling, TGFβ signaling, NO (nitric oxide), 1,25 $(OH)_2$-D_3 (1,25-dihydroxyvitamin D_3) signaling, estrogen–ER (estrogen receptor) signaling, androgen–AR (androgen receptor) signaling, mTOR signaling all converge to β-catenin activation to drive c-myc expression [61]. In addition, E2F, Smads (Smad1, Smad2, Smad3, Smad4), METS, BMAL1, CYR1, C/EBPα, STATs (STAT1, STAT3, STAT4), FBP (FUSE binding protein), NF-κB, AP1, CTCF, and FOXOs (FOXM1c, FOXO3) have been reported to directly bind to c-myc promoter and govern its gene expression

[61]. Therefore, signaling pathways that influence the activities of these upstream transcription factors could lead to c-myc upregulation in cancers.

Genetic mutations of c-myc are relatively infrequent, but some studies have found functional mutations in c-myc Homology Box I (HBI) region in Burkitt lymphoma [62, 63]. For example, T58A mutation increased the stability of c-Myc [64–66]. Additional mutations have been found on T244 and P245 residues in lymphomas, among which P245A mutation increased the turnover half-life and stability of c-Myc [67].

3.1.2 Post-transcriptional Regulation of c-Myc

In addition to transcriptional regulation of c-myc gene expression, the post-transcriptional regulation of c-Myc protein was altered in cancers, including phosphorylation, ubiquitination, and acetylation. c-Myc heterodimerizes with Max and then binds to specific E-boxes with the consensus gene sequence 5′-CACGTG-3′ [68]. The post-transcriptional modification of c-Myc, such as phosphorylation and acetylation, would affect its binding with Max and their regulatory roles in downstream gene expressions.

c-Myc Phosphorylation

It was first reported that protein kinase CK2 phosphorylates c-Myc at the acidic domain and near the basic region to stabilize c-Myc protein [69]. Phosphorylation of c-Myc at the transactivation domain (TAD) on Thr58 and Ser62 is important for c-Myc stability and activity [70, 71]. GSK3 and proline-directed kinases, respectively, phosphorylate Thr58 and Ser62 and modulate c-Myc stability [72]. Moreover, Ser62 phosphorylation is prerequisite for GSK3-mediated phosphorylation, which promotes c-Myc ubiquitination and proteasomal degradation, mediated by the binding and recruiting of the SCFFBW7 ubiquitin ligase complex [70, 71]. In addition, mitogen-activated protein kinase (MAPK), c-JUN N-terminal kinase (JNK), and cyclin-dependent kinase 1 (CDK1) have also been implicated in c-Myc Ser-62 phosphorylation [73–79], suggesting that phosphorylation is a common regulation of c-Myc function.

c-Myc Ubiquitination

Besides recruitment of the SCF(FBW7) ubiquitin ligase complex to direct c-Myc ubiquitination, c-Myc is also polyubiquitinylated by the SCF-SKP2 ubiquitin ligase complex [80, 81]. SKP2 and other subunits of the SCF-SKP2 complex initially interacted with c-Myc, which synergistically leads to c-Myc ubiquitination, proteasomal degradation, and inhibition of its transcriptional activity, thereby governing its regulatory downstream under a tightly controlled fashion [82]. Accordingly, SKP2 is recognized as an oncogene and amplified in a subset of cancers [83]. However, the direct evidence of SKP2 and c-Myc protein level correlation is missing, which should be further explored.

c-Myc Acetylation

Lys323, located within the nuclear localization sequence domain (NLS), is modified by both p300 and mGCN5 [84]. However, it remains unclear whether the lysine acetylation of c-Myc affects its binding sites for specific interaction partners, including TRRAP (transformation/transcription domain-associated protein), STAGA (SPT3-TAF9-GCN5L acetylase), and TIP60 (Tat-interactive protein 60 kDa, also termed KAT5) histone acetyltransferase complexes. Since lysine residue can be modified by both ubiquitination and acetylation, these two modifications can potentially interfere with each other. Indeed, activation of lysine acetylation reduces lysine ubiquitination of c-Myc and enhances its stability [85–87]. Thus, ubiquitination and acetylation are tightly interconnected, not only in regulating c-Myc protein stability but potentially also in controlling its association of co-factors.

3.2 YAP/TAZ

The Hippo pathway plays significant roles in modulating cell proliferation, cell fate, and organ size under normal physiological conditions [88–90]. It has been emerging as critical players of tumorigenesis. The deregulation of transcriptional coactivators YAP (Yes-associated protein) and WWTR1 (TAZ) is critical for cancers [91], and the hyperactivation and overexpression of YAP/TAZ have been tightly linked to various cancer types, including breast cancer [92, 93], bladder cancer [94], liver cancer [95], squamous cell carcinoma [96], ovarian cancer [97], and non-small cell lung cancer [98]. Several mechanisms, including transcriptional upregulation and post-transcriptional activation, could lead to hyperactivation of YAP/TAZ.

3.2.1 Transcriptional Regulation of YAP/TAZ

It has been shown that NF-κB transcription factors directly bind to YAP and TAZ promoters and regulate YAP and TAZ transcription in U2OS cells [54]. However, additional evidence in other cancer types is needed to confirm that NF-κB could regulate YAP/TAZ expression. In addition, in pancreatic ductal adenocarcinoma (PDAC) cells, eIF5A-PEAK1 signaling has been shown to contribute to the elevated YAP/TAZ protein levels, but the intermediate factor responsible for YAP/TAZ gene expression remains unknown [99].

3.2.2 Post-transcriptional Regulation of YAP/TAZ

YAP/TAZ Phosphorylation

YAP/TAZ is post-transcriptionally phosphorylated and deactivated by kinases LATS1 and LATS2, which are phosphorylated and activated by MST1 and MST2, as the core regulation of the canonical Hippo pathway (*see* Fig. 2). Deregulation of these kinases leads to YAP/TAZ dephosphorylation and persistent accumulation in the nucleus [89, 100]. Once in the nucleus, YAP/TAZ binds to DNA-binding transcription factors, most notably TEADs (TEAD1, 2, 3, 4) [101], and could also associate with AP1

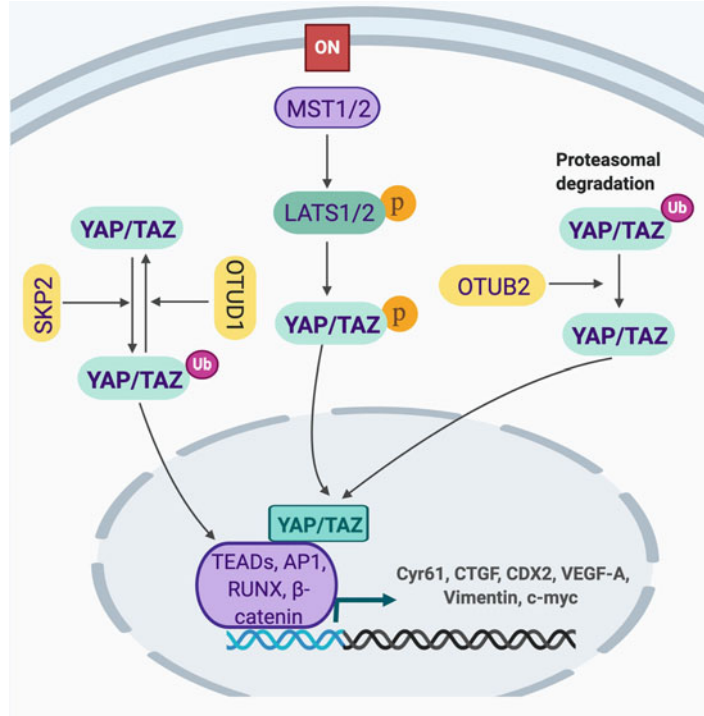


Fig. 2 The regulation of YAP/TAZ. YAP/TAZ could be phosphorylated by LATS1/2, which was controlled by upstream of Hippo pathway. Phosphorylation of YAP/TAZ affects its nuclear translocation and binding with transcription factors, including TEADs, AP1, RUNX, and β-catenin, and their target genes, including Cyr61, CTGF, CDX2, VEGF-A, vimentin, and c-myc. The ubiquitination of YAP/YAZ was dynamically controlled by SKP2, OTUD1, and OTUB2. Ubiquitination of YAP/TAZ affects its proteasomal degradation and translocation to nucleus

[102], RUNXs [90], p73, β-catenin, and ERBB-4 (EGFR family member v-Erb-b2 avian erythroblastic leukemia viral oncogene homolog 4) (*see* Fig. 2) [103]. Several new upstream regulators of Hippo pathways have also been revealed recently, which might offer new targets as potential cancer therapeutics. For example, NUAK2 has been identified as a direct suppressor of Hippo pathway and functions in a feed forward loop and promptly induces YAP/TAZ nuclear translocation, binding with transcriptional partners and concurrent cancer cell and tumor growth [6].

YAP/TAZ Ubiquitination

YAP/TAZ has also been shown to be dynamically ubiquitinated and deubiquitinated. Ubiquitination of YAP/TAZ could direct the proteins for proteinase degradation. In human glioma cells, YAP is ubiquitinated by β-TrCP E3 ubiquitin ligase, and the interaction could be disrupted by ACTL6A, which leads to YAP stabilization and nuclear accumulation [104]. Hence, the hyperactivation of YAP may be responsible for ACTL6A's role in promoting glioma

cell proliferation, migration, and invasion [104]. OTUB2, a deubiquitinating cysteine protease, has been shown to deubiquitinate and activate YAP/TAZ in RAS-transformed MCF10A cells, which is dependent on poly-SUMOylation of OTUB2 on lysine 233 (*see* Fig. 2) [105]. A yet-unknown SUMO-interacting motif (SIM) in YAP and TAZ was required for the association of YAP/TAZ with SUMOylated OTUB2. Importantly, EGF and oncogenic KRAS induce OTUB2 poly-SUMOylation and thereby activate YAP/TAZ. The study revealed a novel mechanism, in which YAP/TAZ activity is induced by oncogenic KRAS [105]. Furthermore, YAP undergoes nonproteolytic, lysine 63 (K63)-linked polyubiquitination by the SCF(SKP2) E3 ligase complex (SKP2) and deubiquitination by the deubiquitinase OTUD1 (*see* Fig. 2). The non-proteolytic ubiquitination of YAP induces its binding with transcription factor TEAD1, thereby retaining YAP's nuclear localization, transcriptional activity, and growth-promoting activity, which is independent of classical Hippo pathway [106].

3.3 β-Catenin

Nuclear accumulation of β-catenin has been manifested in various tumors and is inevitably associated with tumor progression and metastasis. Therefore, precise and highly orchestrated regulation of β-catenin at the transcriptional and posttranslational levels is critical for cancer.

3.3.1 Oncogenic Mutations of β-Catenin

Gain-of-function mutations of β-catenin that lead to stabilized β-catenin have been frequently found in cancers of skin, prostate, ovary, liver, colon, and the endometrium [107–110]. For example, in pilomatricomas, mutations in the N-terminal segment of β-catenin including S33F (TCT→ TTT), S33Y (TCT→ TAT), S37C (TCT→ TGT), S37F (TCT→ TTT), and T41I (ACC→ ATC) lead to the inhibition of GSK-3-mediated phosphorylation of β-catenin and its subsequent ubiquitination and degradation. Stabilized β-catenin leads to persistent accumulation in the cells [109]. In castrate-resistant prostate cancer (CRPC), β-catenin forms complex with AR and potentiates AR signaling [111]. In addition, MDA PCa 118a and MDA PCa 118b prostate cancer cells carry β-catenin D32G mutation, which leads to enhanced nuclear localization of β-catenin and increase of its downstream target gene HAS2 (hyaluronan synthase 2) expression [110].

3.3.2 Posttranslational Regulation of β-Catenin

β-catenin Phosphorylation

In the absence of WNT ligands, WNT receptor complexes (Fz/LRP/CKI γ/Axin/GSK3) fail to bind β-catenin, and CK1 and GSK3α/β sequentially phosphorylate β-catenin (*see* Fig. 3) [112]. Phosphorylated β-catenin is then ubiquitinated by the F box/WD repeat protein β-TrCP, a component of a dedicated E3 ubiquitin ligase complex, subsequently leading to its rapid degradation by the proteasome (*see* Fig. 3) [112]. In contrast, in the presence of WNT ligands, β-catenin degradation is blocked,

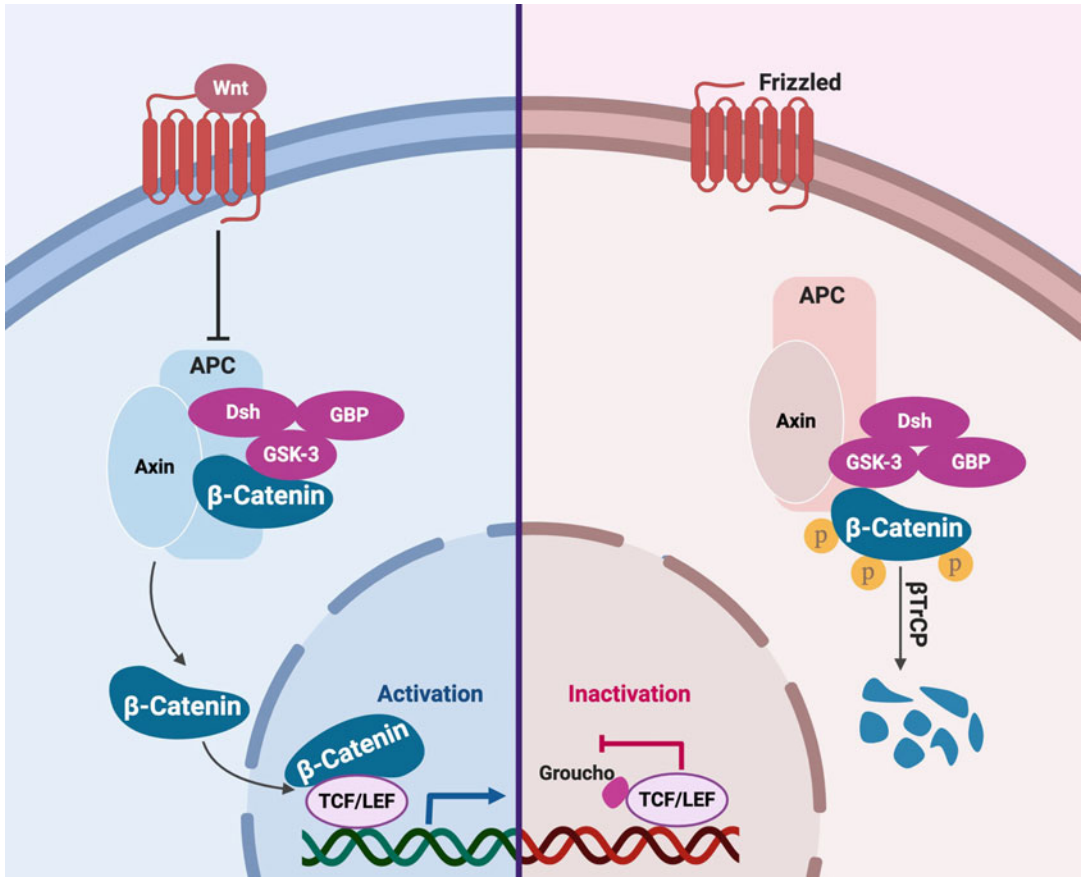


Fig. 3 The regulation of β-catenin. β-catenin was tightly controlled by Wnt signaling, which affects its association with complexes APC, Dsh, GBP, and GSK-3. When Wnt signaling is triggered, β-catenin will be shuttled into the nucleus to initiate its binding with transcription factors TCF/LEF and target gene expression. When Wnt signaling is frizzled, β-catenin will be phosphorylated and subjected to degradation

which leads to nuclear translocation of β-catenin and its binding with T-cell factor (TCF) and lymphoid enhancer-binding protein (LEF) and activation of their target gene transcriptions, including c-myc, cyclin D-1, and metalloproteinase, which are essential regulators of cell growth, proliferation, and EMT transition [10, 109, 113].

3.4 Interplays Among YAP/TAZ, c-MYC, and β-Catenin

β-Catenin has been shown to induce c-Myc expression by activating c-myc promoter, which harbors several TCF-4 (T-cell factor 4) binding sites (*see* Fig. 4). It is also known that Wnt signaling, TGFβ signaling, NO (nitric oxide), 1,25 $(OH)_2$-D_3 (1,25-dihydroxyvitamin D_3) signaling, estrogen-ER (estrogen receptor) signaling, Androgen-AR (androgen receptor) signaling, and mTOR signaling all converge to β-catenin to regulate c-Myc expression [61]. Meanwhile, YAP/TAZ could bind to β-catenin, which is vital for β-catenin-TCF-medicated c-myc transcription

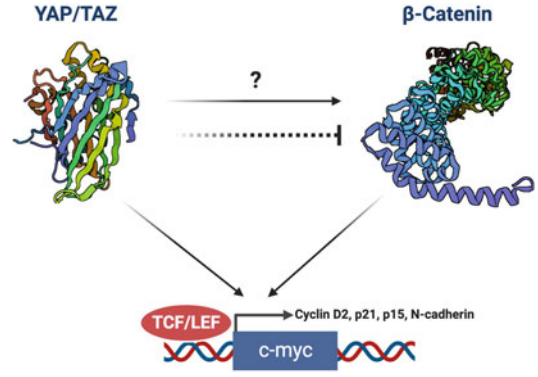


Fig. 4 Interplays among YAP/TAZ, β-catenin, and c-Myc. YAP/TAZ could affect β-catenin protein levels, either promotion or suppression. However, the mechanism remains unknown, which has been labeled with question mark. At the same time, YAP/TAZ and β-catenin could promote c-myc gene expression

[103, 114, 115]. At the protein level, cytoplasmic YAP may directly sequester β-catenin into the cytoplasm (*see* Fig. 4). On the other hand, cytoplasmic TAZ may sequester DVL2 to impede its activity in promoting β-catenin accumulation in the condition of Wnt stimulation [116]. Additionally, YAP directly increased β-catenin level, which may be due to the blocking of β-Trcp-dependent β-catenin degradation (*see* Fig. 4) [115]. Recently, we and others have shown that nuclear YAP/TAZ could interact with Groucho/TLE to inhibit T-cell factor (TCF)-mediated transcription in intestinal stem cells, suggesting that the crosstalk between these pathways is extensive and complicated [117].

Taken together, targeting these signaling nodes may lead to promising therapeutic strategies in cancer treatment. Traditionally, transcription factors were considered as "undruggable" due to the difficulties of targeting protein–DNA binding and protein–protein interaction where defined small molecule binding pockets might be lacking [2, 3]. With the elucidation of above-mentioned new knowledge of YAP/TAZ, c-Myc, and β-catenin modifications and regulation, the inventions of potential therapeutic agents could be possible.

4 Targeting Transcription Factors for Drug Discovery

4.1 Targeting Posttranslational Regulation

4.1.1 YAP/TAZ

In renal cell carcinoma (RCC), dasatinib, a second-generation tyrosine kinase inhibitor, suppressed RCC cell viability in vitro and decreased tumor growth in vivo. Mechanistically, dasatinib directly inhibits Src kinase and subsequently activates Src-JNK-LIMD1-LATS signaling cascade, leading to YAP phosphorylation and suppression of YAP/TAZ-TEAD target genes (such as CTGF, Cyr61, and AJUBA) (Table 1) [118]. In MDA-MB-231, H1299, and HCT-116 cells, statins blocked YAP/TAZ nuclear localization

Table 1
Summary of drugs targeting on YAP/TAZ, β-catenin, and c-Myc

Inhibitors	Target	References
Targeting posttranslational regulation		
Dasatinib	YAP phosphorylation	118
Statins	YAP phosphorylation	119,120,121
Norcantharidin (NCTD)	YAP phosphorylation	122
Dobutamine	YAP phosphorylation	123,124
GDK-100017	β-catenin phosphorylation	125,126
Genistein	β-catenin ubiquitination	127,128
Z86	β-catenin phosphorylation	129,130
Targeting protein–protein interactions		
MYCMI-6	c-Myc–MAX interaction	131
10074-G5	c-Myc–MAX interaction	132
JY-3-094	c-Myc–MAX interaction	133
3jc48-3	c-Myc–MAX interaction	134
KJ-Pyr-9	c-Myc–MAX interaction	135
KSI-3716	c-Myc–MAX interaction	136
sAJM589	c-Myc–MAX interaction	137
Super-TDU	YAP–TEAD interaction	138
Verteporfin	YAP–TEAD interaction	139,140
MGH-CP1	YAP–TEAD interaction	117
ICG-001	β-catenin–CBP interaction	145
NLS-StAx-h	β-catenin–TCF interaction	146
CRT inhibitors	β-catenin–TCF interaction	147
PKF115-584	β-catenin–TCF interaction	148
CGP049090	β-catenin–TCF interaction	148
Henryin	β-catenin–TCF4 interaction	149
Peptoid–peptide	β-catenin–TCF interaction	152
Targeting new allosteric or ligandable site		
Celastrol	c-Myc–DNA interaction	153
α-helix mimetics	c-Myc–DNA interaction	154
Targeting TF degradation		
Dihydroartemisinin	c-Myc	155

(continued)

Table 1
(continued)

Inhibitors	Target	References
JW55	c-Myc	156
MSAB	β-catenin	157
YW2065	β-catenin	158
Nucleic acid-based drugs		
c-myc-As-ODN	c-Myc	159
PMO	c-Myc	161
Se2SAP	c-Myc	165
DC-34	c-Myc	166
IZCZ-3	c-Myc	167
DCR-BCAT	β-catenin	168

and transcriptional responses via inhibition of HMG-CoA reductase, the rate-limiting enzyme of the mevalonate cholesterol biosynthesis pathway. Such inhibition leads to reduction of geranylgeranyl pyrophosphate levels, which is required for membrane localization and activation of RHO GTPases, a key upstream regulator of YAP [119–121]. Norcantharidin (NCTD) inhibited non-small cell lung carcinoma (NSCLC) progression and metastasis via cell cycle arrest, enhancing apoptosis and inducing senescence dependent on the modulation of YAP's translocation between the cytoplasm and nucleus (Table 1) [122]. In addition, the clinical drug dobutamine has been demonstrated to induce the YAP accumulation in the cytosol and YAP-dependent gene transcription in human osteoblastoma U2OS cells independent of Hippo pathway [123], which has been recapitulated in human gastric adenocarcinoma SGC-7901 cells [124].

4.1.2 *β-Catenin*

In non-small cell lung cancer A549/Wnt2 cells (with overexpression of human Wnt2), GDK-100017, a 2,3,6-trisubstituted quinoxaline derivative, suppressed cell proliferation via arresting cell cycle, which is associated with its reduction on β-catenin nuclear localization, β-catenin-TCF/LEF-dependent transcriptional activity, and target gene expression (cyclin D1, *etc.*) (Table 1) [125, 126]. In colorectal cancer cell line SW480, the natural flavonoid genistein inhibited cell proliferation via suppressing β-catenin/TCF transcriptional activity. Mechanistically, genistein promoted the ubiquitination and degradation of β-catenin through targeting the phosphorylation of AKT-GSK3β-β-catenin signaling cascade (Table 1) [127, 128]. In colorectal cancer HCT116 cells and associated xenografted tumor, isopropyl 9-ethyl-1-

(naphthalen-1-yl)-9H-pyrido[3,4-b]indole-3-carboxylate (Z86) inhibited cell growth and tumor growth through suppression of GSK3β (Ser9) phosphorylation and activation of its activity and subsequently promoting the phosphorylation and degradation of β-catenin (Table 1) [129, 130].

4.2 Targeting Protein-Protein Interactions

4.2.1 c-Myc

In a panel of neuroblastoma cell lines and xenograft tumor models, MYCMI-6 was identified as a potent and selective inhibitor of c-Myc/MAX interaction, binding exclusively to the c-Myc bHLHZip domain and suppressing c-Myc-driven transcription [131]. Phenotypically, MYCMI-6 inhibits tumor cell growth in a c-Myc-dependent manner and promotes massive apoptosis in tumor tissue. c-Myc inhibitor 10074-G5 (N-([1,1′- biphenyl]-2-yl)-7-nitrobenzo[c][1,2,5]oxadiazol-4-amine) targets a hydrophobic domain of c-Myc and perturbs the interaction of c-Myc and Max (Table 1). The ortho-biphenyl group 10074-G5 replaced by a para-carboxyphenyl group yielded the new inhibitor JY-3-094, which exhibits improved selectivity over Max–Max homodimers and physicochemical properties [132]. Another analogue of 10074-G5, named 3jc48-3, is 5 times more potent in blocking c-Myc–Max dimerization, leading to inhibition of the proliferation of c-Myc hyperactive human leukemia HL60 and Burkitt's lymphoma Daudi cells (Table 1) [133]. A novel small molecule inhibitor of c-Myc, KJ-Pyr-9, has been identified from a Kröhnke pyridine library. KJ-Pyr-9 disrupted c-Myc–MAX complex formation in cells, leading to blockage of c-Myc-induced oncogenic transformation in cell culture and suppression of the growth of a xenotransplant of MYC-amplified human cancer cells (Table 1) [134]. In bladder cancer cells and xenograft tumor model, c-Myc/MAX binding inhibitor KSI-3716 decreased the expression of c-Myc target genes, such as cyclin D2, CDK4, and hTERT; exerted cytotoxic effects by inducing cell cycle arrest and apoptosis; and blocked tumor growth [135]. In a Burkitt lymphoma P493-6 cell model, sAJM589, a novel small molecule c-Myc inhibitor, potently perturbs the formation of c-Myc−Max heterodimer, preferentially inhibits transcription of c-Myc target genes, and inhibited proliferation of P493-6 cells (Table 1) [136]. However, all these inhibitors lack sufficient potency, selectivity, and toxicity profile to be advanced to human clinical testing. Further efforts to develop specific inhibitors are still needed.

4.2.2 YAP/TAZ

In gastric cancer, downregulation of VGLL4 was correlated with upregulation of YAP and YAP/TEADs target genes, and VGLL4 directly competes with YAP for binding to TEADs. Importantly, VGLL4's tandem Tondu (TDU) domains are not only necessary but also sufficient for its inhibitory activity toward YAP. A peptide mimicking this function of VGLL4 (super-TDU) potently suppressed tumor growth in vitro and in vivo [137]. Mechanistically, super-TDU disrupts YAP–TEAD interaction and YAP–TEAD

target genes, including *CTGF*, *Cyr61*, and *CDX2* (Table 1) [137]. In human hepatocellular carcinoma (HCC), verteporfin, a benzoporphyrin derivative, inhibited cell growth of HCC cells via disruption of YAP–TEAD binding and expression of their target genes (Table 1) [138, 139]. In human retinoblastoma cell lines (Y79 and WERI), verteporfin dose-dependently suppressed cell proliferation and migration and inhibited tumor angiogenesis through inhibition of YAP-TEAD binding and downstream target genes, such as c-myc, CTGF, Cyr61, and VEGF-A (Table 1) [140]. YAP/TAZ drives cancer cell survival and BRAF inhibitor resistance in melanoma [141, 142]. Compared with BRAF inhibitor (BRAFi) sensitive melanoma cancer stem cells (MCS cells), YAP1, TAZ, and TEAD protein levels were significantly increased in BRAFi-resistant MCS cells, which is accompanied by elevated cell survival, spheroid formation, invasion in Matrigel assays, and tumor formation [143]. In xenograft tumor model, verteporfin mitigated YAP1/TAZ level induced by BRAFi resistance, restored BRAF inhibitor suppression of ERK1/2 signaling, and reduced tumor growth in BRAFi-resistant tumors [143]. However, it remains unknown how verteporfin modulates YAP/TAZ and TEAD functions, which warrants future explorations before its further applications.

Most recently, our lab has identified a specific inhibitor of TEAD palmitoylation MGH-CP1, which attenuated the interaction between YAP and TEAD and its downstream regulatory events through inhibition of TEAD autopalmitoylation, which provided new insights of targeting these transcription factors [117].

4.2.3 β-Catenin

ICG-001, a small molecule that downregulates β-catenin-TCF signaling, specifically binds to cyclic AMP response element-binding protein (CBP) and disrupts the β-catenin–CBP interaction (Table 1) [144]. Phenotypically, ICG-001 induces apoptosis and reduces growth of human colon carcinoma SW480, SW620, and HCT116 cells, but not normal colon cells in vitro, and is efficacious in the xenograft mouse models of colon cancer. Likewise, KRAS activation has been found to induce the CBP–β-catenin interaction in pancreatic cancer, and ICG-001 sensitizes pancreatic cancer cells and tumors to gemcitabine (deoxycytidine analog) treatment, possibly through antagonizing CBP–β-catenin interaction [145]. NLS-StAx-h, a selective, cell-permeable, stapled peptide inhibitor, suppresses the interaction between β-catenin and TCF/LEF transcription factors and inhibition of target gene transcription (Table 1). It showed good cellular uptake and profound inhibitory effects on proliferation and migration of colorectal cancer cell lines DLD-1 and SW-480 [146].

An RNAi-based modifier screening strategy was exploited for the identification of specific β-catenin-responsive transcription (CRT) inhibitors without affecting the degradation of β-catenin.

These inhibitory compounds functioned specifically in antagonizing the transcriptional function of nuclear β-catenin, such as blocking β-catenin-TCF-induced target genes and phenotypes in various mammalian and cancer cell lines (Table 1) [147]. It is of great interest to note that these CRT inhibitors are specifically cytotoxic to human colon tumor biopsy cultures as well as colon cancer cell lines with deregulated Wnt signaling. Novartis collections yielded eight compounds with a dose-dependent inhibition of β-catenin-TCF binding and target gene transcription from approximately 7000 natural products and 45000 synthetic compounds. Two structurally related compounds (PKF115-584 and CGP049090) proved to be effective in suppressing Wnt reporter gene activity and colon cancer cell proliferation, among which PKF115-584's inhibitory effect on Wnt signaling was confirmed in xenograft models of human multiple myeloma (Table 1) [148]. Henryin, an ent-kaurane diterpenoid isolated from *Isodon rubescens* var. *lushanensis*, selectively inhibits the proliferation of human colorectal cancer HCT116 cells through inhibiting the association of β-catenin/TCF4 transcriptional complex and the transcription of target genes, such as Cyclin D1 and C-myc (Table 1) [149]. In 2019, a group of small molecule inhibitors specifically disrupting the β-catenin–TCF protein–protein interaction without affecting the β-catenin–E-cadherin and β-catenin–APC interactions have been synthesized and were reported to inhibit migration and invasiveness of Wnt/β-catenin-dependent cancer cells [150].

Peptoids, or poly-N-substituted glycines, are a series of peptidomimetic oligomers in which the side chains are presented to the nitrogen atom of the peptide backbone instead of the α-carbons as they are in amino acids. They have been proposed as capable of curbing protein–protein interactions through mimicking motifs of protein secondary structures [151]. Using the Rosetta suite of protein design algorithms, a small library of peptoid–peptide macrocycles has been designed in silico, based on the prediction of binding to β-catenin. Cell-based luciferase assays were further used to test their inhibitory effects on Wnt signaling [152]. Interestingly, inhibitors which are potently blocking β-catenin–TCF interaction have been identified and significantly inhibit the proliferation of prostate cancer cells in vitro and inhibit Wnt signaling in vivo in a zebrafish model [152].

4.3 Targeting New Allosteric or Ligandable Site of TF

Allosteric modulation is generally recognized as one of the most direct and efficient ways to govern protein functions. Targeting allosteric or ligandable sites has attracted great attentions for drug development due to their high selectivity and potential to target many previously "undruggable" targets.

Most c-Myc inhibitors perturb the binding of c-Myc and its obligate heterodimerization partner Max through their respective bHLH-ZIP domains. However, the natural triterpenoid celastrol

and its derivatives bind to and alter the quaternary structure of the preformed dimer and abrogate its DNA binding (Table 1). Phenotypically, the triterpenoids suppressed the proliferation of multiple myeloma, non-small cell lung cancer, and breast cancer cell lines [153]. Using biophysical methods including NMR spectroscopy and surface plasmon resonance, novel, low-molecular-weight, synthetic α-helix mimetics have been designed, which could bind to helical c-Myc in its transcriptionally active coiled-coil structure in association with Max. These compounds disrupted the heterodimer's binding to its canonical E-box DNA sequence without causing protein–protein dissociation and blocked the proliferation of c-Myc-overexpressing cell lines (Table 1) [154].

4.4 Targeting TF Degradation

In colorectal cancer HCT116 cells, dihydroartemisinin (DHA), the main active metabolite of artemisinin, induced significant apoptosis through promoting the degradation of c-Myc protein (Table 1), which was mitigated by proteasome inhibitor MG-132 or GSK 3β inhibitor LiCl [155]. However, the precise mechanisms of how DHA induces c-Myc degradation require further studies.

4.4.1 c-Myc

4.4.2 β-Catenin

Colon carcinoma cells with mutations in the APC (adenomatous polyposis coli) locus or in an allele of β-catenin have been related to hyperactivation of Wnt signaling. JW55 (a novel TNKS inhibitor) has been identified to stimulate β-catenin degradation, which is fulfilled through inhibition of the PARP domain of tankyrase 1 and tankyrase 2 (TNKS1/2) and induction of the β-catenin destruction complex (Table 1) [156]. In concrete, the inhibitory effect of JW55 on TNKS1/2 poly(ADP-ribosyl)ation activity contributes to stabilization of AXIN2 and increased degradation of β-catenin [156]. With a TCF-dependent luciferase-reporter assay, MSAB (methyl 3-{[(4-methylphenyl)sulfonyl]amino}benzoate) was identified as a selective inhibitor of Wnt/β-catenin signaling through its binding and targeting on β-catenin degradation, which is accompanied by downregulation of Wnt/β-catenin target genes and tumor inhibitory effects selectively on Wnt-dependent cancer cells in vitro and in mouse cancer models (Table 1) [157]. In colorectal cancer SW480 and SW620 cells and mice xenograft model, YW2065 (1c) exerted excellent anti-tumor effects by stabilizing Axin-1, a scaffolding protein that induces proteasome degradation of β-catenin [158].

4.5 Nucleic Acid-Based Therapy

Genetic methods, such as antisense RNA-, RNAi-, or CRISPR/Cas-based gene therapy, can be achieved through inhibition or replacement of a mutated gene and inactivation or reconstruction of a deregulated gene to combat the disease. Due to its specificity, they have drawn great attentions in the past two decades for the treatment of a wide spectrum of cancers.

In human prostate cancer cell lines, such as LNCaP, PC3, and DU145 cells, c-myc-antisense-oligonucleotide and c-myc-As-ODN treatment has shown to time- and dose-dependently reduce DNA synthesis and cell viability (Table 1) [159]. Similarly, in human prostate cancer cell lines, such as LNCaP, PC3, and DU145 cell lines, and PC-3 androgen-independent human prostate cancer xenograft murine model, a novel antisense phosphorodiamidate morpholino oligomer AVI-4126 directly targets c-myc mRNA and reduced its translation, leading to significant apoptosis and growth inhibition in prostate cancer cells and in subcutaneous tumor xenografts (Table 1) [160]. In the LLC1 syngeneic murine lung metastasis tumor model, AVI-4126, a neutral antisense phosphorodiamidate morpholino oligomer (PMO), specifically inhibits c-myc expression and decreased tumor burden and number of tumorlets formed in the lung, decreased mitotic activity, but increased rate of apoptosis (Table 1) [161].

G-quadruplexes (G4s) are noncanonical DNA structures that frequently occur in the promoter regions of oncogenes, such as c-myc, and c-myc G4 stabilizer has been demonstrated to mitigate c-myc expression [162–164]. A core-modified expanded porphyrin analogue, 5,10,15,20-[tetra(N-methyl- 3-pyridyl)]-26,28-diselenasapphyrin chloride (Se2SAP), selectively binds with the c-myc G-quadruplex and inhibits its expression (Table 1) [165]. In multiple myeloma (MM) cells, DC-34, a small molecule, significantly decreases c-myc transcription in a G4-dependent manner [166]. The specific contact responsible for affinity and selectivity of MYC G4 and DC-34 was confirmed by NMR spectroscopy. Furthermore, with the aid of structural modification of aryl-substituted imidazole/ carbazole conjugates, a brand-new, four-leaf clover-like ligand IZCZ-3 was synthesized to preferentially bind and stabilize the c-myc G-quadruplex and suppress c-myc expression (Table 1) [167]. Cellular and physiological studies revealed IZCZ-3's promotive role in cell cycle arrest and apoptosis, thus inhibiting cell growth in squamous cell carcinoma SiHa cells and suppressing tumor growth in SiHa xenograft model, mainly through curbing c-myc transcription by exclusive targeting on the promoter G-quadruplex structure.

Wnt/β-catenin signaling mediates cancer immune evasion and resistance to immune checkpoint therapy, in part by blocking cytokines that trigger immune cell recruitment. DCR-BCAT, a nanoparticle drug product containing a chemically optimized RNAi, triggers silencing of β-catenin, which significantly increases T cell infiltration and potentiated the sensitivity of the tumors to checkpoint inhibition (Table 1). The combination of DCR-BCAT and immunotherapy yielded significantly greater tumor growth inhibition (TGI) compared to monotherapy in B16F10 melanoma, 4T1 mammary carcinoma, Neuro2A neuroblastoma, and Renca renal adenocarcinoma [168].

Thus, nucleic acid-based approach would be a promising strategy in curbing cancers due to its direct effects on silencing oncogenes. However, the challenge that to what extent an oncogene should be downregulated or upregulated in a controlled manner without side effect ought to be overcome before its clinical trial and application. In addition, delivery and stability of these agents need further improvement. Off-target effects of RNAi-based approach should also be very careful evaluated. With the advancement of cutting-edge gene technologies, like the state-of-the-art technology CRISPR (clustered regularly interspaced short palindromic repeats), specific and precise modulation of genes to antagonize cancers would prosper in the near future.

5 Conclusion

Transcription factors play vital roles in tumorigenesis. Although technically challenging, directly targeting these important proteins is critical for development of promising therapeutics. The current review summarized the deregulation of three leading transcription factors and co-factors, including YAP/TAZ, c-Myc, and β-catenin, such as their transcriptional and post-transcriptional regulation in cancer, which provides insights to design drugs to target these traditionally "undruggable" spots and make it "druggable." However, future exploitation of their precise mechanisms of action, dosage and duration, and effectiveness and efficacy in clinical settings is needed to further make them "druggable."

Acknowledgments

The authors would like to thank Melanoma Research Alliance (MRA)-Samuel M. Fisher Memorial Established Investigator award and NIH R01s (5R01DK107651 and 1R01CA238270-01) for the support of our work. The figures are created with BioRender.com.

References

1. Latchman DS (1997) Transcription factors: an overview. Int J Biochem Cell Biol 29(12): 1305–1312
2. Darnell JE (2002) Transcription factors as targets for cancer therapy. Nat Rev Cancer 2(10):740–749
3. Bushweller JH (2019) Targeting transcription factors in cancer—from undruggable to reality. Nat Rev Cancer 19(11):611–624
4. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K et al (2010) An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140(5):744–752
5. Pobbati AV, Hong W (2020) A combat with the YAP/TAZ-TEAD oncoproteins for cancer therapy. Theranostics 10(8):3622
6. Gill MK, Christova T, Zhang YY, Gregorieff A, Zhang L, Narimatsu M et al (2018) A feed forward loop enforces YAP/TAZ signaling during tumorigenesis. Nat Commun 9(1):1–13

7. Meyer N, Penn LZ (2008) Reflecting on 25 years with MYC. Nat Rev Cancer 8(12): 976–990
8. Kalkat M, De Melo J, Hickman KA, Lourenco C, Redel C, Resetca D et al (2017) MYC deregulation in primary human cancers. Gene 8(6):151
9. Clevers H (2006) Wnt/β-catenin signaling in development and disease. Cell 127(3): 469–480
10. Moon RT, Kohn AD, De Ferrari GV, Kaykas A (2004) WNT and β-catenin signalling: diseases and therapies. Nat Rev Genet 5(9): 691–701
11. Wang T, Zheng L, Wang Q, Hu YW (2018) Emerging roles and mechanisms of FOXC2 in cancer. Clin Chim Acta 479:84–93
12. Alasiri G, Fan LYN, Zona S, Goldsbrough IG, Ke HL, Auner HW, Lam EWF (2018) ER stress and cancer: the FOXO forkhead transcription factor link. Mol Cell Endocrinol 462:67–81
13. Li Y, Zhang Y, Yao Z, Li S, Yin Z, Xu M (2016) Forkhead box Q1: a key player in the pathogenesis of tumors. Int J Oncol 49(1): 51–58
14. Han B, Bhowmick N, Qu Y, Chung S, Giuliano AE, Cui X (2017) FOXC1: an emerging marker and therapeutic target for cancer. Oncogene 36(28):3957–3963
15. Nestal de Moraes G, Carneiro LDT, Maia RC, Lam EWF, Sharrocks AD (2019) FOXK2 transcription factor and its emerging roles in cancer. Cancer 11(3):393
16. Gartel AL (2017) FOXM1 in cancer: interactions and vulnerabilities. Cancer Res 77(12): 3135–3139
17. Farhan M, Wang H, Gaur U, Little PJ, Xu J, Zheng W (2017) FOXO signaling pathways as therapeutic targets in cancer. Int J Biol Sci 13(7):815
18. Lee H, Jeong AJ, Ye SK (2019) Highlighted STAT3 as a potential drug target for cancer therapy. BMB Rep 52(7):415
19. Kaltschmidt C, Banz-Jansen C, Benhidjeb T, Beshay M, Förster C, Greiner J et al (2019) A role for NF-κB in organ specific cancer and cancer stem cells. Cancer 11(5):655
20. Otálora-Otálora BA, Henríquez B, López-Kleine L, Rojas A (2019) RUNX family: Oncogenes or tumor suppressors. Oncol Rep 42(1):3–19
21. Khachigian LM (2018) The Yin and Yang of YY 1 in tumor growth and suppression. Int J Cancer 143(3):460–465
22. Atsaves V, Leventaki V, Rassidakis GZ, Claret FX (2019) AP-1 transcription factors as regulators of immune responses in cancer. Cancer 11(7):1037
23. Muller PA, Vousden KH (2013) p53 mutations in cancer. Nat Cell Biol 15(1):2–8
24. Whibley C, Pharoah PD, Hollstein M (2009) p53 polymorphisms: cancer implications. Nat Rev Cancer 9(2):95–107
25. de la Vega MR, Chapman E, Zhang DD (2018) NRF2 and the Hallmarks of Cancer. Cancer Cell 34(1):21–43
26. Arkin MR, Tang Y, Wells JA (2014) Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chem Biol 21(9):1102–1114
27. Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. Nat Rev Drug Discov 3(4):301–317
28. Silvian LF, Friedman JE, Strauch K, Cachero TG, Day ES, Qian F et al (2011) Small molecule inhibition of the TNF family cytokine CD40 ligand through a subunit fracture mechanism. ACS Chem Biol 6(6):636–647
29. Illendula A, Gilmour J, Grembecka J, Tirumala VSS, Boulton A, Kuntimaddi A et al (2016) Small molecule inhibitor of CBF-β-RUNX binding for RUNX transcription factor driven cancers. EBioMedicine 8:117–131
30. Li T, Kang G, Wang T, Huang H (2018) Tumor angiogenesis and anti-angiogenic gene therapy for cancer. Oncol Lett 16(1): 687–702
31. Zaimy MA, Saffarzadeh N, Mohammadi A, Pourghadamyari H, Izadi P, Sarli A et al (2017) New methods in the diagnosis of cancer and gene therapy of cancer based on nanoparticles. Cancer Gene Ther 24(6):233–243
32. Hiemer SE, Zhang L, Kartha VK, Packer TS, Almershed M, Noonan V et al (2015) A YAP/TAZ-regulated molecular signature is associated with oral squamous cell carcinoma. Mol Cancer Res 13(6):957–968
33. Bisso A, Filipuzzi M, Figueroa GPG, Brumana G, Biagioni F, Doni M et al (2019) Cooperation between MYC and β-catenin in liver tumorigenesis requires. Yap/Taz bioRxiv:819631
34. Wu Q, Li J, Sun S, Chen X, Zhang H, Li B, Sun S (2017) YAP/TAZ-mediated activation of serine metabolism and methylation regulation is critical for LKB1-deficient breast cancer progression. Biosci Rep 37(5):1–6
35. Wang C, Jeong K, Jiang H, Guo W, Gu C, Lu Y, Liang J (2016) YAP/TAZ regulates the insulin signaling via IRS1/2 in endometrial cancer. Am J Cancer Res 6(5):996

36. Horie M, Saito A, Ohshima M, Suzuki HI, Nagase T (2016) YAP and TAZ modulate cell phenotype in a subset of small cell lung cancer. Cancer Sci 107(12):1755–1766
37. Chanvorachote P, Sriratanasak N, Nonpanya N (2020) C-myc contributes to malignancy of lung cancer: a potential anticancer drug target. Anticancer Res 40(2):609–618
38. Elbadawy M, Usui T, Yamawaki H, Sasaki K (2019) Emerging roles of C-Myc in Cancer stem cell-related signaling and resistance to cancer chemotherapy: a potential therapeutic target against colorectal cancer. Int J Mol Sci 20(9):2340
39. Miyoshi K, Hennighausen L (2003) β-Catenin: a transforming actor on many stages. Breast Cancer Research 5(2):63
40. Kypta RM, Waxman J (2012) Wnt/β-catenin signalling in prostate cancer. Nature Rev Urol 9(8):418
41. Elian FA, Yan E, Walter MA (2018) FOXC1, the new player in the cancer sandbox. Oncotarget 9(8):8165
42. Nandi D, Cheema PS, Jaiswal N, Nag A (2018) FoxM1: repurposing an oncogene as a biomarker. In: Seminars in cancer biology. Academic Press, pp 74–84
43. Laissue P (2019) The forkhead-box family of transcription factors: key molecular players in colorectal cancer pathogenesis. Mol Cancer 18(1):1–13
44. Grossi V, Fasano C, Celestini V, Lepore Signorile M, Sanese P, Simone C (2019) Chasing the FOXO3: Insights into Its New Mitochondrial Lair in Colorectal Cancer Landscape. Cancer 11(3):414
45. Zhang J, Niu Y, Huang C (2017) Role of FoxM1 in the progression and epithelial to mesenchymal transition of gastrointestinal Cancer. Recent Patents Anti-Cancer Drug Dis 12(3):247–259
46. Brenner O, Levanon D, Negreanu V, Golubkov O, Fainaru O, Woolf E, Groner Y (2004) Loss of Runx3 function in leukocytes is associated with spontaneously developed colitis and gastric mucosal hyperplasia. Proc Natl Acad Sci 101(45):16016–16021
47. Araki K, Osaki M, Nagahama Y, Hiramatsu T, Nakamura H, Ohgi S, Ito H (2005) Expression of RUNX3 protein in human lung adenocarcinoma: implications for tumor progression and prognosis. Cancer Sci 96(4): 227–231
48. Blyth K, Cameron ER, Neil JC (2005) The RUNX genes: gain or loss of function in cancer. Nat Rev Cancer 5(5):376–387
49. Endo T, Ohta K, Kobayashi T (2008) Expression and function of Cbfa-1/Runx2 in thyroid papillary carcinoma cells. J Clin Endocrinol Metabol 93(6):2409–2412
50. Chuang LSH, Ito K, Ito Y (2017) Roles of RUNX in solid tumors. In: RUNX Proteins in Development and Cancer. Springer, Singapore, pp 299–320
51. Sun J, Li B, Jia Z, Zhang A, Wang G, Chen Z et al (2018) RUNX3 inhibits glioma survival and invasion via suppression of the β-catenin/TCF-4 signaling pathway. J Neuro-Oncol 140(1):15–26
52. Sweeney K, Cameron ER, Blyth K (2020) Complex Interplay between the RUNX Transcription Factors and Wnt/β-Catenin Pathway in Cancer: A Tango in the Night. Mol Cell 43(2):188
53. Tang C, Zhu G (2019) Classic and novel signaling pathways involved in cancer: targeting the NF-KB and syk signaling pathways. Curr Stem Cell Res Ther 14(3):219–225
54. Ferraiuolo M, Pulito C, Finch-Edmondson-M, Korita E, Maidecchi A, Donzelli S et al (2018) Agave negatively regulates YAP and TAZ transcriptionally and post-translationally in osteosarcoma cell lines. Cancer Lett 433: 18–32
55. Koh CM, Sabò A, Guccione E (2016) Targeting MYC in cancer therapy: RNA processing offers new opportunities. BioEssays 38(3): 266–275
56. Collins S, Groudine M (1982) Amplification of endogenous myc-related DNA sequences in a human myeloid leukaemia cell line. Nature 298(5875):679–681
57. Schwab M, Alitalo K, Klempnauer KH, Varmus HE, Bishop JM, Gilbert F et al (1983) Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature 305(5931):245–248
58. Nau MM, Brooks BJ, Battey J, Sausville E, Gazdar AF, Kirsch IR et al (1985) L-myc, a new myc-related gene amplified and expressed in human small cell lung cancer. Nature 318(6041):69–73
59. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C (2013) Emerging landscape of oncogenic signatures across human cancers. Nat Genet 45(10): 1127–1133
60. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B et al (2013) Pan-cancer patterns of somatic copy number alteration. Nat Genet 45(10):1134–1140

61. Wierstra I, Alves J (2008) The c-myc promoter: still MysterY and challenge. Adv Cancer Res 99:113–333
62. Cowling VH, Turner SA, Cole MD (2014) Burkitt's lymphoma-associated c-Myc mutations converge on a dramatically altered target gene response and implicate Nol5a/Nop56 in oncogenesis. Oncogene 33(27):3519–3527
63. Bonilla X, Parmentier L, King B, Bezrukov F, Kaya G, Zoete V et al (2016) Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. Nat Genet 48(4):398
64. Bhatia K, Huppi K, Spangler G, Siwarski D, Iyer R, Magrath I (1993) Point mutations in the c–Myc transactivation domain are common in Burkitt's lymphoma and mouse plasmacytomas. Nat Genet 5(1):56–61
65. Bahram F, von der Lehr N, Cetinkaya C, Larsson LG (2000) c-Myc hot spot mutations in lymphomas result in inefficient ubiquitination and decreased proteasome-mediated turnover. Blood J Am Soc Hematol 95(6): 2104–2110
66. Symonds G, Hartshorn A, Kennewell A, O'Mara MA, Bruskin A, Bishop JM (1989) Transformation of murine myelomonocytic cells by myc: point mutations in v-myc contribute synergistically to transforming potential. Oncogene 4(3):285–294
67. Chakraborty AA, Scuoppo C, Dey S, Thomas LR, Lorey SL, Lowe SW, Tansey WP (2015) A common functional consequence of tumor-derived mutations within c-MYC. Oncogene 34(18):2406–2409
68. Nair SK, Burley SK (2006) Structural aspects of interactions within the Myc/Max/Mad network. In: The Myc/Max/Mad Transcription Factor Network. Springer, Berlin, Heidelberg, pp 123–143
69. Channavajhala P, Seldin DC (2002) Functional interaction of protein kinase CK2 and c-Myc in lymphomagenesis. Oncogene 21(34):5280–5288
70. Hann, S. R. (2006). Role of post-translational modifications in regulating c-Myc proteolysis, transcriptional activity and biological function. In Seminars in cancer biology(Vol. 16, No. 4, pp. 288-302). Academic Press
71. Vervoorts J, Lüscher-Firzlaff J, Lüscher B (2006) The ins and outs of MYC regulation by posttranslational mechanisms. J Biol Chem 281(46):34725–34729
72. Sears RC (2004) The life cycle of C-myc: from synthesis to degradation. Cell Cycle 3(9): 1131–1135
73. Benassi B, Fanciulli M, Fiorentino F, Porrello A, Chiorino G, Loda M et al (2006) c-Myc phosphorylation is required for cellular response to oxidative stress. Mol Cell 21(4): 509–519
74. Henriksson M, Bakardjiev A, Klein G, Lüscher B (1993) Phosphorylation sites mapping in the N-terminal domain of c-myc modulate its transforming potential. Oncogene 8(12): 3199–3209
75. Lutterbach B, Hann SR (1994) Hierarchical phosphorylation at N-terminal transformation-sensitive sites in c-Myc protein is regulated by mitogens and in mitosis. Mol Cell Biol 14(8):5510–5522
76. Lutterbach B, Hann SR (1999) c-Myc transactivation domain-associated kinases: Questionable role for map kinases in c-Myc phosphorylation. J Cell Biochem 72(4): 483–491
77. Noguchi K, Kitanaka C, Yamana H, Kokubu A, Mochizuki T, Kuchino Y (1999) Regulation of c-Myc through phosphorylation at Ser-62 and Ser-71 by c-Jun N-terminal kinase. J Biol Chem 274(46):32580–32587
78. Sears R, Leone G, DeGregori J, Nevins JR (1999) Ras enhances Myc protein stability. Mol Cell 3(2):169–179
79. Sears R, Nuckolls F, Haura E, Taya Y, Tamai K, Nevins JR (2000) Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. Genes Dev 14(19):2501–2514
80. Kim SY, Herbst A, Tworkowski KA, Salghetti SE, Tansey WP (2003) Skp2 regulates Myc protein stability and activity. Mol Cell 11(5): 1177–1188
81. Von Der Lehr N, Johansson S, Wu S, Bahram F, Castell A, Cetinkaya C et al (2003) The F-box protein Skp2 participates in c-Myc proteosomal degradation and acts as a cofactor for c-Myc-regulated transcription. Mol Cell 11(5):1189–1200
82. Molinari E, Gilman M, Natesan S (1999) Proteasome-mediated degradation of transcriptional activators correlates with activation domain potency in vivo. EMBO J 18(22): 6439–6447
83. Gstaiger M, Jordan R, Lim M, Catzavelos C, Mestan J, Slingerland J, Krek W (2001) Skp2 is oncogenic and overexpressed in human cancers. Proc Natl Acad Sci 98(9):5043–5048
84. Zeng L, Zhou MM (2002) Bromodomain: an acetyl-lysine binding domain. FEBS Lett 513(1):124–128
85. Vervoorts J, Lüscher-Firzlaff JM, Rottmann S, Lilischkis R, Walsemann G, Dohmann K et al (2003) Stimulation of c-MYC transcriptional activity and acetylation

by recruitment of the cofactor CBP. EMBO Rep 4(5):484–490

86. Patel JH, Du Y, Ard PG, Phillips C, Carella B, Chen CJ et al (2004) The c-MYC oncoprotein is a substrate of the acetyltransferases hGCN5/PCAF and TIP60. Mol Cell Biol 24(24):10826–10834
87. Faiola F, Liu X, Lo S, Pan S, Zhang K, Lymar E et al (2005) Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription. Mol Cell Biol 25(23):10220–10234
88. Piccolo S, Dupont S, Cordenonsi M (2014) The biology of YAP/TAZ: hippo signaling and beyond. Physiol Rev 94(4):1287–1312
89. Pan D (2010) The hippo signaling pathway in development and cancer. Dev Cell 19(4): 491–505
90. Meng Z, Moroishi T, Guan KL (2016) Mechanisms of Hippo pathway regulation. Genes Dev 30(1):1–17
91. Gill MK, Christova T, Zhang YY, Gregorieff A, Zhang L, Narimatsu M et al A feed forward loop enforces YAP/TAZ signaling during tumorigenesis. Nat Commun 9(1):1–13
92. Johnson R, Halder G (2014) The two faces of Hippo: targeting the Hippo pathway for regenerative medicine and cancer treatment. Nat Rev Drug Discov 13(1):63–79
93. Cordenonsi M, Zanconato F, Azzolin L, Forcato M, Rosato A, Frasson C et al (2011) The Hippo transducer TAZ confers cancer stem cell-related traits on breast cancer cells. Cell 147(4):759–772
94. Liu JY, Li YH, Lin HX, Liao YJ, Mai SJ, Liu ZW et al (2013) Overexpression of YAP 1 contributes to progressive features and poor prognosis of human urothelial carcinoma of the bladder. BMC Cancer 13(1):349
95. Zender L, Spector MS, Xue W, Flemming P, Cordon-Cardo C, Silke J et al (2006) Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. Cell 125(7):1253–1267
96. Schlegelmilch K, Mohseni M, Kirak O, Pruszak J, Rodriguez JR, Zhou D et al (2011) Yap1 acts downstream of α-catenin to control epidermal proliferation. Cell 144(5):782–795
97. St John MA, Tao W, Fei X, Fukumoto R, Carcangiu ML, Brownstein DG et al (1999) Mice deficient of Lats1 develop soft-tissue sarcomas, ovarian tumours and pituitary dysfunction. Nat Genet 21(2):182–186
98. Wang Y, Dong Q, Zhang Q, Li Z, Wang E, Qiu X (2010) Overexpression of yes-associated protein contributes to progression and poor prognosis of non-small-cell lung cancer. Cancer Sci 101(5):1279–1285
99. Strnadel J, Choi S, Fujimura K, Wang H, Zhang W, Wyse M et al (2017) eIF5A-PEAK1 signaling regulates YAP1/TAZ protein expression and pancreatic cancer cell growth. Cancer Res 77(8):1997–2007
100. Felley-Bosco E, Stahel R (2014) Hippo/YAP pathway for targeted therapy. Translat Lung Cancer Res 3(2):75
101. Chan P, Han X, Zheng B, DeRan M, Yu J, Jarugumilli GK et al (2016) Autopalmitoylation of TEAD proteins regulates transcriptional output of the Hippo pathway. Nat Chem Biol 12(4):282
102. Liu X, Li H, Rajurkar M, Li Q, Cotton JL, Ou J et al (2016) Tead and AP1 coordinate transcription and motility. Cell Rep 14(5): 1169–1180
103. Kim MK, Jang JW, Bae SC (2018) DNA binding partners of YAP/TAZ. BMB Rep 51(3):126
104. Ji J, Xu R, Zhang X, Han M, Xu Y, Wei Y et al (2018) Actin like-6A promotes glioma progression through stabilization of transcriptional regulators YAP/TAZ. Cell Death Dis 9(5):1–16
105. Zhang Z, Du J, Wang S, Shao L, Jin K, Li F et al (2019) OTUB2 promotes cancer metastasis via hippo-independent activation of YAP and TAZ. Mol Cell 73(1):7–21
106. Yao F, Zhou Z, Kim J, Hang Q, Xiao Z, Ton BN et al (2018) SKP2-and OTUD1-regulated non-proteolytic ubiquitination of YAP promotes YAP nuclear localization and activity. Nat Commun 9(1):1–16
107. Polakis P (2000) Wnt signaling and cancer. Genes Dev 14(15):1837–1851
108. Miyoshi K, Hennighausen L (2003) β-Catenin: a transforming actor on many stages. Breast Cancer Research. 5(2):63
109. Chan E, Gat U, McNiff JM, Fuchs E (1999) A common human skin tumour is caused by activating mutations in β-catenin. Nat Genet 21(4):410–413
110. Wan X, Liu J, Lu JF, Tzelepi V, Yang J, Starbuck MW et al (2012) Activation of β-catenin signaling in androgen receptor–negative prostate Cancer cells. Clin Cancer Res 18(3): 726–736
111. Khurana N, Sikka SC (2019) Interplay between SOX9, Wnt/β-catenin and androgen receptor signaling in castration-resistant prostate cancer. Int J Mol Sci 20(9):2066
112. Aberle H, Bauer A, Stappert J, Kispert A, Kemler R (1997) β-catenin is a target for the

ubiquitin–proteasome pathway. EMBO J 16(13):3797–3804

113. Tetsu O, McCormick F (1999) β-Catenin regulates expression of cyclin D1 in colon carcinoma cells. Nature 398(6726):422–426
114. Nussinov R, Tsai CJ, Jang H, Korcsmáros T, Csermely P (2016) Oncogenic KRAS signaling and YAP1/β-catenin: Similar cell cycle control in tumor initiation. In: Seminars in cell & developmental biology. Academic Press, p 85
115. Deng F, Peng L, Li Z, Tan G, Liang E, Chen S et al (2018) YAP triggers the Wnt/β-catenin signalling pathway and promotes enterocyte self-renewal, regeneration and tumorigenesis after DSS-induced injury. Cell Death Dis 9(2):1–16
116. Liu H, Du S, Lei T, Wang H, He X, Tong R, Wang Y (2018) Multifaceted regulation and functions of YAP/TAZ in tumors. Oncol Rep 40(1):16–28
117. Li Q, Sun Y, Jarugumilli GK, Liu S, Dang K, Cotton JL et al (2020) Lats1/2 Sustain Intestinal Stem Cells and Wnt Activation through TEAD-Dependent and Independent Transcription. Cell Stem Cell 26(5):675–692
118. Sun J, Wang X, Tang B, Liu H, Zhang M, Wang Y et al (2018) A tightly controlled Src-YAP signaling axis determines therapeutic response to dasatinib in renal cell carcinoma. Theranostics 8(12):3256
119. Moroishi T, Hansen CG, Guan KL (2015) The emerging roles of YAP and TAZ in cancer. Nat Rev Cancer 15(2):73–79
120. Sorrentino G, Ruggeri N, Specchia V, Cordenonsi M, Mano M, Dupont S et al (2014) Metabolic control of YAP and TAZ by the mevalonate pathway. Nat Cell Biol 16(4):357–366
121. Wang Z, Wu Y, Wang H, Zhang Y, Mei L, Fang X et al (2014) Interplay of mevalonate and Hippo pathways regulates RHAMM transcription via YAP to modulate breast cancer cell motility. Proc Natl Acad Sci 111(1):E89–E98
122. Guo J, Wu Y, Yang L, Du J, Gong K, Chen W et al (2017) Repression of YAP by NCTD disrupts NSCLC progression. Oncotarget 8(2):2307
123. Bao Y, Nakagawa K, Yang Z, Ikeda M, Withanage K, Ishigami-Yuasa M et al (2011) A cell-based assay to screen stimulators of the Hippo pathway reveals the inhibitory effect of dobutamine on the YAP-dependent gene transcription. J Biochemi 150(2):199–208
124. Zheng HX, Wu LN, Xiao H, Du Q, Liang JF (2014) Inhibitory effects of dobutamine on human gastric adenocarcinoma. World J Gastroenterol: WJG 20(45):17092
125. Lee SB, Gong YD, Park YI, Dong MS (2013) 2, 3, 6-Trisubstituted quinoxaline derivative, a small molecule inhibitor of the Wnt/beta-catenin signaling pathway, suppresses cell proliferation and enhances radiosensitivity in A549/Wnt2 cells. Biochem Biophys Res Commun 431(4):746–752
126. Lee SB, Park YI, Dong MS, Gong YD (2010) Identification of 2, 3, 6-trisubstituted quinoxaline derivatives as a Wnt2/β-catenin pathway inhibitor in non-small-cell lung cancer cell lines. Bioorg Med Chem Lett 20(19): 5900–5904
127. Park S, Choi J (2010) Inhibition of β-catenin/Tcf signaling by flavonoids. J Cell Biochem 110(6):1376–1385
128. Amado NG, Predes D, Moreno MM, Carvalho IO, Mendes FA, Abreu JG (2014) Flavonoids and Wnt/β-catenin signaling: potential role in colorectal cancer therapies. Int J Mol Sci 15(7):12094–12106
129. Li X, Bai B, Liu L, Ma P, Kong L, Yan J et al (2015) Novel β-carbolines against colorectal cancer cell growth via inhibition of Wnt/β-catenin signaling. Cell Death Dis 1(1):1–9
130. Kong L, Mao B, Zhu H, Li Y (2015) Novel β-carbolines inhibit Wnt/β-catenin signaling.
131. Castell A, Yan Q, Fawkner K, Hydbring P, Zhang F, Verschut V et al (2018) A selective high affinity MYC-binding compound inhibits MYC: MAX interaction and MYC-dependent tumor cell proliferation. Sci Rep 8(1):1–17
132. Yap JL, Wang H, Hu A, Chauhan J, Jung KY, Gharavi RB et al (2013) Pharmacophore identification of c-Myc inhibitor 10074-G5. Bioorg Med Chem Lett 23(1):370–374
133. Chauhan J, Wang H, Yap JL, Sabato PE, Hu A, Prochownik EV, Fletcher S (2014) Discovery of Methyl 4′-Methyl-5-(7-nitrobenzo [c][1, 2, 5] oxadiazol-4-yl)-[1, 1′-biphenyl]-3-carboxylate, an Improved Small-Molecule Inhibitor of c-Myc–Max Dimerization. ChemMedChem 9(10): 2274–2285
134. Hart JR, Garner AL, Yu J, Ito Y, Sun M, Ueno L et al (2014) Inhibitor of MYC identified in a Kröhnke pyridine library. Proc Natl Acad Sci 111(34):12556–12561
135. Jeong KC, Kim KT, Seo HH, Shin SP, Ahn KO, Ji MJ et al (2014) Intravesical instillation of c-MYC inhibitor KSI-3716 suppresses orthotopic bladder tumor growth. J Urol 191(2):510–518
136. Choi SH, Mahankali M, Lee SJ, Hull M, Petrassi HM, Chatterjee AK et al (2017)

Targeted disruption of Myc–Max oncoprotein complex by a small molecule. ACS Chem Biol 12(11):2715–2719

137. Jiao S, Wang H, Shi Z, Dong A, Zhang W, Song X et al (2014) A peptide mimicking VGLL4 function acts as a YAP antagonist therapy against gastric cancer. Cancer Cell 25(2):166–180
138. Liu-Chittenden Y, Huang B, Shim JS, Chen Q, Lee SJ, Anders RA et al (2012) Genetic and pharmacological disruption of the TEAD–YAP complex suppresses the oncogenic activity of YAP. Genes Dev 26(12):1300–1305
139. Tschaharganeh DF, Chen X, Latzko P, Malz M, Gaida MM, Felix K et al (2013) Yes-associated protein up-regulates Jagged-1 and activates the Notch pathway in human hepatocellular carcinoma. Gastroenterology 144(7):1530–1542
140. Brodowska K, Al-Moujahed A, Marmalidou A, Zu Horste MM, Cichy J, Miller JW, Vavvas DG (2014) The clinically used photosensitizer Verteporfin (VP) inhibits YAP-TEAD and human retinoblastoma cell growth in vitro without light activation. Exp Eye Res 124:67–73
141. Lin L, Sabnis AJ, Chan E, Olivas V, Cade L, Pazarentzos E et al (2015) The Hippo effector YAP promotes resistance to RAF-and MEK-targeted cancer therapies. Nat Genet 47(3):250–256
142. Kim MH, Kim J, Hong H, Lee SH, Lee JK, Jung E, Kim J (2016) Actin remodeling confers BRAF inhibitor resistance to melanoma cells through YAP/TAZ activation. EMBO J 35(5):462–478
143. Fisher ML, Grun D, Adhikary G, Xu W, Eckert RL (2017) Inhibition of YAP function overcomes BRAF inhibitor resistance in melanoma cancer stem cells. Oncotarget 8(66): 110257
144. Emami KH, Nguyen C, Ma H, Kim DH, Jeong KW, Eguchi M et al (2004) A small molecule inhibitor of β-catenin/cyclic AMP response element-binding protein transcription. Proc Natl Acad Sci 101(34): 12682–12687
145. Manegold P, Lai KK, Wu Y, Teo JL, Lenz HJ, Genyk YS et al (2018) Differentiation therapy targeting the β-catenin/CBP interaction in pancreatic cancer. Cancer 10(4):95
146. Dietrich L, Rathmer B, Ewan K, Bange T, Heinrichs S, Dale TC et al (2017) Cell permeable stapled peptide inhibitor of Wnt signaling that targets β-catenin protein-protein interactions. Cell Chem Biol 24(8):958–968
147. Gonsalves FC, Klein K, Carson BB, Katz S, Ekas LA, Evans S et al (2011) An RNAi-based chemical genetic screen identifies three small-molecule inhibitors of the Wnt/wingless signaling pathway. Proc Natl Acad Sci 108(15): 5954–5963
148. Sukhdeo K, Mani M, Zhang Y, Dutta J, Yasui H, Rooney MD et al (2007) Targeting the β-catenin/TCF transcriptional complex in the treatment of multiple myeloma. Proc Natl Acad Sci 104(18):7516–7521
149. Li X, Pu J, Jiang S, Su J, Kong L, Mao B et al (2013) Henryin, an ent-kaurane diterpenoid, inhibits Wnt signaling through interference with β-catenin/TCF4 interaction in colorectal cancer cells. PLoS One 8(7): 1–10
150. Wang Z, Zhang M, Wang J, Ji H (2019) Optimization of Peptidomimetics as Selective Inhibitors for the β-Catenin/T-Cell Factor Protein–Protein Interaction. J Med Chem 62(7):3617–3635
151. Simon RJ, Kania RS, Zuckermann RN, Huebner VD, Jewell DA, Banville S et al (1992) Peptoids: a modular approach to drug discovery. Proc Natl Acad Sci 89(20): 9367–9371
152. Schneider JA, Craven TW, Kasper AC, Yun C, Haugbro M, Briggs EM et al (2018) Design of Peptoid-peptide Macrocycles to Inhibit the β-catenin TCF Interaction in Prostate Cancer. Nat Commun 9(1):1–10
153. Wang H, Teriete P, Hu A, Raveendra-Panickar D, Pendelton K, Lazo JS et al (2015) Direct inhibition of c-Myc-Max heterodimers by celastrol and celastrol-inspired triterpenoids. Oncotarget 6(32):32380
154. Jung KY, Wang H, Teriete P, Yap JL, Chen L, Lanning ME et al (2015) Perturbation of the c-Myc–max protein–protein interaction via synthetic α-helix mimetics. J Med Chem 58(7):3002–3024
155. Lu JJ, Meng LH, Shankavaram UT, Zhu CH, Tong LJ, Chen G et al (2010) Dihydroartemisinin accelerates c-MYC oncoprotein degradation and induces apoptosis in c-MYC-overexpressing tumor cells. Biochem Pharmacol 80(1):22–30
156. Waaler J, Machon O, Tumova L, Dinh H, Korinek V, Wilson SR et al (2012) A novel tankyrase inhibitor decreases canonical Wnt signaling in colon carcinoma cells and reduces tumor growth in conditional APC mutant mice. Cancer Res 72(11):2822–2832
157. Hwang SY, Deng X, Byun S, Lee C, Lee SJ, Suh H et al (2016) Direct targeting of β-catenin by a small molecule stimulates proteasomal degradation and suppresses

oncogenic Wnt/β-catenin signaling. Cell Rep 16(1):28–36

158. Yang W, Li Y, Ai Y, Obianom ON, Guo D, Yang H et al (2019) Pyrazole-4-Carboxamide (YW2065): A Therapeutic Candidate for Colorectal Cancer via Dual Activities of Wnt/β-Catenin Signaling Inhibition and AMP-Activated Protein Kinase (AMPK) Activation. J Med Chem 62(24):11151–11164
159. Balaji KC, Koul H, Mitra S, Maramag C, Reddy P, Menon M et al (1997) Antiproliferative effects of c-myc antisense oligonucleotide in prostate cancer cells: a novel therapy in prostate cancer. Urology 50(6): 1007–1015
160. Iversen PL, Arora V, Acker AJ, Mason DH, Devi GR (2003) Efficacy of antisense morpholino oligomer targeted to c-myc in prostate cancer xenograft murine model and a Phase I safety study in humans. Clin Cancer Res 9(7):2510–2519
161. Sekhon HS, London CA, Sekhon M, Iversen PL, Devi GR (2008) c-MYC antisense phosphosphorodiamidate morpholino oligomer inhibits lung metastasis in a murine tumor model. Lung Cancer 60(3):347–354
162. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S (2006) Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res 34(19):5402–5415
163. Ohnmacht SA, Neidle S (2014) Small-molecule quadruplex-targeted drug discovery. Bioorg Med Chem Lett 24(12): 2602–2612
164. Chatterjee J, Mierke DF, Kessler H (2008) Conformational preference and potential templates of N-methylated cyclic pentaalanine peptides. Chem Eur J 14(5):1508–1517
165. Seenisamy J, Bashyam S, Gokhale V, Vankayalapati H, Sun D, Siddiqui-Jain A et al (2005) Design and synthesis of an expanded porphyrin that has selectivity for the c-MYC G-quadruplex structure. J Am Chem Soc 127(9):2944–2959
166. Calabrese DR, Chen X, Leon EC, Gaikwad SM, Phyo Z, Hewitt WM et al (2018) Chemical and structural studies provide a mechanistic basis for recognition of the MYC G-quadruplex. Nat Commun 9(1):1–15
167. Hu MH, Wang YQ, Yu ZY, Hu LN, Ou TM, Chen SB et al (2018) Discovery of a new four-leaf clover-like ligand as a potent c-MYC transcription inhibitor specifically targeting the promoter G-quadruplex. J Med Chem 61(6):2447–2459
168. Ganesh S, Shui X, Craig KP, Park J, Wang W, Brown BD, Abrams MT (2018) RNAi-mediated β-catenin inhibition promotes T cell infiltration and antitumor activity in combination with immune checkpoint blockade. Mol Ther 26(11):2567–2579

Chapter 10

A Survey of Transcription Factors in Cell Fate Control

Emal Lesha, Haydy George, Mark M. Zaki, Cory J. Smith, Parastoo Khoshakhlagh, and Alex H. M. Ng

Abstract

Transcription factors (TFs) play a cardinal role in the development and maintenance of human physiology by acting as mediators of gene expression and cell state control. Recent advancements have broadened our knowledge on the potency of TFs in governing cell physiology and have deepened our understanding of the mechanisms through which they exert this control. The ability of TFs to program cell fates has gathered significant interest in recent decades, and high-throughput technologies now allow for the systematic discovery of forward programming factors to convert pluripotent stem cells into numerous differentiated cell types. The next generation of these technologies has the potential to improve our understanding and control of cell fates and states and provide advanced therapeutic modalities to address many medical conditions.

Key words Transcription Factor (TF), Reprogramming, Engineering, iPSC, Cell therapy

1 Transcription Factors: Mediators of Gene Expression

Transcription factors (TFs) play an indispensable role in the evolution and development of human biology. Their involvement in organ development, cell physiology, and pathologic disease makes the understanding of TFs important for the advancement of biomedical research. TFs are defined as proteins that bind to DNA regulatory sequences to modulate gene transcription. Specifically, TFs can bind distal to or at the promoter region of a DNA sequence to alter gene expression through transcriptional control. Although most TF binding sites are usually small at 6–12 bases [1, 2], the distance between a TF binding site and the transcription start site of the gene regulated by a specific TF can include regions as large as several megabases; as such TFs can also alter the structural organization of chromatin [2, 3]. Based on the modulating effect they can

Emal Lesha, Haydy George and Mark M. Zaki contributed equally with all other contributors.

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_10,

have, TFs can be categorized as enhancing or silencing, thus aiding in upregulation or downregulation in the expression of target genes.

The interactions between a TF and its binding site are more complex than once thought, and our understanding of such interactions is evolving with continued research. TFs can directly bind DNA via a specific set of DNA sequences called binding motifs and can recognize multiple binding motifs. Additionally, TFs have different binding affinities for their binding motifs. The differential usage of multiple DNA binding domains within the same TF, as well as multiple docking conformations between TF and DNA, could explain such complex TF–DNA interactions [4]. TFs can also interact or compete with each other. For example, binding of a TF to an inactive region of chromatin can cause a conformational change which would then allow another TF to bind to the DNA sequence of interest [5]. Competition between TFs for binding sites and vice versa has been described in the literature [6, 7]; this mechanism potentially helps in regulating the activity and abundance of TFs and has been elucidated to play a role in cancer development [8]. Lastly, recent studies have shown that eukaryotic TFs can actively track their target genes and control site occupancy [9].

Multiple computational models have been developed that can assist in mapping and characterizing TF binding motifs. The positional weight matrix (PWM), for example, is a well-described model that predicts the binding affinity of a TF to a specific DNA sequence and thus can help predict binding domains for that TF. The ability to find TF binding motifs has also evolved in the last decade through in vivo techniques such as ChIP-seq or Dnase-seq and in vitro techniques such as microfluidics-based mechanically induced trapping of molecular interactions (MITOMI) and in vitro selection-based approaches (SELEX) [4]. In vivo techniques are helpful in determining the binding of TFs to DNA at particular cellular or treatment conditions, while in vitro techniques are more helpful for large-scale characterization of TF binding preferences[4].

Comprehensive databases of TFs and TF binding profiles have been developed in recent years. These resources provide invaluable opportunities for further advancement of research on TFs and their role in cell fate. The Animal Transcription Factor DataBase, for example, is a resource that contains a catalog of >125,000 TF genes from 97 animal genomes [10]. In regard to human TFs, the most comprehensive analysis published to date compiled a set of 1639 TFs [2]. Multiple databases exist that compile information on TF binding sites, TF binding profiles, PWMs, and more. JASPAR and TRANSFAC include some of the more commonly used platforms, and new databases have been developed recently focusing on human TF profiles [11, 12]. A classification system for

human TFs has been described that classifies 1558 TFs according to their DNA binding domains [13]. Based on this classification, the human TF repertoire is categorized into ten different superclasses of DNA binding domains. The majority of the TFs fall into three of the superclasses, specifically the zinc finger, helix-turn-helix, and basic domains. This classification system has also been extended to mouse and other mammalian TF orthologs [14, 15].

2 Physiological Function of Transcription Factors and Clinical Significance

Our understanding of the role and function of TFs has accelerated in the past decade, enabled by the development of the aforementioned technologies and databases. Studies of tissue-specific expression of TFs across the different human tissues and organ systems have helped characterize the role and function of each TF and their relation to human physiology and pathological disorders. Human Cell Atlas projects, using single cell transcriptomics, are significantly advancing these efforts [16]. Physiologically, TFs play a major role in the regulation of embryological development. While only a small number of TFs dictate the first stages of embryogenesis, during later stages a dynamic shift occurs that leads to the overexpression of thousands of genes, also referred to as zygotic genome activation (ZGA) [17]. Recent studies have shown DUX4, for instance, to be a key regulator of ZGA in mammals [18].

As TFs underlie major developmental pathways of organ systems, mutations of these TFs have also been linked to congenital disease. For instance, NKX2–5 is a TF involved in the development of the cardiac system, including proliferation of cardiac precursor cells and formation of the cardiac outflow tract. Mutations in NKX2–5 have been detected in patients with congenital cardiac disorders such as Tetralogy of Fallot, Transposition of the Great Arteries, and Patent Ductus Arteriosus [19]. Mutations in AFF4 lead to altered transcriptional elongation and are linked with patients developing cognitive impairment, heart defects, obesity, pulmonary deficits, and short stature, also known as CHOPS syndrome [20, 21]. Disruptions in other transcription factors may have more localized effects as well. For example, a frameshift mutation in GRHL2/TFCP2L3 has been associated with non-syndromic autosomal dominant sensorineural hearing loss [22]. Overall, transcription factors have essential roles in gene expression that when impaired can lead to major localized or systemic disease.

TFs play a critical role in cell response pathways, including activation and regulation of immune response in the body, which have pathologic implications. The Interferon Regulatory Factor (IRF) and Signal Transducer and Activator of Transcription

(STAT) families of TFs, for instance, are involved in transcriptional regulation of interferons, proteins that regulate inflammatory activation and are implicated in the pathogenesis of autoimmune disease. Mutations of these TFs have been linked to increased susceptibility to severe fungal and viral infections [23]. NF-κB is a ubiquitous TF that plays a general role in the regulation of inflammatory genes. Upregulation of NF-κB has been shown in many autoinflammatory conditions, including asthma, inflammatory bowel disease, and atherosclerosis [24]. Further, mutations in the transcription factor AIRE lead to impaired immune tolerance, causing autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED) [25].

Finally, TFs control and regulate cell fate and differentiation. A landmark study showed that through the combination of TFs Oct4, Sox2, Klf4, and c-Myc, it was possible to activate pluripotency in fibroblast cells, thereby reprogramming them into induced pluripotent stem cells (iPSCs) [26]. Furthermore, through the combination of specific TFs, fibroblasts have been trans-differentiated into cells of interest, including myoblasts and neurons [27, 28]. TFs have also been over-expressed directly in iPSCs for forward programming into desired cell types, such as neurons, oligodendrocytes, and more [29, 30]. This aspect of TF function is of major interest in current biomedical research. Discovery of such TF combinations has contributed to the knowledge of transcriptional pathways of tissue development and cell regeneration and has applications for producing cell types for disease modeling, drug discovery, and cell therapies.

3 Clinical Applications of Transcription Factor-Based Programming

The ability to program cell identity, by using reprogramming technology to convert somatic cells into iPSCs and especially forward programming of iPSCs to cell types of interest through the use of TFs, is a particularly exciting aspect of TFs with clinical implications. This technology has the potential to revolutionize cell therapy and change the landscape of medical management for many diseases. In fact, several former and ongoing clinical studies have utilized these technologies for treating various conditions. For example, autologous iPSC-derived dopaminergic neurons have successfully been implanted into a patient with Parkinson's disease, who showed improvement in the clinical endpoints of the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and 39-item Parkinson's Disease Questionnaire (PDQ-39) at 2-year follow-up [31]. Another study assessed the use of iPSC-derived retinal pigment epithelial (RPE) cell injection for a patient with macular degeneration [32]. This study showed not only safety but also stable best corrected visual acuity and

improved visual function and quality of life as assessed by the National Eye Institute Visual Functioning Questionnaire (VFQ)-25 at 1-year follow-up [32].

Early clinical experience with iPSC-derived cells has supported further investigation of iPSC-derived cell therapies in clinical trials. As of 2022, the National Eye Institute is recruiting for a Phase 1/2 study to assess the effect of autologous iPSC-derived RPEs in patients with dry age-related macular degeneration (NCT04339764). Another clinical trial sponsored by Osaka University in Japan is recruiting patients for a Phase 1 study to investigate allogeneic iPSC-derived cardiomyocytes in patients with ischemic cardiomyopathy (NCT04696328). The first-in-human clinical trial assessing iPSC-derived neural stem/progenitor cells in complete spinal cord injury will also soon be underway in Japan [33]. TFs are at the forefront in reprogramming cells to iPSCs as used in these trials, but current work to differentiate iPSCs into target cell types uses directed differentiation approaches. TF-based forward programming of iPSCs into differentiated cell types is an emerging modality for cell-based therapies [30].

Many other cell types are being derived from iPSCs and may soon be in clinical studies. This is particularly exciting in disciplines where cells have limited potential to regenerate, such as neurodegenerative diseases. Specific neurological cell types have been produced and include but are not limited to sensory hair cells, astrocytes, and GABAergic neurons. These cells may improve sensory deficits, stabilize neuronal cell and blood–brain barrier function, and enhance inhibitory drive of neuronal circuits, respectively.

4 Current Technologies of Transcription Factor Programming and Future Directions

TF-based programming is of major interest in current scientific research; thus the need and potential to control and manipulate the expression level of TFs, particularly to reprogram cell fate toward pluripotency and produce cells of interest from pluripotent stem cells, have led to the development of multiple technologies in genetic engineering, computational biology, and cell therapy. These advances, which will be discussed here, have made possible the tackling of various aspects of TF programming both in vitro and in vivo.

The first report of reprogramming of somatic cells to pluripotent stem cells was published in 2006, where murine fibroblasts were reprogrammed into iPSCs using the combination of Oct4, Sox2, Klf4, and c-Myc TFs [26]. The protocol used retroviral transduction and required cell culture of up to 21 days for the detection of reprogrammed cells and resulted in reprogramming of only 0.02% of the cultured fibroblast cells [26]. Since this breakthrough, multiple studies have attempted to improve the efficacy

and process required to develop iPSCs in vitro. Recent published work has reported efficiencies of reprogramming of human primary fibroblasts into iPSCs with efficiencies as high as 90% [34]. The source of somatic cells used for reprogramming into iPSCs has expanded beyond dermal fibroblasts to a plethora of cell types; for instance, a recent publication described reprogramming of long-term hematopoietic stem cells into iPSCs with efficiency of as high as 50% [35].

Differentiation of iPSCs into target differentiated cells of interest has the potential to innovate therapeutics and tackle medical conditions that remain without cure. Directed differentiation has been the traditional approach to produce differentiated cells from pluripotent stem cells. By altering the culturing conditions in a stepwise manner by mimicking natural developmental pathways, pluripotent stem cells can be coaxed into the desired differentiated cell type of interest. However, this approach typically requires many steps, has lengthy timelines, gives low differentiation efficiency of the target cell type with high off-target production of undesired cell types, and has high variability when applied to different iPSC lines. In contrast, TF-based forward programming has overcome many of these concerns; however, one of the main areas of explorations is discovering the combination of TFs needed to program iPSCs into a target cell type. The few combinations of TFs known to derive specific cell types were obtained mainly via trial and error experiments using prior knowledge, with only a small percentage of the total human TF repertoire known to have the ability to induce differentiation in vitro. Recent work published from our group addresses this issue by building the TFome™ technology platform for unbiased discovery of TFs for cell fate programming. The TFome™ includes an open reading frame (ORF) library containing >1700 human TFs at splice-isoform resolution and the ability to express them in hiPSCs to perform the first genome-scale screen to identify TFs to induce forward programming in vitro in 4 days without changes to the matrix or media in a single step [30]. The TFome™ approach identified 290 TFs that induced differentiation, 241 of which had never been described in the literature for this purpose. As the screen was cell type-agnostic by using loss of pluripotency as a readout instead of a particular cell type marker, the top hits were validated individually, and the cell type was determined by transcriptomic profiling. ATOH1 programmed iPSCs into neurons, NKX3-1 into fibroblasts, ETV2 into vascular endothelial cells, and SOX9 into oligodendrocytes. Given the cell autonomous nature of this forward programming approach, where TFs alone induced programming and alterations to the media or matrix are not required, we successfully produced up to three distinct cell types in the same culture and in a common media in what was termed “parallel programming.” Furthermore, we showed that iPSCs engineered with TFs could produce the desired

cell types even in directed differentiation conditions used to produce organoid tissues of another lineage, thereby demonstrating the cell autonomous nature of TFome™-based programming in what we call "orthogonal programming." We accelerated the production of oligodendrocytes and therefore myelin sheaths within cerebral organoids [30]. The TFome™ platform is being expanded to incorporate combinatorial studies and genome-scale readouts, which should yield a vast number of novel recipes to produce desired cell types and expand the range of medical conditions that can be tackled via cell therapy. Importantly, inducing forward programming at a faster rate without the need for complex culture conditions can lead to efficient cell production, making cell therapies more easily manufactured at scale and more accessible to patients.

Advancements in computational biology, synthetic biology, and genetic engineering, among other disciplines, continue to have a major impact in advancing TF-based cell therapy research. For instance, synthetic zinc-finger TFs can aid in iPSC reprogramming [36] and warrant further studies into whether this unique approach can be generalized to the production of differentiated cell types. CRISPR-Cas-based library screening has been described to screen for cell fate inducing TFs by utilizing a library of guide RNAs designed based on known TFs [37]. Limitations of this emerging approach, such as Cas proteins being non-human in origin and therefore immunogenic, CRISPR activators generally having lower potency at inducing gene expression and cell conversion compared to direct over-expression, and the need to identify and optimize guide RNAs for each target gene, may eventually be overcome. In summary, advancements in the discovery and characterization of TFs have major implications in our understanding of developmental biology, synthetic biology, and disease pathophysiology, leading to the development of technologies for cell fate engineering that have the potential to provide advanced therapeutic modalities for tackling congenital and pathologic disease.

References

1. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML (2009) Diversity and complexity in DNA recognition by transcription factors. Science 3245935: 1720–1723
2. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT (2018) The human transcription factors. Cell 1724:650–665
3. Dekker J, Heard E (2015) Structural and functional diversity of topologically associating domains. FEBS Lett 58920(Pt A):2877–2884
4. Inukai S, Kock KH, Bulyk ML (2017) Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev 43:110–119
5. Boeva V (2016) Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front Genet 7:24
6. Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R (2014) The

transcription factor titration effect dictates level of gene expression. Cell 1566:1312–1323

7. Darieva Z, Clancy A, Bulmer R, Williams E, Pic-Taylor A, Morgan BA, Sharrocks AD (2010) A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression. Mol Cell 381:29–40
8. Karreth FA, Tay Y, Pandolfi PP (2014) Target competition: transcription factors enter the limelight. Genome Biol 154:114
9. Castellanos M, Mothi N, Munoz V (2020) Eukaryotic transcription factors can track and control their target genes using DNA antennas. Nat Commun 111:540
10. Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res 47:D33–DD8
11. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res 45:D61–DD7
12. Zhang Q, Liu W, Zhang HM, Xie GY, Miao YR, Xia M, Guo AY (2020) hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. Genomics Proteomics Bioinformatics 182:120–128
13. Wingender E, Schoeps T, Donitz J (2013) TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res 41(Database issue):D165
14. Wingender E, Schoeps T, Haubrock M, Donitz J (2015) TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic Acids Res 43:D97–D102
15. Wingender E, Schoeps T, Haubrock M, Krull M, Donitz J (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. Nucleic Acids Res 46(D1):D343–D3D7
16. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, Zhou Y, Ye F, Jiang M, Wu J, Xiao Y, Jia X, Zhang T, Ma X, Zhang Q, Bai X, Lai S, Yu C, Zhu L, Lin R, Gao Y, Wang M, Wu Y, Zhang J, Zhan R, Zhu S, Hu H, Wang C, Chen M, Huang H, Liang T, Chen J, Wang W, Zhang D, Guo G (2020) Construction of a human cell landscape at single-cell level. Nature 5817808:303–309
17. Schulz KN, Harrison MM (2019) Mechanisms regulating zygotic genome activation. Nat Rev Genet 204:221–234
18. De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D (2017) DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nat Genet 496:941–945
19. McCulley DJ, Black BL (2012) Transcription factor pathways and congenital heart disease. Curr Top Dev Biol 100:253–277
20. Izumi K, Nakato R, Zhang Z, Edmondson AC, Noon S, Dulik MC, Rajagopalan R, Venditti CP, Gripp K, Samanich J, Zackai EH, Deardorff MA, Clark D, Allen JL, Dorsett D, Misulovin Z, Komata M, Bando M, Kaur M, Katou Y, Shirahige K, Krantz ID (2015) Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin. Nat Genet 474:338–344
21. Raible SE, Mehta D, Bettale C, Fiordaliso S, Kaur M, Medne L, Rio M, Haan E, White SM, Cusmano-Ozog K, Nishi E, Guo Y, Wu H, Shi X, Zhao Q, Zhang X, Lei Q, Lu A, He X, Okamoto N, Miyake N, Piccione J, Allen J, Matsumoto N, Pipan M, Krantz ID, Izumi K (2019) Clinical and molecular spectrum of CHOPS syndrome. Am J Med Genet A 1797: 1126–1138
22. Peters LM, Anderson DW, Griffith AJ, Grundfast KM, San Agustin TB, Madeo AC, Friedman TB, Morell RJ (2002) Mutation of a transcription factor, TFCP2L3, causes progressive autosomal dominant hearing loss, DFNA28. Hum Mol Genet 1123:2877–2885
23. Mogensen TH (2018) IRF and STAT transcription factors - from basic biology to roles in infection, protective immunity, and primary Immunodeficiencies. Front Immunol 9:3047
24. Lee TI, Young RA (2013) Transcriptional regulation and its misregulation in disease. Cell 1526:1237–1251
25. Abramson J, Giraud M, Benoist C, Mathis D (2010) Aire's partners in the molecular control of immunological tolerance. Cell 1401:123–135
26. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 1264:663–676
27. Davis RL, Weintraub H, Lassar AB (1987) Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell 516:987–1000
28. Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Sudhof TC, Wernig M (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 4637284: 1035–1041
29. Iwafuchi-Doi M, Zaret KS (2014) Pioneer transcription factors in cell reprogramming. Genes Dev 2824:2679–2692

30. Ng AHM, Khoshakhlagh P, Rojo Arias JE, Pasquini G, Wang K, Swiersy A, Shipman SL, Appleton E, Kiaee K, Kohman RE, Vernet A, Dysart M, Leeper K, Saylor W, Huang JY, Graveline A, Taipale J, Hill DE, Vidal M, Melero-Martin JM, Busskamp V, Church GM (2021) A comprehensive library of human transcription factors for cell fate engineering. Nat Biotechnol 394:510–519
31. Schweitzer JS, Song B, Herrington TM, Park TY, Lee N, Ko S, Jeon J, Cha Y, Kim K, Li Q, Henchcliffe C, Kaplitt M, Neff C, Rapalino O, Seo H, Lee IH, Kim J, Kim T, Petsko GA, Ritz J, Cohen BM, Kong SW, Leblanc P, Carter BS, Kim KS (2020) Personalized iPSC-derived dopamine progenitor cells for Parkinson's disease. N Engl J Med 38220:1926–1932
32. Mandai M, Kurimoto Y, Takahashi M (2017) Autologous induced stem-cell-derived retinal cells for macular degeneration. N Engl J Med 3778:792–793
33. Sugai K, Sumida M, Shofuda T, Yamaguchi R, Tamura T, Kohzuki T, Abe T, Shibata R, Kamata Y, Ito S, Okubo T, Tsuji O, Nori S, Nagoshi N, Yamanaka S, Kawamata S, Kanemura Y, Nakamura M, Okano H (2021) First-in-human clinical trial of transplantation of iPSC-derived NS/PCs in subacute complete spinal cord injury: study protocol. Regen Ther 18:321–333
34. Kogut I, McCarthy SM, Pavlova M, Astling DP, Chen X, Jakimenko A, Jones KL, Getahun A, Cambier JC, Pasmooij AMG, Jonkman MF, Roop DR, Bilousova G (2018) High-efficiency RNA-based reprogramming of human primary fibroblasts. Nat Commun 91: 745
35. Wang K, Guzman AK, Yan Z, Zhang S, Hu MY, Hamaneh MB, Yu YK, Tolu S, Zhang J, Kanavy HE, Ye K, Bartholdy B, Bouhassira EE (2019) Ultra-high-frequency reprogramming of individual long-term hematopoietic stem cells yields low somatic variant induced pluripotent stem cells. Cell Rep 2610:2580–2592
36. Eguchi A, Wleklinski MJ, Spurgat MC, Heiderscheit EA, Kropornicka AS, Vu CK, Bhimsaria D, Swanson SA, Stewart R, Ramanathan P, Kamp TJ, Slukvin I, Thomson JA, Dutton JR, Ansari AZ (2016) Reprogramming cell fate with a genome-scale library of artificial transcription factors. Proc Natl Acad Sci U S A 11351:E8257–E8E66
37. Liu S, Striebel J, Pasquini G, Ng AHM, Khoshakhlagh P, Church GM, Busskamp V (2021) Neuronal cell-type engineering by transcriptional activation. Front Genome Ed 3: 715697

Chapter 11

Single-Cell mRNA-Seq of In Vitro-Derived Human Neurons Using Smart-Seq2

Christoph Schweingruber, Jik Nijssen, Julio Aguila Benitez, and Eva Hedlund

Abstract

Single-cell mRNA sequencing can dissect heterogeneous cell populations as it can identify cell types and cellular states based on their unique transcriptional signatures. We use fluorescence-activated cell sorting (FACS) to isolate individual cultured neurons derived from human-induced pluripotent stem cells (hiPSCs) followed by polyA-based Smart-Seq2 RNA sequencing to analyze the single-cell transcriptional profiles. We provide protocols and guidelines on dissociation, cell selection, and library preparation that can be readily adapted to other cell types or tissue samples.

Key words Single-cell RNA sequencing, Smart-Seq2, FACS, Motor neuron

1 Introduction

Single-cell RNA sequencing methods are rapidly developing, because they permit an unprecedented opportunity to dissect heterogeneous cell populations and to probe the transcriptome of a multitude of cells individually and simultaneously. Thus scarce biological materials from in vivo and in vitro sources can be interrogated for specific cell types along with their unique molecular characteristics. The Smart-seq2 protocol remains one of the most sensitive methods for single-cell transcriptomics, albeit at comparatively low throughput, but with the resulting cDNA libraries covering the whole transcript length [1–3]. Another advantage is its versatility, as it can be readily adjusted to different input materials and amounts, such that transcriptomes from single cells of different sizes and even from human post mortem material can be obtained [4–6]. Here we have carefully adjusted the protocol to obtain single-cell transcriptomes of spinal motor neurons derived from human-induced pluripotent stem cells (hiPSCs) that are isolated

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_11,

by fluorescence-activated cell sorting (FACS). We also provide comprehensive guidelines to adapt the protocol to different input materials so that the procedure can be used for other cell populations as well.

2 Materials

Cultured neurons. We routinely derive motor neurons from human-induced pluripotent stem cells in vitro using the protocol detailed here [7].

2.1 Preparations and Cleaning

Absolute ethanol.

70% (v/v) ethanol.

DNA-Off wipes.

RNase Zap wipes.

2.2 Cell Dissociation and FACS Reagents

HBSS buffer with Ca^{2+} and Mg^{2+}.

TrypLE Express cell dissociation reagent.

Neurobasal medium.

B-27 (custom) supplement.

KnockOut Serum Replacement.

Bovine serum albumin (BSA).

Y-27632 dihydrochloride (ROCK inhibitor).

D-Glucose.

Recombinant Human/Murine/Rat Brain-Derived Neurotrophic Factor (BDNF).

Recombinant Human Glial cell-Derived Neurotrophic Factor (GDNF).

TO-PRO-3 iodide.

Neurobasal+B27: Neurobasal medium with 2% v/v B-27 supplement.

Tip blocking solution: Neurobasal medium with 2% (v/v) B-27 supplement and 2% (w/v) BSA.

FACS buffer: HBSS with 2% (w/v) BSA, 2% (v/v) KnockOut Serum Replacement, 2% (v/v) B-27 supplement, 25 mM D-Glucose, and 5 μM Y-27632.

2.3 Oligonucleotides

ERCC RNA Spike-In Control Mixes: Prepare tenfold serial dilutions to 4e-9 from original 10 μL stock.

SMARTer_oligo-dTVN: 5′-/5Biosg/ AAG CAG TGG TAT CAA CGC AGA GTA CTT TTT TTT TTT TTT TTT TTT TTT TTT TVN-3′. Prepare 100 μM stocks and 60 μL aliquots at 10 μM (working concentration). Store them at −20 °C.

SMARTer_TSO-LNA: 5′-/5Biosg/ AAG CAG TGG TAT CAA CGC AGA GTA C rG rG + G − 3′. Prepare 12 μL aliquots at 100 μM. Store them at −80 °C and thaw only once afterward.

SMARTer_ISPCR: 5′-AAG CAG TGG TAT CAA CGC AGA GT-3′. Prepare 100 μM stock, prepare 12 μL aliquots at working concentration, and store at −20 °C.

Nextera Index Primers, Kit v2 Set A, 96 indexes, 384 samples (Illumina): Dilute fivefold with ultrapure water to obtain the working concentration.

Nextera Index Primers, Kit v2 Set D, 96 indexes, 384 samples (Illumina): Dilute fivefold with ultrapure water to obtain the working concentration.

2.4 Smart-Seq2

Due to the high sensitivity of the single-cell library protocol, we recommend the following articles that work in our hands to ensure purity of the preparations.

UltraPure DNase/RNase-free water.

1 M Tris–HCl (pH 8.0).

0.5 M EDTA pH 8.0, RNase-free

2% (v/v) Triton X-100 in ultrapure water.

5 M Betaine.

Recombinant RNase Inhibitor (RRI, Takara/Clonetech).

SuperScript II Reverse Transcriptase (Thermo Fisher Scientific).

KAPA HiFi Hot Start Ready Mix (Roche Diagnostics).

25 mM each deoxynucleotide mix.

1 M magnesium chloride.

24-well Piko PCR Plate Frames (Thermo Fisher Scientific).

24-well Slidetiter Piko PCR plates (Thermo Fisher Scientific).

Adhesive Sealing Sheets (Thermo Fisher Scientific).

PCR tube strip, 8-tube chain (Sarstedt).

PCR lid strip, 8-lid chain (Sarstedt).

Single cell lysis buffer (scLB1): Prepare for each experiment according to Table 1.

First strand synthesis master mix (scSS1.1): Prepare freshly in each library preparation according to Table 2.

Second strand synthesis master mix (scSS2.1): Prepare freshly in each library preparation according to Table 3.

Table 1
Single cell lysis buffer (scLB1) (*see* Note 2)

Component	1	rxn	110	rxn
0.2% triton X-100 + 0.5% RRI	2.02	µL	222.2	µL
10-µM Smarter_oligo-dTVN	0.5	µL	55.0	µL
25-mM dNTP	0.4	µL	44.0	µL
100-mM DTT	0.03	µL	3.3	µL
ERCC spike-ins, 4e-5 dilution	0.05	µL	5.5	µL
Total	3	µL	330	µL

Table 2
First strand synthesis master mix (scSS1.1)

Component	1	rxn	110	rxn
5× SSII buffer	1.0	µL	110.0	µL
5-M betaine	1.0	µL	110.0	µL
100-mM DTT	0.25	µL	27.5	µL
1-M $MgCl_2$	0.035	µL	3.9	µL
200 units/µL SSII RT	0.25	µL	27.5	µL
40 units/µL RRI	0.125	µL	13.8	µL
100-µM Smarter_TSO-LNA	0.1	µL	11.0	µL
Total	2.76	µL	303.6	µL

Table 3
Second strand synthesis master mix (scSS2.1)

Component	1	rxn	110	rxn
Ultrapure water	1.15	µL	126.5	µL
2-µM SMARTer_ISPCR	0.1	µL	11.00	µL
2× KAPA HiFi HS mix	6.25	µL	825.0	µL
Total	7.5	µL	687.5	µL

2.5 Tagmentation

Nextera XT Index Kit, 96 indexes, 384 samples (Illumina).

Phusion HF polymerase (Thermo Fisher Scientific).

25 mM deoxynucleotide mix (dATP, dCTP, dGFT and dTTP, each at a final concentration of 25 mM) (Thermo Fisher Scientific).

0.2% (w/v) sodium dodecyl sulfate (SDS) in ultrapure water.

Dimethyl sulfoxide (DMSO), PCR grade.

2.6 SPRI Bead Purification

80% (v/v) ethanol.

TE (pH 8.0): 10 mM Tris–HCl pH 8.0, 1 mM EDTA, in ultrapure water.

5 M sodium chloride.

10% (v/v) IGEPAL CA630: Dilute from concentrate (Sigma-Aldrich) with ultrapure water (Thermo Fisher Scientific).

10% (w/v) sodium azide.

PEG 8000 flakes.

SPRI bead buffer. Prepare according to Table 4.

Deep Well Plate, V-bottom (Axygen).

Storage Plates, V-bottom (Fisher Scientific).

2.7 Quantification of cDNA and Quality Control

Qubit 1× dsDNA HS assay (Thermo Fisher Scientific).

Qubit Assay Tubes.

Black Well Assay Plate.

High Sensitivity DNA Kit.

High Sensitivity DNA Reagent.

Table 4
SPRI bead buffer composition

Component	19.5%	20%	24%
5-M sodium chloride	10 mL		
1-M Tris–HCl, pH 8.0	500 µL		
0.5-M ETDA	100 µL		
10% v/v IGEPAL CA630	50 µL		
10% w/v sodium azide	250 µL		
PEG 8000	9.75 g	10 g	12 g
Ultrapure water	Up to 49 mL		

2.8 Equipment

Thermocycler.
Fluorescence microplate reader.
Qubit Fluorometer (Thermo Fisher Scientific).
Bioanalyzer 2100 (Agilent Technologies).
FACSAria Fusion (BD Biosciences).

3 Methods

3.1 Workflow

The single-cell RNA sequencing presented here encompasses (1) the dissociation of the sample to a single-cell suspension, (2) the isolation of individual cells by FACS, (3) the parallel processing of single-cell lysates to single-cell cDNA libraries using the Smart-Seq2 protocol, and (4) the indexing of these libraries using tagmentation before (5) pooling of the final libraries for sequencing (*see* Fig. 1). For setting up the flow cytometry, we recommend three control samples (*see* Subheading 3.4.2), and we provide guidelines for adapting the protocol for smaller cells/lower mRNA inputs (*see* Subheading 3.6.2).

3.2 Preparation of Collection Plates

[Time]: These steps take about 30 min of which 20 min are for thawing reagents and chilling racks.

3.2.1 Cleaning of the Workspace and Preparations

1. Clean working area with 70% (v/v) ethanol and DNA-Off wipes.
2. Thaw aliquots of the reagents in Table 1 (single-cell lysis buffer) on ice or at 4 °C.
3. Cool metal stands on ice or at 4 °C.
4. Once all reagents are thawed, mix them vigorously.

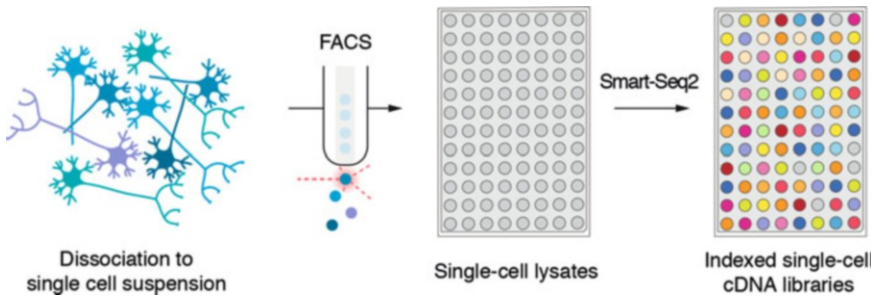


Fig. 1 Workflow from single-cell isolation to final indexed single-cell mRNA library pools

3.2.2 Place Single-Cell Lysis Solution in Plates

[Time]: This step takes about 15 min for preparing the first plate and 5 min for each additional plate.

1. Prepare the single-cell lysis buffer (*see* Table 1) in a microfuge tube. Mix it by inverting the tube at least ten times, and collect at the bottom by a quick spin in a microfuge (5–10 sec at less than 1000 g). Avoid vortexing the single-cell lysis solution because the detergent will foam up and sequester cells unnecessarily.
2. Assemble the Piko PCR frame plates (four 24-well plates per frame for a 96-well frame plate), and place them onto an ice-cold metal stand to cool.
3. Transfer 3 μL of the single-cell lysis buffer into each well using a multichannel pipette.
4. Seal the Piko PCR plates. Collect the lysis solution at the bottom of the wells by centrifugation for 15 sec at <2000 g.
5. Keep the plates on ice for use within the day (*see* **Note 1**).

3.3 Dissociation of Neurons in Monolayer Culture

The following steps should be adaptable to most cultured cells, although cells from different sources might require additional changes, e.g., use of other proteases for dissociation, different incubation times, or additional filtration steps to remove debris. We strongly recommend the inclusion of three controls for each sorting round to properly set up cytometric gates (*see* Subheading 3.4.2); live cells stained with TO-PRO-3 iodide, unstained live cells, and stained dead cells (*see* Subheading 3.3.3).

3.3.1 Preparations

[Time]: This step takes 1 h for pre-incubation of cells. During the last 15 min, start with step 2.

1. One hour before starting the dissociation, supplement the neuron culture with 5 μM Y-27632 ROCK inhibitor. It is included in all subsequent steps to prevent apoptosis during dissociation and processing.
2. Prepare sufficient FACS buffer, working dilution of TO-PRO-3 iodide, and tip blocking solution.
3. Warm up TrypLE Express supplemented with 5 μM Y-27632 ROCK inhibitor.

3.3.2 Dissociation and Staining

[Time]: This step takes 30–60 min depending on the dissociation time.

1. Gently remove the media from the cells cultured in 12-well plates with a micropipette.
2. Gently and without agitation, wash with 2 mL HBSS.
3. Add 0.5 mL TrypLE Express supplemented with 5 μM Y-27632 ROCK inhibitor per well.

4. Incubate at 37 °C on an orbital shaker at ≤100 rpm. After 20 min gently pipette up and down the cells 10–12 times, followed by further incubation with agitation. Repeat the pipetting every 5 min until single-cell suspension forms (*see* **Notes 3**, **4**, and **5**).
5. In the meanwhile, add TO-PRO-3 iodide to the FACS buffer to 1 μM final concentration (except for the unstained control).
6. Collect cells in 6 ml of pre-warmed Neurobasal+B27, and spin down at 200 g for 4 min.
7. For the first sample, also include the ethanol-fixed cells and spin them down too.
8. Resuspend in 300 μl FACS buffer and filter through a FACS filter mesh cap.
9. Maintain the single-cell suspension in FACS buffer on ice.
10. Proceed to FACS the cells within 30 min (*see* **Note 6**).

3.3.3 Stained Dead Cell Control Sample

[Time]: This sample can be harvested the day before and is used to set up the cytometric gates for separating live/dead cells. It requires one additional hour fixation time compared to live cell samples.

1. Prepare fresh 70% ethanol with absolute ethanol and distilled water. Transfer 3 ml to a FACS tube and chill it on ice.
2. Dissociate cells with TrypLE Express as above (*see* Subheading 3.3.2).
3. Resuspend in 5 ml Neurobasal medium and spin at 200 g for 4 min.
4. Collect cells in 500 μl HBSS.
5. Drip the cell suspension slowly into the ice-cold 70% ethanol while vortexing the FACS tube vigorously to avoid cell aggregates. Do not allow drops of cell suspension to fall onto the plastic directly (*see* **Note 7**).
6. Leave the samples on ice for at least 1 h to permeabilize.
7. On the day of the single-cell collection (*see* Subheading 3.3.2), spin down, and resuspend the fixed and permeabilized cells in 300 μL FACS buffer with TO-PRO-3 iodide.

3.4 Fluorescence-Assisted Cell Sorting (FACS) of Live Cells

This step needs to be optimized to the target cell population. Our set-up can serve as a starting point.

[Time]: This step takes 1 h of preparation time prior to the flow cytometry, followed by 30 min setup of the gating strategy and 15–30 min of collection time per 96-well plate.

3.4.1 Preparations and Flow Cytometer Settings

We sort neurons on a FACSAria Fusion (BD Biosciences) equipped with a 100 μm ceramic nozzle at maximum 20 psi pressure under sterile conditions. Both collection plates and fluids are cooled to 4 °C. Also, prepare the following items:

1. Ice box with the prepared collection plates.
2. Dry ice box with chilled metal stand for snap-freezing the single-cell lysates.
3. Centrifuge to spin down the collection plates.
4. Adhesive seals.
5. Ice box with the samples.

3.4.2 Sorting Strategy

In order to avoid extending the interval from harvest to lysis for the sensitive live cells, we do not mark cells by immunofluorescence staining, but mainly sort them based on morphological characteristics. However, we recommend to confirm the selection of the proper cell population with a reporter line if possible. For motor neurons we use a Hb9:GFP reporter line (control 1, prepared as in Subheading 3.3.2) [8].

It is important to exclude multiplets (cell clumps) and any dead cell material in the sorting to ensure a high-quality single-cell collection. We prefer to stain dead cells instead of live cells with the membrane-impermeant nuclear dye TO-PRO-3 iodide in order to exclude nucleic acid binding dyes in the lysates that may impair the Smart-seq2 preparation. In order to avoid other cellular fragments without nuclei (from neurites, axons, and apoptotic bodies), we adjust our gates with two additional controls: non-stained live Hb9:GFP motor neurons (control 2; *see* Subheading 3.3.2) and ethanol-fixed dead-stained Hb9:GFP motor neurons (control 3; *see* Subheading 3.3.3). The latter will reveal the position of the non-fluorescent cellular fragments without nuclei.

Thus, the following gating strategy can be set up with these three controls (*see* Fig. 2):

1. From all events: In FSC-H × SSC-H, select cells of the appropriate size and granularity (gate P1) to exclude debris.
2. From gate P1: In FSC-A × FSC-W, select singlets (gate P2).
3. From gate P2: In SSC-A × SSC-W, select singlets again (gate P3).
4. From gate P3: In FSC-H × TO-PRO-3 iodide, separate live soma (gate P4) from dead soma, apoptotic bodies, and other large debris (gate P5).

3.4.3 Collection

1. From gate P4 (live soma), index sort one event (cell) per well, and dispense 5-nL volume into 3 μL lysis buffer (*see* **Note 8**).
2. After finishing collecting a plate, put it into an ice-cold metal stand within the laminar flow unit, and seal it immediately with sealing foil.

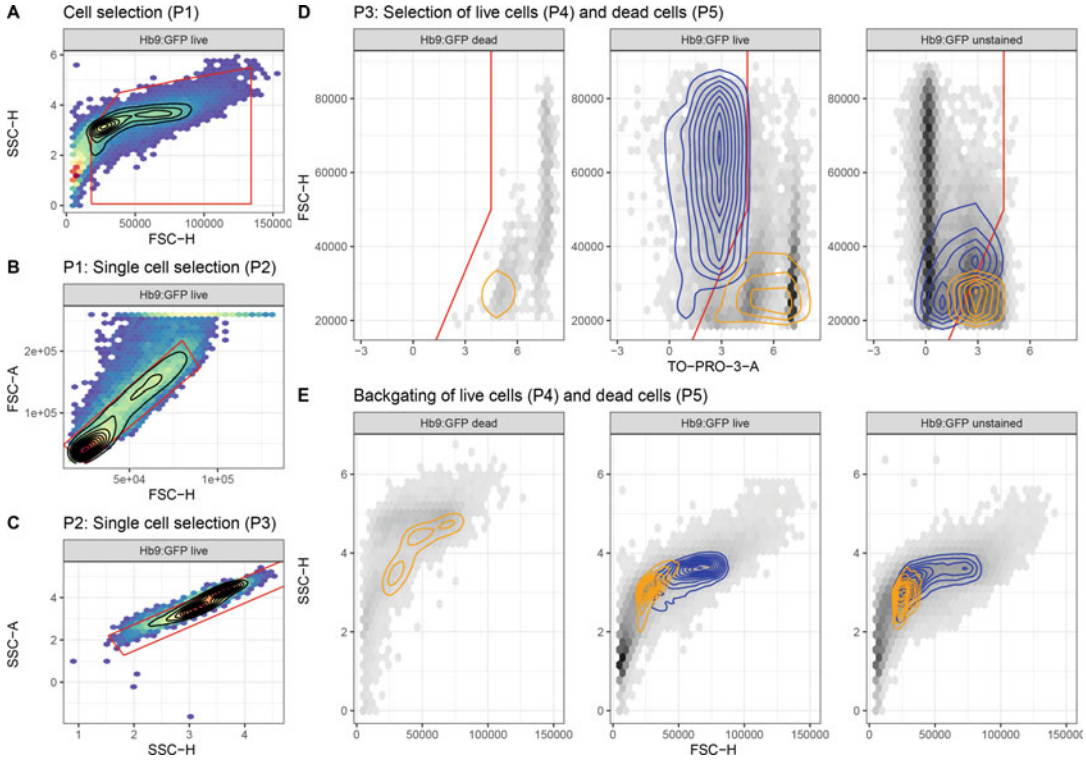


Fig. 2 Flow cytometry gates to sort live single neurons. (**a**) Cells are selected based on FSC-H × SSC-H over small debris (gate P1). (**b**) Cell clumps flair out and are biased toward the area dimension in FSC-H × FSC-A plots. Thus selection of near diagonal events ensures sorting of single cells (gate P2). (**c**) Similarly, single cells will line up diagonally in SSC-H × SSC-A plots. This gate re-affirms the selection (gate P3). (**d**) TO-PRO-3 only penetrates perforated cell membranes of dead cells to stain nuclei. Thus somas of dead cells can be excluded based on a fluorescence cut-off that is established with live stained and unstained samples (vertical segment). We find that dissociated neuron preparations often also contain relatively large fragments without nuclei (axon and neurite segments or apoptotic bodies). Thus we use a control sample with stained dead cells to adjust the gates (diagonal segment) sorting live cells (blue, gate P4) and dead fragments (orange, gate P5). (**e**) The live cells (blue, gate P4) and dead fragments (orange, gate P5) populations can be highlighted in the whole sample indicating the partial overlap based on simple parameters such as cell size (FSC-H) and internal structure (SSC-H) alone

3. Collect lysates at the bottom of the wells by centrifugation for 15 sec at <2000 g.
4. Immediately place the plate onto a metal stand cooled on dry ice to snap-freeze the single-cell lysates.
5. Place the plates into a cooled zip-lock bag on dry ice (*see* **Note 9**).
6. Store at −80 °C until processing the lysates.

3.5 Single-Cell Smart-Seq2 cDNA Libraries

3.5.1 Preparations

[Time]: These take about 30 min.

5. Clean working area with 70% ethanol and DNA-Off wipes.
6. Thaw aliquots of reagents on ice.
7. Cool metal racks on ice.
8. Once all reagents are thawed, mix vigorously.

3.5.2 First Strand Synthesis

[Time]: This step takes about 20 min hands-on time and 2 h for the reverse transcription reaction.

1. Prepare the first strand synthesis reaction mix (*see* Table 2). Mix it by inverting the tube, and collect at the bottom by a quick spin in a microfuge.
2. Set the thermocycler to 72 °C.
3. Place the Piko PCR frame plate with the frozen lysates onto an ice-cold metal stand. Using a plastic scraper, seal the plastic cover shut (*see* **Note 10**).
4. Collect the cell lysates at the bottom of the wells by centrifugation for 15 sec at <2000 g.
5. Remove the Piko PCR tube strips from the 96-well framed plate. The strips are held together by the plastic seal. Seal tightly and cut off the excess border of the plastic seal.
6. Place the tube strips in a thermocycler, put the cover on, and incubate at 72 °C for 3 min. This is the denaturation step of the reverse transcription.
7. Put the Piko PCR tube strips back into the 96-well framed plate. Immediately cool on ice for at least 2 min.
8. While the samples are cooling, start up the thermocycler program "1_RT_10" (*see* Table 5) to warm up to 42 °C.
9. Collect the samples by brief centrifugation. Remove seal carefully, and place Piko PCR framed plate onto an ice-cold metal stand.
10. Distribute the first strand reaction mix into an eight-well strip (36.5 μL per well). Add 2.76 μL first strand reaction mix to the side of each well with single-cell lysate using a multichannel pipette.
11. Seal the Piko PCR framed plate.
12. Combine the first strand synthesis reaction with the lysates by brief centrifugation for 15 sec at <2000 g.
13. Seal the plate with a plastic cover.
14. Remove Piko PCR tube strips from the 96-well framed plate. The strips are held together by the plastic seal.
15. Seal tightly and cut off the excess border of the plastic seal.

Table 5
Thermocycler program "1_RT_10"

Step	Action	Temp	Time	Cycles
1.	Warm up	42 °C	Inf	1 ×
2.	Initial reverse transcription	42 °C	90 min	1 ×
3.	RNA unfolding Template switch	50 °C 42 °C	2 min 2 min	10 ×
4.	Inactivation	70 °C	15 min	1 ×
5.	Final hold	4 °C	Inf	1 ×

16. Place the plate into the thermocycler, and add cover.
17. Run the "1_RT_10" program (continue to the second step, Table 5). This is the reverse transcription with template switching (*see* **Note 11**).

3.5.3 Second Strand Synthesis and Amplification

[Time]: This step takes 30 min for reagents to thaw, 20-min hands-on time, and subsequently 2.5–4 h for the PCR amplification.

1. Prepare the second strand synthesis reaction mix on ice (*see* Table 3) in a microfuge tube. Mix the second strand synthesis reaction mix by inversion, and collect by centrifugation for 15 sec at <2000 g.
2. Distribute the second strand reaction mix into an eight-well strip (100 µL per well) as a reservoir for pipetting.
3. Remove the first strand synthesis reaction samples from the thermocycler. Immediately place the Piko PCR tube strips back into the 96-well plate frame and onto an ice-cold metal stand.
4. Start up the thermocycler program "2_ISP_21" (*see* Table 6).
5. Seal with a plastic cover sheet, and collect reactions by brief centrifugation for 15 sec at <2000 g. Place the Piko PCR framed plate onto an ice-cold metal stand, and remove the seal carefully.
6. Add 7.5 µL second strand reaction mix to the side of each well with the samples using a multichannel pipette.
7. Seal the Piko PCR framed plate. Combine the reaction by brief centrifugation for 15 sec at <2000 g.
8. Remove Piko PCR tube strips from the framed plate. The strips are held together by the plastic seal. Seal tightly and cut off the excess border of the plastic seal.

Table 6
Thermocycler program "2_ISP_21"

Step	Action	Temp	Time	Cycles
1.	Hot start	98 °C	Inf	1 ×
2.	Initial denaturation	98 °c	3 min	1 ×
3.	Denaturation Annealing Elongation	98 °C 67 °C 72 °C	20 sec 15 sec 6 min	21 ×
4.	Final elongation	72 °C	5 min	1 ×
5.	Final hold	12 °C	Inf	1 ×

9. Place the second strand reaction tubes into the thermocycler, add cover, and run the "2_ISP_21" program (continue to the second step, Table 6). This is the second strand synthesis and amplification step.

3.5.4 cDNA Library Purification

[Time]: This step takes 45–60 min in total of which the first 20 min is for warming up reagents.

1. Equilibrate the SPRI beads (19.5% PEG-8 k) to room temperature for at least 20 min.
2. Distribute the magnetic beads into an eight-well strip (220 μL per well).
3. Column by column, add 12.5 μL beads per well to a 96-well deep well plate using a multichannel pipette.
4. Remove the samples from the thermocycler. Place the Piko PCR tube strips into the 96-well plate frame and onto an ice-cold metal stand.
5. Seal the plate with a plastic cover, and collect reaction by brief centrifugation for 15 sec at <2000 g. Remove seal carefully, and place the 96-well framed plate onto an ice-cold metal stand.
6. Column by column using a multichannel pipette, transfer the samples to the deep well plates with the magnetic beads. After each transfer, mix beads and reaction by pipetting up and down 10 times.
7. After the last transfer, cover the samples, and incubate at room temperature for 8 min to allow the magnetic beads to bind the samples.
8. Place the covered plate with the samples onto a magnetic stand to collect the beads for 5 min.
9. Discard the supernatant without disturbing the beads. Dry the beads for 10 min at room temperature without cover while the plate is still on the magnetic stand.

10. Distribute elution solution into an eight-well strip (160 μL per well) as reservoir for the multichannel pipette. While the samples are still on the magnetic stand, add 13 μL elution solution per well.
11. Remove the plate from the magnetic stand. Column by column, resuspend the beads in the elution solution by pipetting up and down 10 times.
12. Cover the deep well plate, and incubate it at room temperature for 5 min to release the samples from the beads.
13. Place the deep well plate onto the magnetic stand and bind beads for >2 min.
14. Label 12 low nucleic acid-binding eight-well strips, and place them into an ice-cold metal stand.
15. Column by column, transfer 12 μL of the eluate from the plate on the magnetic stand to eight-well strips.

While transferring, do not disturb the beads.

After the transfer and taking samples for quantification, seal the eight-well strips, and freeze at −20 °C.

3.6 Single-Cell cDNA Library Quantification and Quality Control

3.6.1 Single-Cell cDNA Library Quantification

[Time]: This step takes 15–30 min per plate.

Following the preparation, the concentrations of the individual single-cell cDNA libraries are determined, and their quality is assessed by profiling their molecular size distribution. The cDNA concentration can be measured in several ways. We use the Qubit assay system based on a fluorescent dsDNA intercalating dye (*see* **Note 12**) and record the fluorescent signal on a microplate spectrofluorometer (SpectraMax, Biotec) because it allows parallel quantification of full plates.

1. Place sufficient Qubit 1× dsDNA HS assay solution for all samples in a reservoir for the multichannel pipette.
2. Pipette 49 μL assay solution per well in black 96-well assay plates for samples. Keep the assay plates in the dark.
3. Collect the single-cell cDNA libraries by centrifugation of the storage plates for 15 sec at <2000 g. Carefully remove the sealing sheet.
4. Column by column, transfer 1 μL of each single-cell cDNA library directly into the assay solution in a black assay plate. Keep the assay plate in the dark.
5. After completing the transfer, seal the storage plate with the single-cell cDNA libraries. Collect them at the bottom of the well by centrifugation of the storage plates for 15 sec at <2000 g, and store at −20 °C.

6. Prepare the standards for assay calibration. Place 47.5 μL assay solution per well and add 2.5 μL of the standards. We use four replicates of each standard 1 (0 ng/μL dsDNA, background) and standard 2 (10 ng/μL dsDNA, upper boundary).
7. After preparing both samples and the standards, we measure the fluorescence signal in a spectrofluorometer (excitation at 485 nm/emission at 530 nm). The concentration of the single-cell cDNA library is calculated according to the following equation with the dilution factor $f_{\text{Dilution}} = 50$ and the average values of the standards.

$$\text{Conc} = f_{\text{Dilution}} \times 0.5 \times \left(\frac{F_{\text{Sample}} - F_{\text{Standard 1}}}{F_{\text{Standard 2}} - F_{\text{Standard 1}}} \right) \text{ ng}/\mu\text{l}$$

Typically, the single-cell cDNA libraries will have concentrations ranging from 0.5 to 5.0 ng/μL (*see* **Note 13**).

8. After completion of measurements, discard the assay plates.

3.6.2 cDNA Quality Control

[Time]: This step takes 30 min for thawing of assay reagents and 1-h run time per chip.

The molecular size distribution of the single-cell cDNA libraries is assessed with the High Sensitivity DNA Kit on a Bioanalyzer 2100 for chip-based capillary electrophoresis in accordance with the kit. We routinely pool 1 μL of each single-cell cDNA library and assess these pools for the overall cDNA quality in a sample set, e.g., over all single-cell libraries from a given cell line.

1. For each single-cell cDNA pool, prepare an eight-tube strip.
2. Collect the single-cell cDNA libraries by centrifugation of the storage plates for 15 sec at <2000 g. Carefully remove the sealing sheet.
3. Column by column, combine 1 μL of each single-cell cDNA library into an eight-tube strip.
4. After completing the transfer, seal the storage plate with the single-cell cDNA libraries. Collect them at the bottom of the well by centrifugation of the storage plates for 15 sec at <2000 g, and store at −20 °C.
5. Combine the contents from each tube of the eight-tube strip into a single tube. This is the pooled single-cell cDNA library sample for profiling on the Bioanalyzer.
6. Follow the High Sensitivity DNA kit instructions (or similar), and analyze the cDNA profile traces.

The aims of the molecular size profiles are (1) to confirm the presence of high-molecular-weight cDNA originating from intact mRNA and (2) to assess contamination with RT/PCR artifacts of low molecular weight such as primer dimers and potential

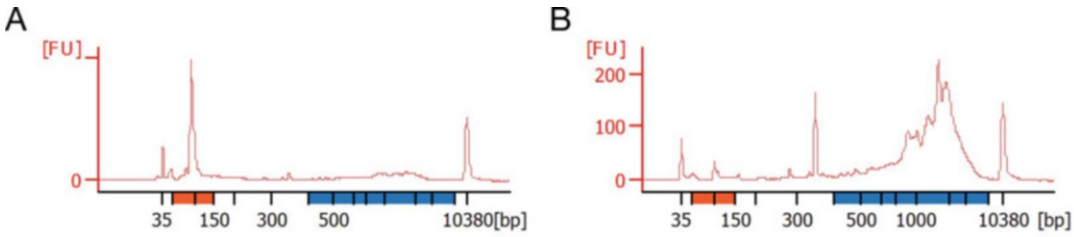


Fig. 3 Quality control for the cDNA library preparation. (**a**) This cDNA library preparation is dominated by primer dimers in the low-molecular-weight range (orange bar), and simple dsDNA quantitation will overestimate the high-molecular-weight cDNA originating from intact mRNA (blue bar). (**b**) This cDNA library is composed mainly of high-molecular-weight cDNA indicative of intact mRNA input and near optimal reaction conditions

concatemers from repeated template switching [2, 9]. The RT/PCR artifacts can make up substantial amounts of dsDNA if the reaction conditions are not optimal. These can interfere with the downstream steps as they mask the true cDNA content. Examples of molecular size profiles for single-cell cDNA libraries are given in Fig. 3.

If RT/PCR artifacts are present in the single-cell library preparations, we recommend i) to optimize reaction conditions to increase the cDNA yield (*see* **Note 14**) and ii) to introduce a molecular size cut-off of ca. 200 bp during purification with SPRI beads (typically use at 1:0.8 DNA:beads ratio).

3.7 Tagmentation

Here we use Nextera XT reagents to uniquely barcode and fragment our libraries. We only use the Tagment DNA buffer (TD buffer) and the Amplicon Tagment Mix (ATM) component. These can also be substituted with own preparations of Tn5 transposase loaded with the suitable adaptor oligonucleotides [10, 11] for Illumina compatible index primers.

3.7.1 Cleaning of the Workspace and Preparations

[Time]: These take about 30 min of which 20 min are for thawing reagents and chilling racks.

1. Clean working area with 70% (v/v) ethanol and DNA-Off wipes.
2. Thaw aliquots of the following reagents on ice or at 4 °C.
3. Cool metal racks on ice or at 4 °C.
4. Once all reagents are thawed, gently mix them by inverting the tubes.
5. Dilute cDNA samples to 200 pg/μL.
6. Set heating block or thermocycler to 55 °C.

Table 7
Thermocycler program "3_TGM_PHUS"

Step	Action	Temp	Time	Cycles
1.	Hot start	72 °C	Inf	1 ×
2.	Inactivation	72 °C	3 min	1 ×
3.	Initial denaturation	95 °C	30 sec	1 ×
4.	Denaturation Annealing Elongation	95 °C 55 °C 72 °C	10 sec 30 sec 30 sec	12 ×
5.	Final elongation	72 °C	5 min	1 ×
6.	Final hold	12 °C	Inf	1 ×

3.7.2 Tagmentation

[Time]: These steps take about 2 h of which the first 30 min is hands-on time.

1. On ice, combine 1.25 μL diluted cDNA sample (250 pg total) with 2.5 μL Nextera TD buffer (*see* **Note 15**).
2. Add 1.25 μL ATM to the sample and mix by pipetting twice up and down.
   Spin down to collect the samples at the bottom of the well/tubes.
3. Incubate the sample for 5 min at 55 °C.
4. Immediately add 1.25 μL 0.2% w/v sodium dodecyl sulfate to stop the reaction. Mix by gently pipetting up and down five times. Inactivate at room temperature for 5 min.
5. Start thermocycler with program "3_TGM_PHUS" (*see* Table 7) to get to starting temperature.
6. In the following enrichment PCR, the libraries get barcoded by unique combinations of Illumina-compatible i5/i7 indices. Add 1.25 μL of desired 0.2 μM index i5 primer and 1.25 μL of desired 0.2 μM index i7 primer to each sample.
7. Add 6.25 μL of the amplification master mix (*see* Table 8) and mix. Spin down the samples to collect the samples at the bottom of the wells.
8. Run the reactions on the thermocycler (continue to second step, Table 7).

3.7.3 Quantification and Purification of Indexed Single-Cell Libraries

[Time]: This step takes 1–2 h depending on the sample numbers that combine into a sequencing pool.

1. Before purification, we use 1 μL of each indexed single-cell library for quantification as in Subheading 3.6.1.

Table 8
Amplification master mix for library enrichment PCR

Component	1	rxn	110	rxn
Ultrapure water	1.675	µL	184.25	µL
5× Phusion HF buffer	3.0	µL	330.0	µL
Absolute DMSO	1.125	µL	123.75	µL
25 mM each dNTP	0.3	µL	33.0	µL
Phusion HF polymerase, 2 units/µL	0.15	µL	16.5	µL
Total	6.25	µL	687.5	µL

2. We combine 2 ng of each uniquely indexed single-cell library into a sequencing pool. The number of total samples depends on the sequencing depth, the available i5/i7 index pairs, and the sequencer. Typically, we pool between 72 and 384 samples per sequencing pool.
3. The final sequencing pool is then purified with SPRI beads as in Subheading 3.6.1 but with 24% SRPI beads at 0.8:1.0 beads to DNA ratio. The sequencing pool is eluted in 50 µL elution solution.

3.7.4 Quantitation of Sequencing Pools (Single Tubes)

[Time]: This step takes 15 min.
(*see* **Note 12**)

1. Distribute 199 µL Qubit 1× dsDNA assay solution onto 0.5 mL PCR tubes.
2. Add 1 µL cDNA pool.
3. Incubate in the dark at room temperature for at least 2 min.
4. Measure the diluted concentration (QF) using a Qubit fluorometer.
5. Determine original sample concentration by $conc = QF \times 200$.

3.8 Recipes

This protocol is an adaption from [12].

3.8.1 Preparation of SRPI Beads

1. In a 50 mL tube, combine the components for the SPRI bead buffer (*see* Table 4) (*see* **Note 16**).
2. Place the tube with the SPRI bead buffer on a rotary wheel for up to 2 h to completely dissolve the PEG-8000 pellets.
3. Then, chill the SPRI bead buffer on ice.
4. Fully resuspend SeraMag Speed Beads by vortexing them vigorously.
5. Place 1 mL of beads into a 2 mL tube, and place it on a magnetic stand for 2 min or until clear to precipitate the beads. Discard the supernatant.

6. Remove the 2 mL tube from the magnetic stand, and resuspend the beads in 1 mL ice-cold TE in order to wash the beads.
7. Repeat **steps 5** and **6** twice for washing the beads.
8. Finally resuspend the beads in 0.9 mL TE buffer (ca. 0.1 mL beads).
9. Add the ca. 1 mL bead suspension to the chilled SPRI bead buffer. Mix completely.
10. Assess the size cut-off of the SPRI beads by purifying 0.25 μg of a suitable DNA ladder at beads-to-DNA volumetric ratios from 0.6 to 2.0. According to the result, the ratio of beads to sample should be adjusted in the purifications for the batch of SPRI beads.
11. Aliquot in 2 mL tubes and store at 4 °C for up to 6 months.

4 Notes

1. It is best to prepare the collection plates on the day of the single-cell collection. However, the plates can also be snap frozen on dry ice. If doing so, vigorously mix up the lysis buffer upon thawing on ice, and collect on the bottom of the wells by centrifugation to avoid concentration gradients of any component.
2. The Recombinant RNase Inhibitor (RRI, Takara) requires 1 mM dithiothreitol (DTT) for full activity, whereas other ribonuclease inhibitors can be inhibited by it. Its inclusion depends on the type of RNase inhibitor used. The deoxynucleotides in the lysis buffer enhance the efficacy of the reverse transcription [1].
3. In our hands, it is essential not to shear or disrupt patches of neurons directly while pipetting. The neurons will come off first as a sheet and only later disband to single cells.
4. We also block/coat the tips before pipetting the cell suspension to minimize absorption of the cells to the plastic. To this end, pipette the tip blocking solution up and down several times with the tips just before using them with the cell suspension.
5. The overall incubation time depends on how quickly the single cell suspension is established. Typically this takes about 20–40 min for neurons.
6. Proceed as quickly as possible. In our hands, isolated motor neurons can be kept for a brief period of time (less than 1 h) on ice before they begin to die. In order to accommodate several samples for collection, it is recommended to work in a pair: one researcher harvests samples sequentially, and the other researcher performs the single-cell collection by FACS.

7. This is to prevent the formation of cell clumps when the aqueous cell suspension comes in contact with the ethanol phase. If the cell suspension flows down the wall of the tube into the ethanol, the aqueous drops would rather be emulsified, and cells would congeal when the phases mix. Vortexing breaks the surface tension of the ethanol phase and mixes in the cell suspension evenly and immediately. Thus the single cell suspension is maintained.
8. Elution in 5 nanoliter drops (5×10^6 μm^3 = 5 times $100 \times 100 \times 100$ μm^3) which leaves plenty of space for the collected neuron.
9. This is to prevent ice from directly contacting the plates as well as to prevent condensation when thawing the samples later because both constitute sources of potential environmental contaminants.
10. Some sealing sheets can detach when the plates are frozen at −80 °C.
11. Proceed at this stage if possible. For less than a day, the first strand libraries might be kept at 4 °C.
12. Several fluorescent and dsDNA-specific intercalating dyes can be used to quantify cDNA yield. We regularly use the Qubit dsDNA High-Sensitivity assay systems (Thermo Fisher Scientific) or the Quant-iT PicoGreen assay (Thermo Fisher Scientific).
13. If lower concentrations are obtained, this can indicate low quality of single cells after sorting, RNA degradation in the single cell lysates, or loss of cDNA during improper purification. If these have been ruled out, it is worthwhile to optimize the reaction conditions and to increase the amplification in the second strand synthesis step.
14. The optimization of two parameters in the Smart-Seq2 library preparation for new sample types is critical: the number of PCR cycles and the concentration of SMARTer_oligo-dTVN primer [5]. As a guideline, the amount of RT primer should be reduced for lower RNA inputs or smaller cells while increasing the cycle number in the second strand synthesis/amplification step. This will remove excessive primers which form the RT/PCR artifacts while simultaneously increasing the cDNA yield through further amplification. A 5′-biotin modification in the oligonucleotides abolishes the formation of concatemers effectively.
15. The ratio between the adaptor-loaded transposase (in the ATM mix) and the cDNA changes the integration frequency and thus fragment size. Using less cDNA will give shorter fragments. Thus an accurate quantification of the cDNA is critical. The protocol here is set up to yield indexed libraries mostly in

the size range from 300 to 500 bp. If libraries are consistently too large, the cDNA input amount should be lowered. If the libraries are too small, the cDNA input might be increased. However also low-quality RNA would yield already fragmented cDNA, which cannot be helped, and these samples should be excluded in the cDNA quality control step (*see* Subheading 3.6).

16. The IGEPAL CA630 acts as a surfactant here and reduces surface tension in the highly viscous bead suspension for easier pipetting. The sodium azide is a bacteriostatic and increases shelf-life of the bead suspension.

Acknowledgments

We thank Mattias Karlen for help with Fig. 1. Flow cytometry was performed in the Biomedicum Flowcytometry Core Facility with support of the Karolinska Institutet. We thank Juan Basile and Belinda Pannagel for assisting with flow cytometry. We are grateful for discussions with other members of the Hedlund laboratory. We are also thankful for discussions with and support from members of Rickard Sandberg's lab. The work in the Hedlund laboratory is supported by grants from the Swedish Research Council (grant number 2020-01049), The Radala Foundation for ALS Research (Switzerland), Ulla-Carin Lindquists Foundation for ALS Research (Ulla-Carin Lindquists stiftelse för ALS forskning), Åhlén-stiftelsen (grant number 213051), Olav Thon Stiftelsen (Norway), The Swedish Brain Foundation (Hjärnfonden) (grant number FO2021-0145) and Parkinsonfonden (grant number 1328/21). C.S. was supported by an Early Postdoc.Mobility fellowship from the Swiss National Science Foundation (P2BEP3_172233).

References

1. Picelli S, Björklund ÅK, Faridani OR et al (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods 10:1096–1098. https://doi.org/10.1038/nmeth.2639
2. Picelli S (2017) Single-cell RNA-sequencing: the future of genome biology is now. RNA Biol 14:637–650. https://doi.org/10.1080/15476286.2016.1201618
3. Ding J, Adiconis X, Simmons SK et al (2020) Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 38:737–746. https://doi.org/10.1038/s41587-020-0465-8
4. Nichterwitz S, Benitez JA, Hoogstraaten R et al (2018) LCM-Seq: a method for spatial transcriptomic profiling using laser capture microdissection coupled with PolyA-based RNA sequencing. Methods Mol Biol 1649:95–110. https://doi.org/10.1007/978-1-4939-7213-5_6
5. Picelli S, Faridani OR, Björklund AK et al (2014) Full-length RNA-seq from single cells using smart-seq2. Nat Protoc 9:171–181. https://doi.org/10.1038/nprot.2014.006
6. Nichterwitz S, Chen G, Aguila Benitez J et al (2016) Laser capture microscopy coupled with smart-seq2 for precise spatial transcriptomic profiling. Nat Commun 7:1–11. https://doi.org/10.1038/ncomms12139
7. Nijssen J, Aguila J, Hedlund E (2019) Axon-seq for in depth analysis of the RNA content of

neuronal processes. Bio-Protocol 9. https://doi.org/10.21769/BioProtoc.3312

8. Di Giorgio FP, Boulting GL, Bobrowicz S, Eggan KC (2008) Human embryonic stem cell-derived motor neurons are sensitive to the toxic effect of glial cells carrying an ALS-causing mutation. Cell Stem Cell 3:637–648. https://doi.org/10.1016/j.stem.2008.09.017
9. Kapteyn J, He R, McDowell ET, Gang DR (2010) Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. BMC Genomics 11. https://doi.org/10.1186/1471-2164-11-413
10. Picelli S, Björklund AK, Reinius B et al (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res 24:2033–2040. https://doi.org/10.1101/gr.177881.114
11. Hennig BP, Velten L, Racke I et al (2018) Large-scale low-cost NGS library preparation using a robust Tn5 purification and Tagmentation protocol. G3#58; Genes Genomes Genetics 8:79–89. https://doi.org/10.1534/g3.117.300257
12. Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946. https://doi.org/10.1101/gr.128124.111

Chapter 12

Computational Analysis of Single-Cell RNA-Seq Data

Byungjin Hwang

Abstract

Single-cell RNA sequencing (scRNA-seq) is gaining popularity as this allows you to profile a large number of individual cells. However, as the volume of the data increases, the need for appropriate computational methods also arises. Here, I will provide an overview of standard computational workflow for scRNA-seq and discuss each step and provide useful tips if applicable.

Key words Single-cell RNA sequencing (scRNA-seq), Unique molecular identifier (UMI), Batch effect, Normalization differential expression

1 Introduction

There has been a rapid progress in single-cell RNA-seq (scRNA-seq)-related technologies in recent years. This advance provided us many valuable insights into complex biological systems of healthy and diseased states. The increase in number of available software tools follows the new single-cell technology development and growing number of reported cells and genes. Therefore, it is becoming more crucial to standardize the analysis pipeline from QC step to downstream analysis and interpretation. Critically, this standardization will be useful for providing best practices for experimentalist and clinicians interested in analyzing their own data. This book chapter covers standard pipelines and popular tools for scRNA-seq, which will guide users from initial QC step to downstream analysis steps.

2 Initial Data QC and Normalization

2.1 Pre-Processing of the Raw Data and QC Visualization

After obtaining raw data from the sequencing machines (e.g., Illumina NovaSeq), FastQC (https://www.obtaining raw data from the sequencing m.babraham.ac.uk/projects/fastqc/) can be

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_12,

used to check the initial QCs such as base quality and overrepresented sequences. Although there are various technologies to generate a scRNA-seq library, here we focus only on droplet-based method that contains cell barcode and unique molecular identifiers (UMIs, for molecular error correction). Cell Ranger from 10× Genomics [1] is the most commonly used genome alignment and quantification tool to produce read count matrices that contain gene by cell barcode (raw by column) UMI counts. Alternatively, Kallisto [2] can be also used for non-standard cell barcode and UMI dimensions as this tool is flexible with configuration of the cell barcode and UMI space. Here, the cell barcode may not be actual "single cell" because two cells can be encapsulated in one droplet (multiplet). Due to the nature of the Poisson process, 10× Genomics protocol recommends loading 10–20 k cells in one channel (5–10% multiplet rate). After quantification, Cell Ranger produces both raw and filtered cell barcode/gene count matrices. Users can inspect whether valid cell barcodes are filtered in from the output HTML by checking the knee plot (*see* Fig. 1a, barcode count rank is plotted against the UMI counts).

If you run multiple reactions per condition (technical replicates), Cell Ranger needs to be run to aggregate data (over multiple GEMs) with depth normalization. For overloading experiments utilizing "cell hash" antibodies (*see* Fig. 1b), "Seurat" tool (recommend using RStudio for interactive visualization)'s HTODemux function can be run to designate the perturbation condition for each cell barcode. Note that, even though cell hashing increases the scRNA-seq throughput, multiplets are "identified" and trashed. If you want ultra-high-throughput scRNA-seq methods for most cost-effectiveness, scifi-RNA-seq [3] and SCITO-seq [4] can be used to profile more than >100 k cells in one channel (*see* Fig. 1c and d). Customized analysis pipelines can be found in the following website (scifi-RNA-seq: https://github.com/epigen/scifiRNA-seq_publication, SCITO-seq: https://github.com/yelabucsf/SCITO-seq_Manuscript).

After obtaining read count matrices from above, we must make sure that the count matrices contain only viable single cell data. We will focus on using two commonly used platform for downstream analysis: "Seurat (R)" and "Scanpy (Python)." Typically, count depth (number of UMIs) distribution per cell, gene number per cell, and fraction of counts from mitochondrial genes per barcode can be visualized (*see* Fig. 1e) with simple histograms. If the cell has low number of gene counts and high mitochondrial fraction, this could potentially mean mRNA leakage due to broken membrane. At this step, user could use multiplet detection tools such as Scrublet [7] or DoubletFinder [8] to filter out cell barcodes that contain multiple cells (*see* Fig. 1f).

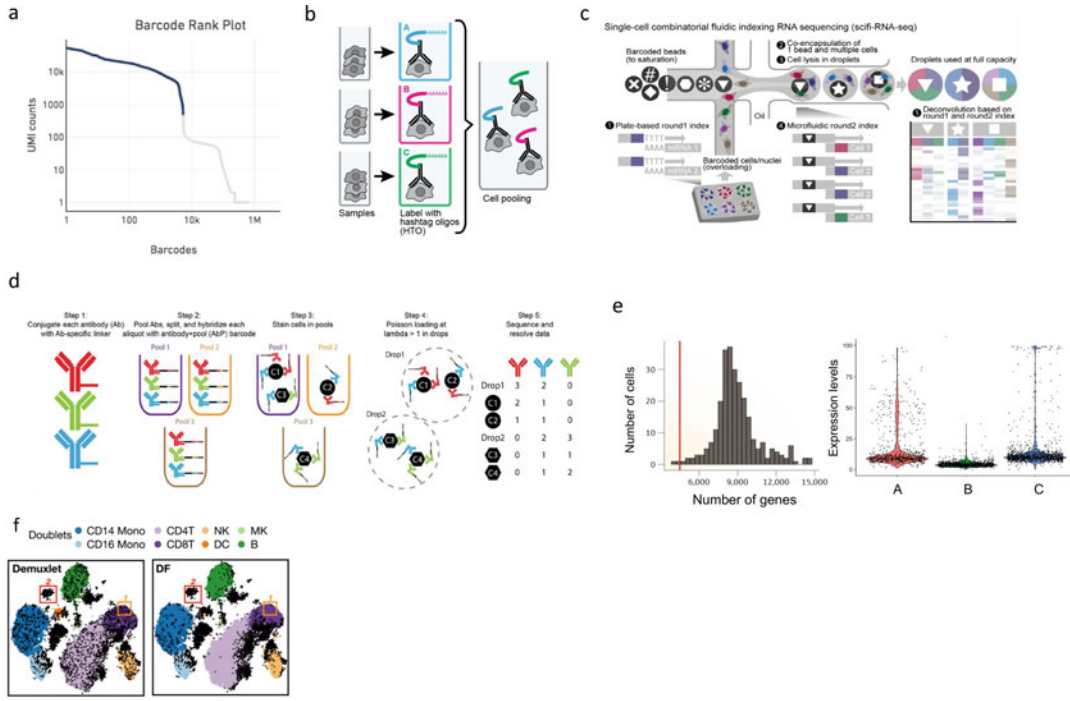


Fig. 1 Overview of the scRNA-seq sample QC approaches. (**a**) Example kneeplot from 10× Genomics showing top barcodes are filtered in (blue) compared to the background barcodes (gray). (**b**) Sample can be hashed with antibodies according to different environmental perturbations and pooled for scRNA-seq. These samples can be demultiplexed using hash barcodes embedded in conjugated oligos. (**c**) scifi-RNA-seq [3] approach utilizing in situ barcoded RT primers for ultra-high-throughput loading. (**d**) SCITO-seq [4] approach using combinatorial indexing to increase throughput of profiling proteins in a single-cell level. (**e**) QC histogram of number of genes expressed per cell and cells expressing certain level of mitochondrial genes can be filtered. (**f**) DoubletFinder [5] compares the doublet detection performance with Demuxlet [6]

The abovementioned QC metrics should not be used in isolation because they could contain biological relevance and artifacts during the library preparation and sequencing can affect the output unintentionally. Joint modeling of these covariates should be considered in the future to better filter out the cell barcodes. In addition to filtering out cell barcodes, genes can be filtered out if these are expressed in less than 20 cells which may affect detecting cell clusters.

As sufficient data quality cannot be determined explicitly a priori based on above filtering, we usually check quality of the further downstream analysis of cluster annotation after making an UMAP projection. We would recommend users do either use very permissive initial filtering and do iterative filtering along with downstream UAMP visualization of the expected clusters or store the QC covariates in the meta information without initial filtering. Unusual clusters can be clustered out separately in the UMAP space so that user can easily rule out in the downstream analysis.

2.2 Data Normalization

Data normalization is a way to normalize the data to overcome low-input and various forms of bias or noises (i.e., from sequencing) presented in the dataset to facilitate the downstream analysis.

The read count matrices are expected to be proportional to the gene-specific expression level and cell-specific scaling factors that are usually random. These nuisance variables include all the cell capture and reverse transcription efficiency and cell-intrinsic factors. Normalization does address sampling effects by scaling the counts data to correct the relative expression between cells. Most people use counts per million (CPM) normalization on their standard analysis. Basically, it uses size factor proportional to the count depth per cell using different factors of 10 (*see* Fig. 2a). The size factor calculation can be done as [raw gene count / sum of total count per cell] $*10^4$. The strong assumption behind this is all cells initially contained the same number of mRNA molecules and count depth difference is only due to sampling. So far, we mentioned per cell normalization (implemented in Python Scanpy), but gene counts can also be scaled the same way as we did in per cell (applied

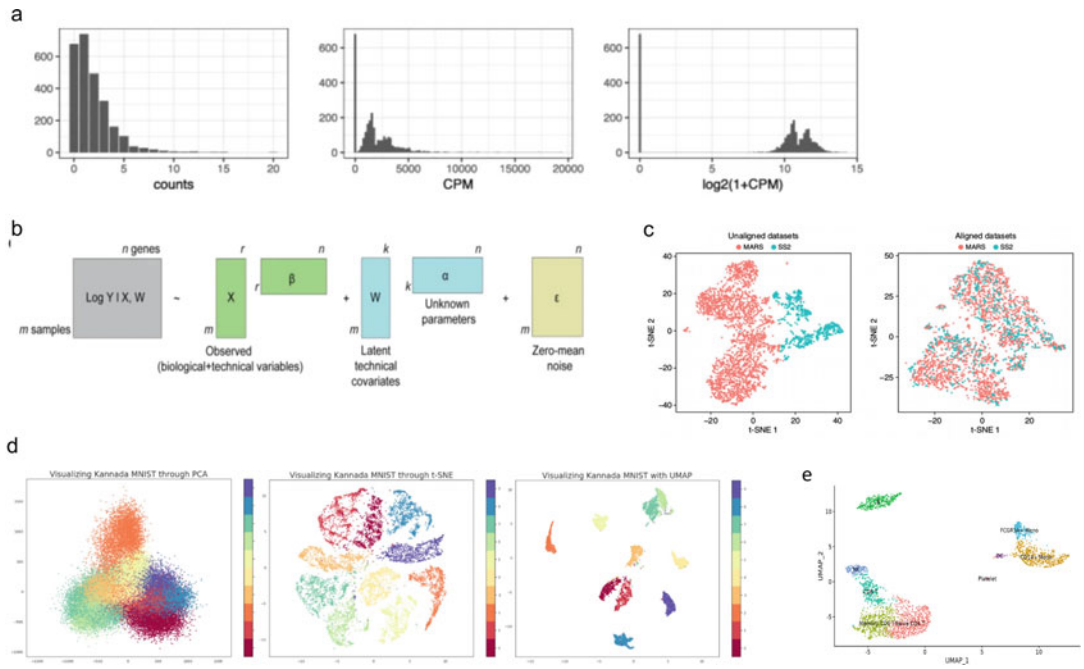


Fig. 2 Data normalization to downstream clustering analysis. (**a**) Normalization of the raw counts using CPM and log transformed distribution [9]. (**b**) Confounding factor modelling: the expression value *y* can be modeled as a linear combination of *r* technical and biological factors and *k* latent factors with a noise matrix [10]. (**c**) t-SNE plots of 3451 hematopoietic progenitor cells from murine bone marrow sequenced using MARS-Seq (2686) and SMART-Seq2 (SS2; 765), before and after alignment [11]. (d) Kannada MNIST dataset comparing PCA, t-SNE, and UMAP projection. Overall, UMAP preserves both global structure and local structure well compared to other methods. (**e**) Standard example of UMAP of PBMC sample in Seurat's guided tutorial. Each population's identity is overlayed on top of each cluster

in R Seurat). We leave this up to user's choice whether one thinks all genes should be equally weight or not. After normalization, we normally log(x + 1) transformed to mitigate mean–variance relationship and reduce the skewness of the data. This also facilitates downstream analysis where the assumption of batch correction and differential expression testing is normally distributed data.

2.3 Correction for Biological and Technical Covariates

Experiments are never done in a perfect way. Batch effects can be often present if you did the different experiment condition in a different day or in a different plate. Gene amplification process can be stochastic and only capturing fraction of those genes. Finally, sequencing can create noises as well. All these factors are called "technical covariates" that can be attributed to difficulties in interpreting the dataset. Another important variable is biological covariates where the cell cycles of captured single cells are different. Linear regression analysis is commonly done to regress out these effects [10].

The most common variable to remove is the effects of cell cycle, and this can be easily performed by running simple linear regression (*see* Fig. 2b) against a cell cycle score (implemented in both Seurat and Scanpy), or one can regress out the effect of mitochondrial gene expression. We don't normally regress out the ribosomal genes although many of these genes come out of highly expressed and sometimes make it difficult to interpret the result. This step is time-consuming, and regressing out the effect sometimes messes up the downstream analysis (i.e., proliferating cell identification); therefore we do recommend iterative QC to make sure in downstream clustering analysis that removing these effects makes sense. Technical artifacts can still be present because of the poor sampling result. Particularly in the text of trajectory analysis, regressing out the count depth factor is known to be helpful.

3 Integration of the Datasets

3.1 Batch Effect and Integration of the Multiple Datasets

Batch effect (technical, non-biological factors that also affect variation in the resulting data, *see* Fig. 2c) is very common in scRNA-seq experiments and can be avoided prior with smarter way of design by pooling of many samples in one batch is feasible. If using hashtag antibodies (e.g., universally expressed surface tags), you can encode multiple environmental perturbations in a different batch [5]. Genetic information (e.g., SNP) can also be used to demultiplex pooled individual donors [6]. However, if cells are inevitably grouped in a distinctive way such as on different 10× chips or harvested in different time points, this different environment may affect the transcript expression profiles. Combat [12] uses simple linear method taking both the mean and variance into account and is generally known to perform well on most scenarios. Integrating

dataset issue such as dataset produced by different specific condition is another distinct matter that has becoming prevalent. It is an encouraging fact that consortium level or higher-level data is becoming more available to the end users so that the user can upload their future dataset to QC individual data. Unlike linear batch effect correction method, this involves non-linear methods to project all cells into a shared embedding space (e.g., mutual nearest neighbors or canonical correlation analysis methods). Although we advise the user to be wary of over-correction issue for these non-linear methods, this opens the door for exciting applications such as mapping the billions of cell reference covering comprehensive set of tissues, organism, and various clinical settings.

4 Downstream Analysis

4.1 Dimensionality Reduction and Data Visualization

Dimensionality reduction is the first necessary step to visualize your dataset in a readable way. Since you cannot visualize all genes in 25,000 dimensions, some famous techniques such as PCA or MDS are often integrated in common softwares like Seurat and Scanpy to visualize only the important ones (highest principal components that distinguish your data clearly; usually the first principal component explains your data with the largest variance).

Starting with initial 25,000 genes coming out along with the count matrices, we need to filter out uninformative genes. Thus, highly variable genes (HVGs) are often selected [13] to facilitate downstream analysis (typically 1000–5000 genes). As implemented in both Seurat and Scanpy, they are simply selected based on the highest variance-to-mean ratio from binned mean gene expression values.

After narrowing the features down, dimension is further deducted by various algorithms. scRNA-seq data is considered to be a low-dimensional where some combination of expression profile can sufficiently describe the biological manifold far fewer than the number of genes. Principal component analysis (PCA) is the most popular method used here that uses linear approach to maximize the captured residual variance in each further dimension. This method chose the top N principal components which can be determined by "elbow plot" where you see ranked orders of the PCs in the descending order of variance ratio. For visualization, people tend to use non-linear reduction methods such as t-SNE [14] and UMAP [15]. t-SNE focuses on the local structure rather than global structure and largely depends on the perplexity parameter. Thus, UMAP is considered to be the best practice for exploratory data visualization as this is computationally fast and scalable to a large number of cells (*see* Fig. 2d).

4.2 Clustering and Annotating Cells

Clustering (*see* Fig. 2e) is the process where you group the cells based on similarity of their gene expression profiles. One of the important goal for scRNA-seq study is to define cell types and characterize them in groups. In this regard, this step is known to be very crucial for further downstream analysis where you dig into genes that are differentially expressed between these clusters. Clusters are assigned to minimize the intra-cluster distances using Euclidean, Cosine similarity [16], or Correlation-based metrics [17]. Another strategy to cluster cells is to use community detection methods such as Louvain and Leiden algorithm (optimized Louvain) that are implemented in both Seurat and Scanpy. These community detection methods are known to be faster than clustering method because of the reduced search space of the K-nearest neighbor (KNN) graph (K is usually chosen between values of 5 and 100).

After clustering, annotation of the cluster is finally performed. Finding this "identity" of this cluster relies on the external source such as the Human Cell Atlas. In the case of absence of the right reference, data-driven approach can be used to compare marker genes with the published datasets. The markers genes are selected by running the differential expression (DE) testing between the clusters either using the simple t-test or Wilcoxon's rank-sum test. Top ranked gene lists are then compared with the reference dataset (e.g., Human Cell Atlas or Mouse Cell Atlas). This iteration of clustering, annotation, sub-clustering, and annotation can be time-consuming. Therefore, one can use automated annotation tools such as scmap [18] or Garnett [19] to transfer annotation between the reference and the given dataset. But we recommend using both manual and automated approach since the reference dataset doesn't always contain the exact same cell identities of your given dataset.

References

1. Zheng GXY, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049
2. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527
3. Datlinger P, Rendeiro AF, Boenke T et al (2021) Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. Nat Methods 18:635–642
4. Hwang B, Lee DS, Tamaki W et al (2021) SCITO-seq: single-cell combinatorial indexed cytometry sequencing. Nat Methods 18:903–911
5. Stoeckius M, Zheng S, Houck-Loomis B et al (2018) Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol 19:224
6. Kang HM, Subramaniam M, Targ S et al (2018) Multiplexed droplet single-cell RNA--sequencing using natural genetic variation. Nat Biotechnol 36:89–94
7. Wolock SL, Lopez R, Klein AM (2019) Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst 8:281–291.e9
8. McGinnis CS, Murrow LM, Gartner ZJ (2019) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst 8:329–337.e4
9. Townes FW, Hicks SC, Aryee MJ, Irizarry RA (2019) Feature selection and dimension

reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol 20:295

10. Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50:96
11. Butler A, Hoffman P, Smibert P et al (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36:411–420
12. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127
13. Brennecke P, Anders S, Kim JK et al (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 10: 1093–1095
14. van der Maaten L (2008) Visualizing Data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=IwAR0Bggl1eA5TFmqOZeCQXsIoL6PKrVXUFaskUKtg6yBhVXAFFvZA6yQiYx-M. Accessed 7 Feb 2021
15. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software 3(29):861. https://doi.org/10.21105/joss.00861
16. Haghverdi L, Lun ATL, Morgan MD, Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36: 421–427
17. Kim T, Chen IR, Lin Y et al (2019) Impact of similarity metrics on single-cell RNA-seq data clustering. Brief Bioinform 20:2316–2326
18. Kiselev VY, Yiu A, Hemberg M (2018) Scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 15:359–362
19. Pliner HA, Shendure J, Trapnell C (2019) Supervised classification enables rapid annotation of cell atlases. Nat Methods 16:983–986

Chapter 13

Database for Plant Transcription Factor Binding Sites

Wen-Chi Chang and Chi-Nga Chow

Abstract

Reconstruction of gene regulatory networks is a very important but difficult issue in plant sciences. Recently, numerous high-throughput techniques, such as chromatin immunoprecipitation sequencing (ChIP-seq) and DNA affinity purification sequencing (DAP-seq), have been developed to identify the genomic binding landscapes of regulatory factors. To understand the relationships among transcription factors (TFs) and their corresponding binding sites on target genes is usually the first step for elucidating gene regulatory mechanisms. Therefore, a good database for plant TFs and transcription factor binding sites (TFBSs) will be useful for starting a series of complex experiments. In this chapter, PlantPAN (version 3.0) is utilized as an example to explain how bioinformatics systems advance research on gene regulation.

Key words Transcription factor, Transcription factor binding site, Promoter, Bioinformatics, Regulatory network, Database

1 Introduction

To reconstruct transcriptional regulatory networks is an important but difficult issue in plant sciences, due to the limited information about transcription factors (TFs) and their binding targets. In the past two decades, several high-throughput technologies have been developed, such as microarray and next-generation sequencing (NGS). Both technologies were applied to investigate the genome binding landscapes of TFs, using chromatin immunoprecipitation (ChIP)-on-chip, chromatin immunoprecipitation sequencing (ChIP-seq), DNase-seq, and DNA affinity purification sequencing (DAP-seq) [1]. Therefore, TFs and their corresponding binding motif could be retrieved from those arrays and sequencing data through bioinformatics methods. Numerous databases were created for integrating TFs and their target genes. For example, Cistrome DB maps the TFs and transcription factor binding sites (TFBSs) based on genome-wide scales in humans and mice by collecting big data from ChIP-seq, DNase-seq, and ATAC-seq [2]. Compared to mammals, the related resources are limited for

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_13,

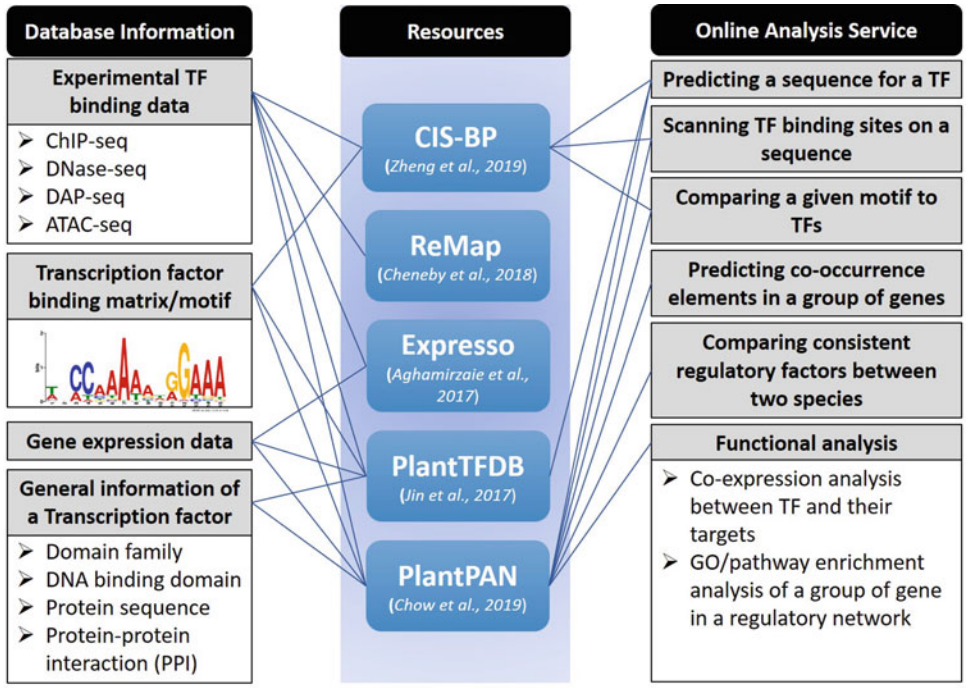


Fig. 1 The analysis functions designed in several databases for plant gene transcriptional regulation. The websites for the above resources are as follows: CIS-BP, http:/cisbp.ccbr.utoronto.ca/; ReMap, http:/remap.univ-amu.fr/; Expresso, https:/bioinformatics.cs.vt.edu/expresso/; PlantTFDB, http:/planttfdb.gao-lab.org/; and PlantPAN, http:/plantpan.itps.ncku.edu.tw/

plants. Recently, CIS-BP Database [3], PlantTFDB [4], Expresso [5], ReMap [6], and PlantPAN3.0 [7] were developed to integrate ChIP-seq, DAP-seq, and SELEX data for reconstruction of transcriptional regulation networks in plants. As shown in Fig. 1, most databases provide "search" and "browse" functions for retrieving general information about a TF. For example, the target genes discovered from the high-throughput sequencing data, binding matrix, protein sequence, expression data, functional domain, and DNA binding motif of a TF can be accessed. In addition, several resources provide online analysis services, so users can predict regulatory elements or corresponding TFs for an input sequence. These platforms are briefly described as follows. The CIS-BP database collects TFs and their corresponding binding DNA motifs from numerous organisms, including 73 plant species, so that users can predict TF binding motifs on their input DNA sequences. PlantTFDB4.0 provides comprehensive information about TFs, such as DNA binding motifs, functional domains, 3D structures, gene ontology (GO), plant ontology (PO), and interactions. It also integrates PlantRegMap [5] to provide regulatory landscapes of TFs from high-throughput data. Expresso is a web server for investigating TFs and their target genes based on ChIP-seq data.

However, only 20 TFs were collected in the current version of Expresso. ReMap is also a database that integrates ChIP-seq, ChIP-exo, and DAP-seq to identify transcriptional regulatory atlases for humans and *Arabidopsis*. Most of the databases mentioned above show the list of target genes for a TF and are limited in further functional assays for a gene or a group of genes. Therefore, comparison analyses across species, co-occurrence TFBS analyses, and functional analyses of a group of genes are designed and programmed in PlantPAN3.0. Since PlantPAN (version 3.0) is the latest and most comprehensive database for plant TFs/TFBSs, PlantPAN is utilized as an example to explain how bioinformatics databases are applied in plant research.

2 Materials

In this chapter, we focus on the application of bioinformatics methods for TF–target gene discovery and gene regulatory network reconstruction. Several materials such as the "gene name" and "sequence" obtained from gene banks can be used for analysis. In the following sections, the database used to demonstrate the utility of bioinformatics resources is PlantPAN system (version 3.0) (http://plantpan.itps.ncku.edu.tw/).

2.1 Gene Name and Keyword

Normally, gene name and keyword can be used as search inputs in numerous resources. General information about gene description, gene symbol, chromosome location, gene structure, and gene and protein sequences can be easily accessed. Furthermore, promoter analysis and functional assays can be processed via the web interface.

2.2 Sequences

The gene sequence, protein sequence, promoter sequence, and motif sequence are allowed for further analysis in PlantPAN, based on different purposes. How to obtain gene regulatory information from these sequences via the PlantPAN system will be illustrated in the Methods.

3 Methods

Eight general questions about reconstructing gene regulatory networks are listed below. The analytic methods used to answer these questions are described point by point in Subheadings 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8.

1. How do I identify TFs/TFBSs on a promoter sequence or multiple promoter sequences?
2. There are many TFs/TFBSs that were predicted on my promoter. Are there any methods that could help me to select high-confidence TFs/TFBSs?

3. How do I identify co-occurrence TFs/TFBSs on a group of promoters?
4. How do I identify conserved regulatory TFs between two species?
5. How do I identify regulatory TFs and their binding sites on a certain gene based on high-throughput experimental methods?
6. How do I identify target genes of a TF based on high-throughput experimental methods?
7. I have identified a TF binding sequence by experimental methods (i.e., deletion analysis); how do I identify a similar motif or its corresponding TF in the database?
8. How do I identify enriched GO terms or pathways in a regulatory network?

3.1 Identification of TFs/TFBSs on Promoters

"Gene Search" and "Promoter Analysis" are available for discovering TFs/TFBSs on promoter sequences. The gene name and keywords can be input into the "Gene Search" function, and TFs/TFBSs can be predicted for selected regions. The TFs/TFBSs located in the conserved regions, tandem repeats, and CpNpG islands will be shown on the sequence with different labels. A promoter sequence can be input into the "Promoter Analysis" function, and then analysis outputs similar to "Gene Search" will be shown on the webpage. "Multiple promoter analysis" is also designed on the webpage for scanning TFs/TFBSs on two or more promoters. An example of the results is displayed in Fig. 2.

3.2 Using Co-expression Analysis to Select High-Confidence TFs/TFBSs

After predicting numerous TFs/TFBSs on a promoter, the "Co-expression Analysis" function may be useful; it is on the output page of the TFs/TFBSs analysis, as shown in Fig. 2. It supposes that TFs co-expressed with target genes might play more critical roles than those without co-expression patterns. Therefore, different correlation methods, protein–protein interaction (PPI) prediction model scores, and conditions are programmed to identify high-confidence TFs. The TFs that are co-expressed with the target gene under different conditions will be selected to reconstruct gene regulatory networks (GRNs) (*see* Fig. 3). The binding sites of those TFs are listed in a table for easy retrieval.

3.3 Identification of Co-occurrence TFs/TFBSs in a Group of Promoters

"Gene Group Analysis" is designed to identify co-occurrence TFs/TFBSs in a group of promoters. Both gene IDs and sequences can be input into the "Gene Group Analysis" by selecting plant species or "Multiple promoter analysis," respectively. After setting the parameters, a list of TFs/TFBSs that co-occurred in the query genes (or sequences) will be shown.

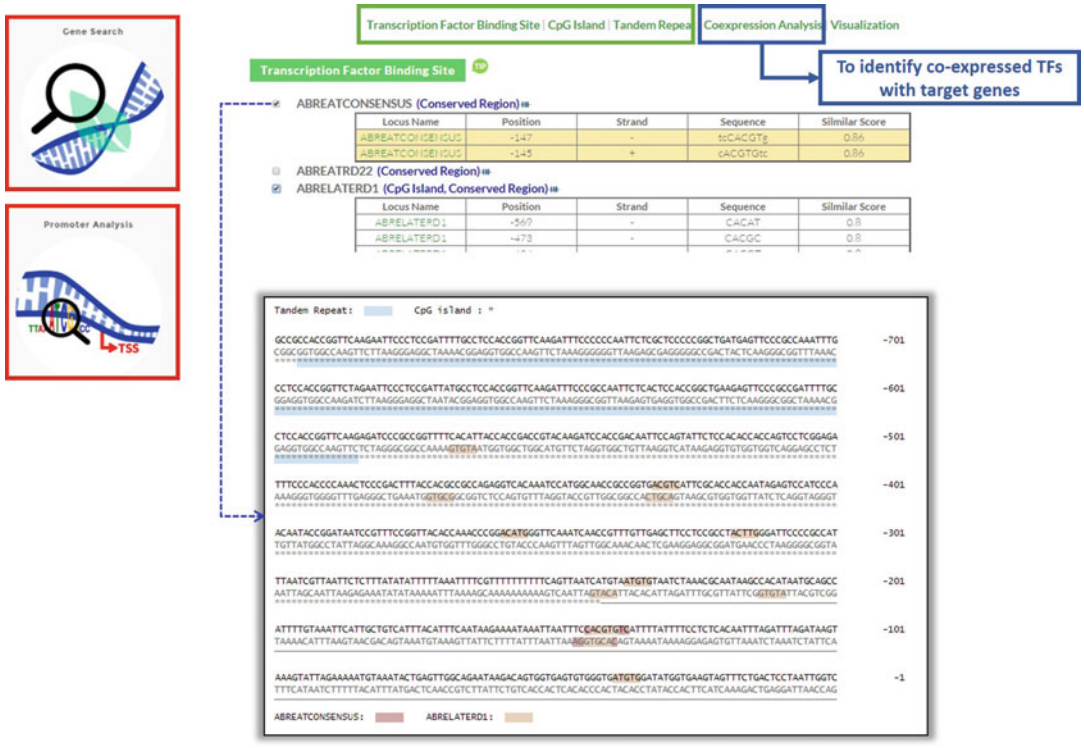


Fig. 2 The web interface of a TFs/TFBSs analysis of a promoter sequence

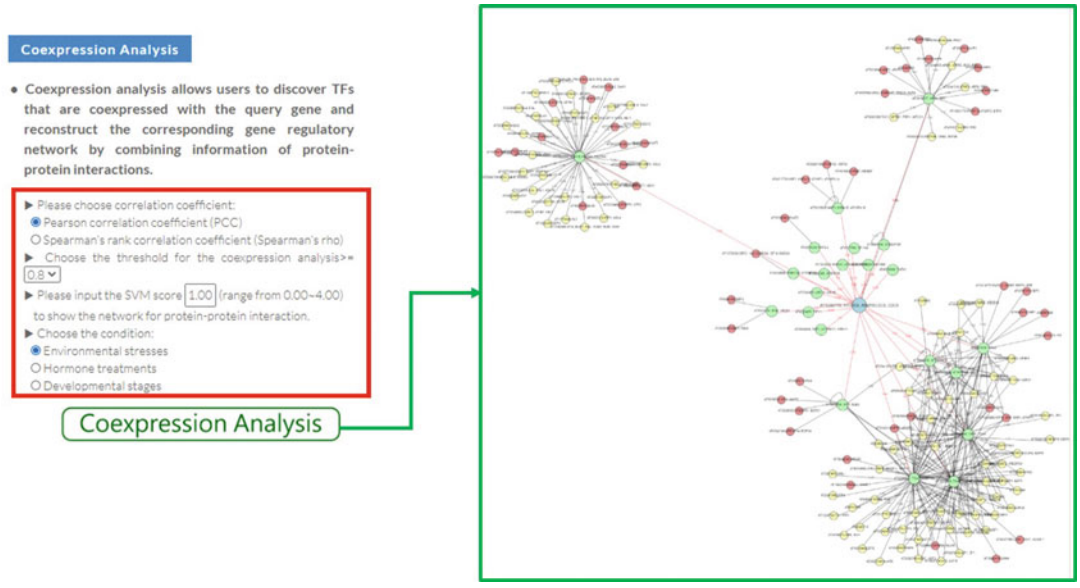


Fig. 3 The blue dot is the target gene (query). Green dots are TFs co-expressed with the target gene under different conditions. Red dots are proteins interacting with TFs based on the experimental method. Yellow dots are proteins interacting with TFs based on the prediction method

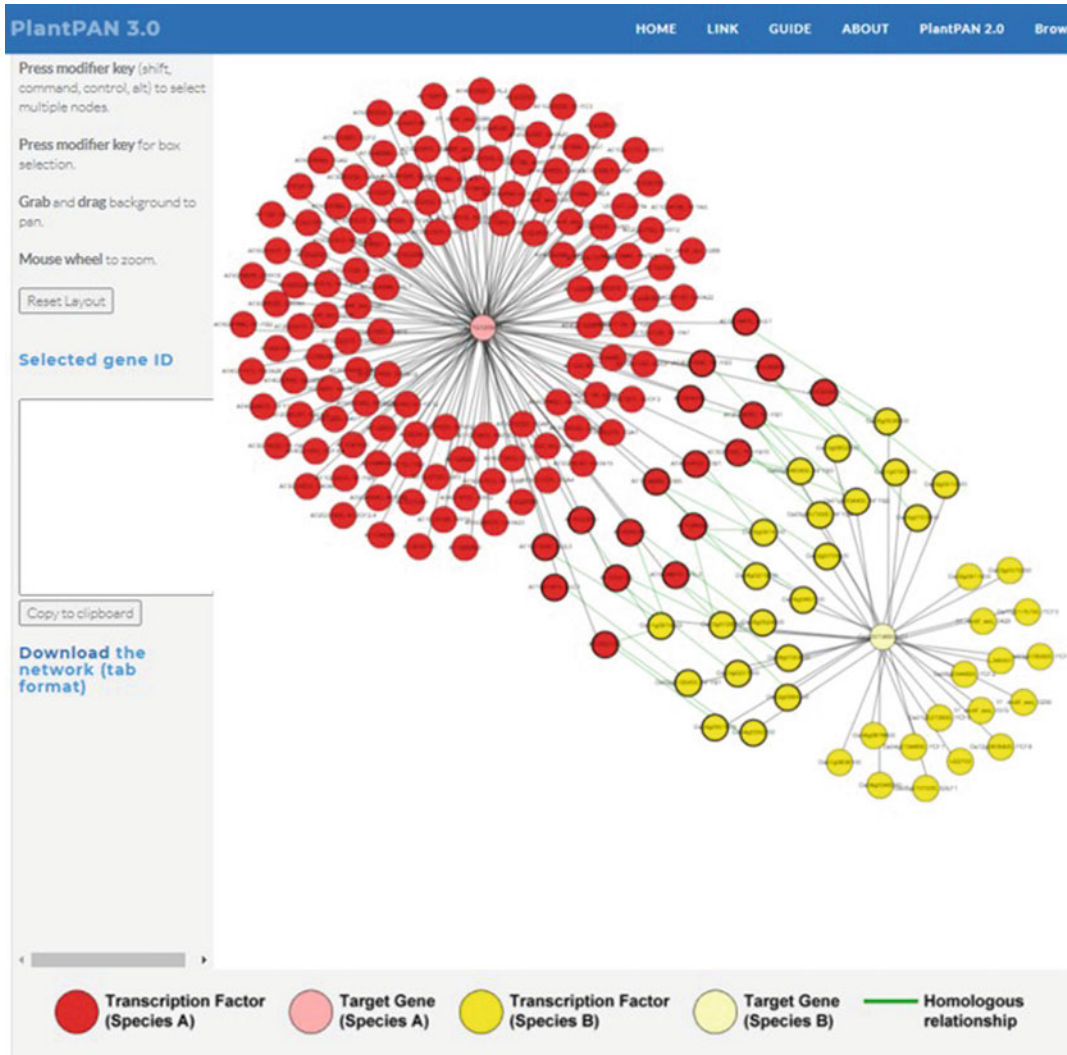


Fig. 4 The pink dot is the target gene (query) from the first species. The light-yellow dot is homologous to the target gene in the other species. Red dots are TFs predicted in the conserved region on the target gene promoter. Dark-yellow dots are TFs predicted in the conserved region on the homologous gene promoter. Each green line connects the homologous gene pairs between the two species

3.4 Identification of Conserved Regulatory TFs Between Two Species

The homologous TFs and proteins collected from seven plant species, including *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Glycine max*, *Solanum lycopersicum*, *G. hirsutum*, and *Arabidopsis lyrata*, are used to investigate conserved TFs and proteins between GRNs of two species in PlantPAN. *Arabidopsis* and rice are utilized as examples to demonstrate the utility of "Cross Species" in PlantPAN. First, a gene ID from *Arabidopsis* is input, and "rice" is selected for comparison. Then, TFs/TFBSs located in conserved regions between *Arabidopsis* and rice are shown in a figure and listed in a table. Finally, GRNs are reconstructed in accordance with various conserved regions. As displayed in Fig. 4, it is easy to compare the conserved TFs between two GRNs from two species.

3.5 Identification of TFs and Their Binding Sites on a Gene Based on High-Throughput Experiments

Plant ChIP-seq Database (PCBase) is a sub-database integrated into PlantPAN. PCBase collects plant ChIP-seq experimental data derived from the Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). The landscapes of TFs and their binding sites are analyzed genome-wide in PCBase. The genes from seven plants can be accessed for their regulatory TFs via "Gene Search" in PCBase. The TF binding sites (regions) on the query gene are listed according to different experiments (*see* Fig. 5). All peaks from experimental data can be downloaded via PCBase. Furthermore, if the query genes are not from the seven species provided from PCBase, sequences can be input for scanning TFBSs based on the TF matrices developed from PCBase via "Promoter Analysis."

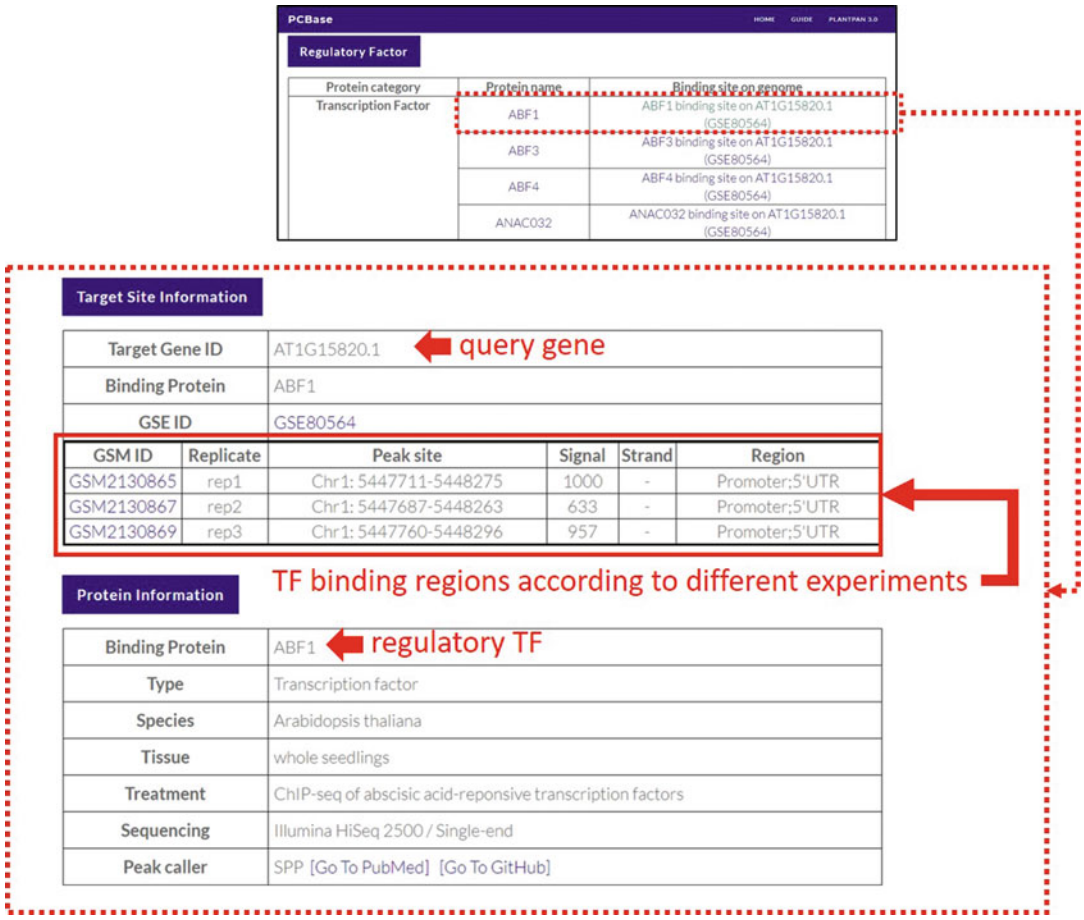


Fig. 5 An example of output results of the "Gene Search" in PCBase. The TFs that regulate the target gene (query) are listed. After selecting a TF, the TF binding regions on the query genes will be listed

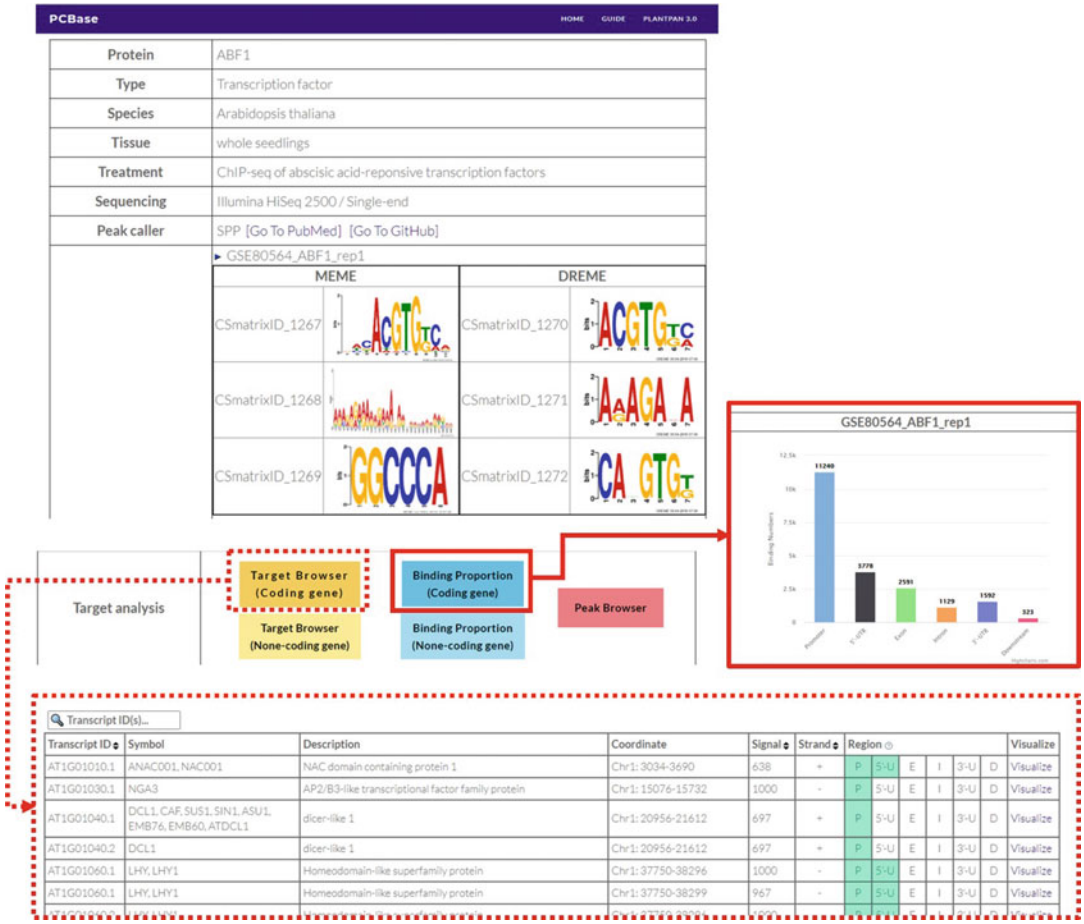


Fig. 6 An example of search results for ABF1 from *Arabidopsis* via "Protein Search" in PCBase

3.6 Identification of Target Genes of a TF from High-Throughput Experiments

All TFs listed in "Protein Search" in PCBase contain ChIP-seq evidence. When selecting a TF from "Protein Search," the binding sequence logos (matrix) of the TF (query) analyzed by ChIP-seq data are illustrated in the results page. Moreover, the targets of the TF can be discovered according to different experiment samples. Figure 6 demonstrates an example of search results of ABF1 from *Arabidopsis*. The binding proportion of ABF1 in a sample (GSE80564_rep1) is displayed. It shows that most ChIP-seq peaks are located in promoters, but several peaks are also identified in 5′-UTR, exon, intron, 3′-UTR, and downstream regions. The target genes of the TF can be accessed via "Target Browser."

3.7 Comparing and Analyzing a Known Motif

A motif comparison is programmed in PlantPAN3.0. A motif encoded in an International Union of Pure and Applied Chemistry (IUPAC) [8] name is used to discover similar TF matrices or its corresponding TF in the database via "TF/TFBS Search" (*see* Fig. 7). Alternatively, a TF name or keyword can also access its

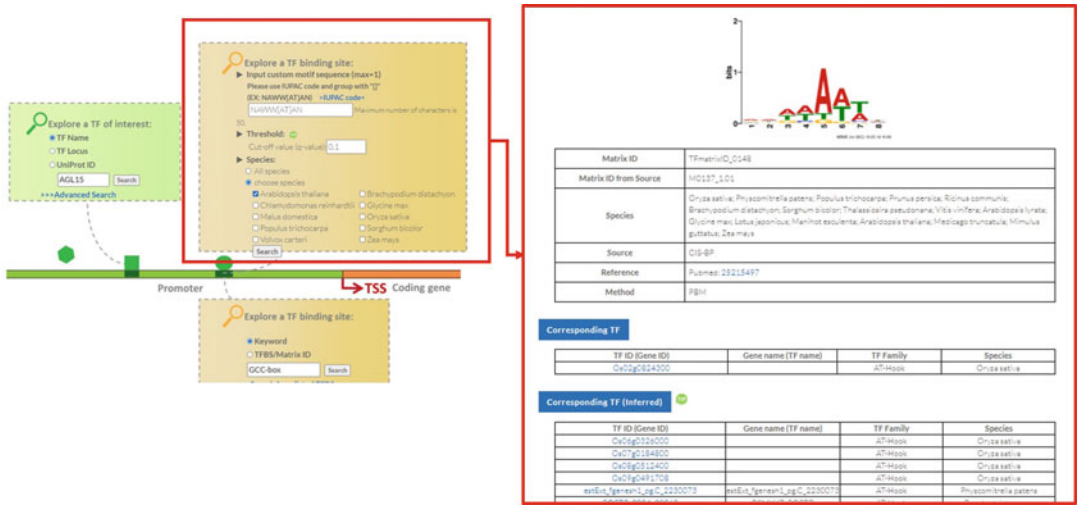


Fig. 7 A motif sequence can be input to discover similar TF matrix and its corresponding TF via "TF/TFBS Search" in PlantPAN

binding matrix. Additionally, user-customized motifs can be input for scanning their binding sites on a promoter sequence by using "Promoter Analysis."

3.8 Identification of Enriched GO Terms and Pathways in a Regulatory Network

After reconstructing a GRN, the related gene functions and pathways are worth investigating. The GRN in PlantPAN is generated by the Cytoscape tool [9], as shown in Fig. 3. Every node in the GRN can be selected for further analysis. An example is illustrated in Fig. 8: nodes in a sub-network or whole network are selected, and a hypergeometric distribution method is applied to evaluate the GO/pathway enrichment of a group of selected genes. A lower p-value indicates higher significance of the GO term or pathway in the GRN.

4 Conclusion

All analysis results and TF matrix data from ChIP-seq in PlantPAN can be downloaded. The detailed guide for each analysis function is displayed in http://plantpan.itps.ncku.edu.tw/index.html#guide. Several analysis functions described above are also available in CIS-BP, ReMap, Expresso, and PlantTFDB. For example, the experimental TF binding data can be revealed from all these databases. Comprehensive protein information of a TF can be accessed by PlantTFDB. Identification of TFs and TFBSs on a gene or promoter sequence is also available in CIS-BP.

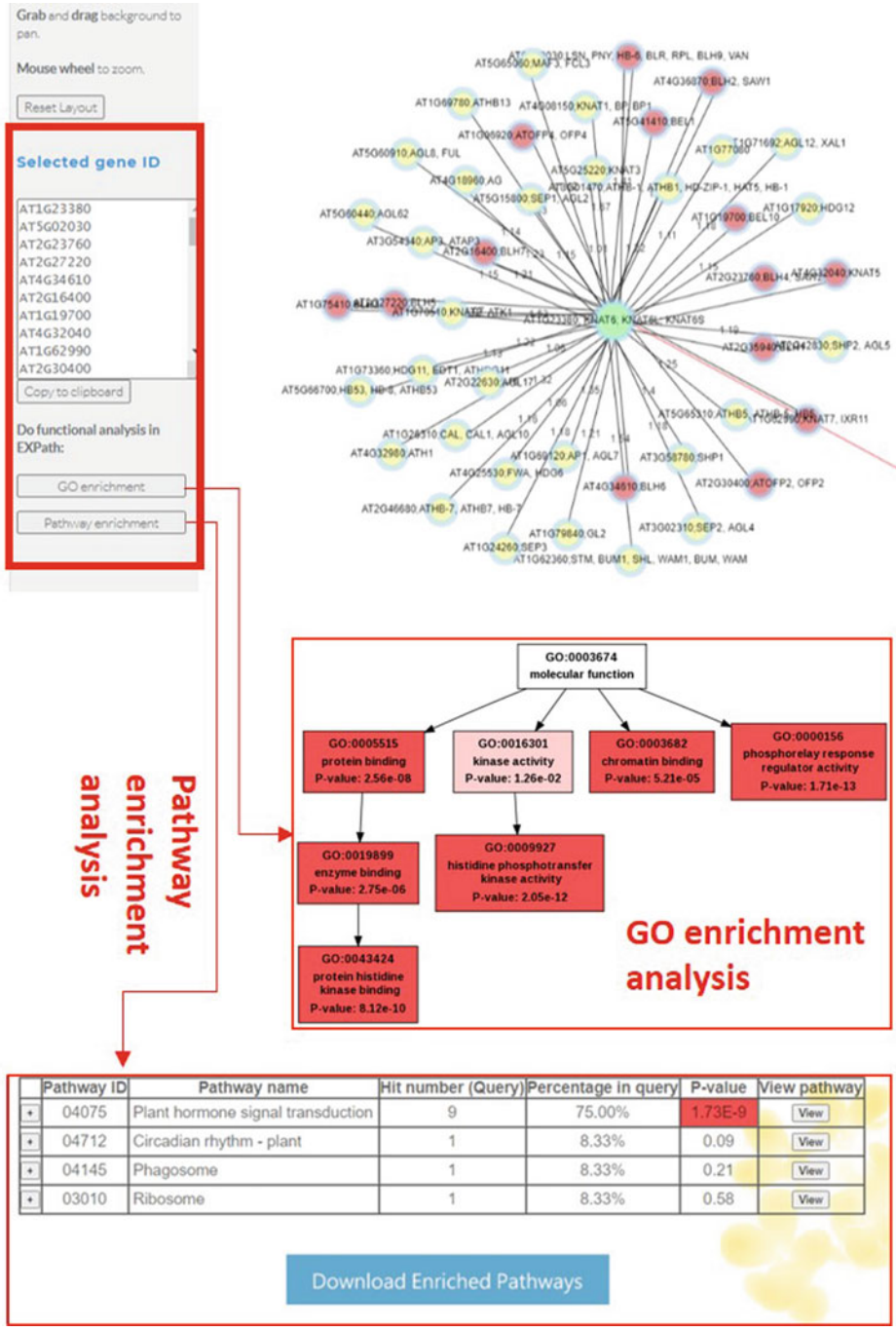


Fig. 8 Every node (gene) in the GRN can be selected as a gene group and analyzed for its enriched GO terms and pathways

Acknowledgments

The authors would like to thank the Ministry of Science and Technology (MOST 108-2311-B-006 -002 -MY3 and MOST 110-2918-I-006-001) of the Republic of China for financially supporting this research. Computational analyses, data mining, and storage were performed using the system provided by the Bioinformatics Core at the National Cheng Kung University and National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs), supported by the Ministry of Science and Technology, Taiwan.

References

1. Bartlett A, O'Malley RC, Huang SC, Galli M, Nery JR, Gallavotti A, Ecker JR (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc 12(8):1659–1672. https://doi.org/10.1038/nprot.2017.055
2. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, Chen CH, Brown M, Zhang X, Meyer CA, Liu XS (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res 47(D1):D729–D735. https://doi.org/10.1093/nar/gky1094
3. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158(6): 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009
4. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45(D1):D1040–D1045. https://doi.org/10.1093/nar/gkw982
5. Aghamirzaie D, Raja Velmurugan K, Wu S, Altarawy D, Heath LS, Grene R (2017) Expresso: a database and web server for exploring the interaction of transcription factors and their target genes in Arabidopsis thaliana using ChIP-Seq peak data. F1000Res 6:372. https://doi.org/10.12688/f1000research.10041.1
6. Cheneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res 46(D1):D267–D275. https://doi.org/10.1093/nar/gkx1092
7. Chow CN, Lee TY, Hung YC, Li GZ, Tseng KC, Liu YH, Kuo PL, Zheng HQ, Chang WC (2019) PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. Nucleic Acids Res 47(D1):D1155–D1163. https://doi.org/10.1093/nar/gky1081
8. Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res 13(9):3021–3030. https://doi.org/10.1093/nar/13.9.3021
9. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics 26(18):2347–2348. https://doi.org/10.1093/bioinformatics/btq430

Chapter 14

MicroRNA Regulatory Network Analysis Using miRNet 2.0

Le Chang and Jianguo Xia

Abstract

MicroRNAs exert their effects in the context of gene regulatory networks. The recent development of high-throughput experimental approaches and the growing availability of gene expression data have permitted comprehensive functional studies of miRNAs. However, the data interpretation is often challenging due to the fact that miRNAs not only act cooperatively with other miRNAs but also participate in complex networks by interacting with other functional elements, including non-coding RNAs or transcription factors that often have extensive effects on cell biology. This chapter provides detailed practical procedures on how to use miRNet 2.0 (https://www.mirnet.ca) to perform miRNA regulatory network analytics to gain functional insights.

Key words miRNAs, Network analysis, Gene regulatory networks, Systems biology

1 Introduction

MicroRNAs (miRNAs) are short (~22 nucleotide length) non-coding RNA molecules that can regulate gene expression at the post-transcriptional level and usually act as control nodes or hubs in gene regulatory networks [1]. Understanding miRNA function is challenging due to the "many-to-many" relationships between miRNAs and their target genes. In addition, complex interplay exists between miRNAs and other functional elements, such as transcription factors (TFs), long non-coding RNAs (lncRNAs), etc. [2]. However, building miRNA regulatory networks and interpreting the results are not a straightforward task. Users need to manually curate data from multiple databases, construct interaction networks, and visualize them in some software package. Many such network visualization tools are not dedicated to miRNA networks and do not offer extra support beyond visualization. Researchers have to resort to additional tools to gain biological insights. Therefore, there is an urgent demand for easy-to-use and one-stop bioinformatics tools to support miRNA regulatory network analysis.

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_14,

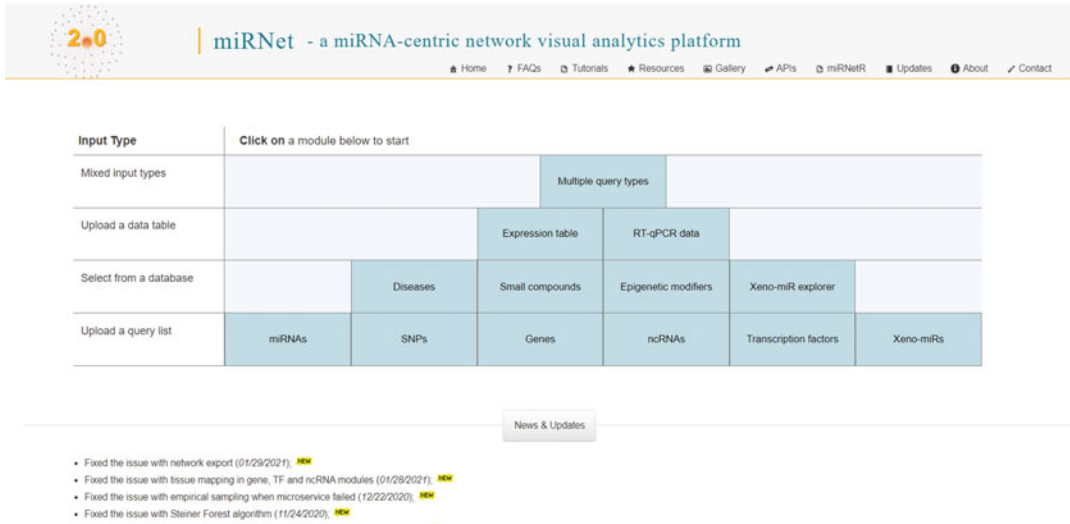


Fig. 1 A screenshot of miRNet 2.0 homepage. The main functions are organized into 13 modules based on input types

The miRNet 2.0 is a user-friendly web-based tool (https://www.mirnet.ca) that allows users to easily create and visually explore miRNA-associated regulatory networks. It was first released in 2016, supporting both computationally predicted and experimentally validated miRNA-target gene network analysis [3], and has since been continuously updated to meet the growing needs from the community [4–6]. The version 2.0 release contains 13 modules which can be divided into 4 different categories based on the input types: (1) upload a query list, (2) select a query list from built-in databases, (3) upload a data table, and (4) mixed input types (see Fig. 1). The miRNet 2.0 adopts a stepwise design concept to guide users through all analysis procedures, including data uploading, network building, network customization, and network visual exploration. Its companion R package, miRNetR, was released recently to complement the web server to allow more flexible or batch analysis for those researchers who are familiar with R language. Additionally, the web application programming interface (API) was also implemented to allow other tool developers to submit their queries programmatically [6].

This chapter contains four sections covering the main features of miRNet 2.0. Subheading 1 focuses on creating and exploring miRNA–gene regulatory networks from a list of miRNAs, including functional enrichment analysis and module analysis. Subheading 2 describes a workflow for network analysis from miRNA RT-qPCR data. Subheading 3 goes through how to construct a composite miRNA network from multiple data types (miRNAs and TFs). Finally, Subheading 4 showcases an example using miRNet 2.0 for miRNA–disease association network analysis.

2 Materials

A personal computer with an Internet connection.

2.1 Browser Requirements

An up-to-date web browser that supports HTML5 with JavaScript enabled, such as Google Chrome (v50+), Firefox (3.0+), and Internet Explorer (9.0+).

2.2 Hardware Requirements

We recommend a ≥ 2 GHz CPU, 4-GB physical RAM with at least 2 GB free, and a minimum of a 15-inch screen with a screen resolution of 1280 × 800 or higher. A mouse with scrolling support is highly recommended for network visualization.

3 Methods

3.1 Network Analysis and Visualization from a List of miRNAs

1. *Starting up*. Go to the miRNet 2.0 homepage (www.mirnet.ca) (*see* Fig.1). There are 13 modules corresponding to 13 different input types (*see* **Note 1**). For miRNA list input, click the "miRNA" button to enter the data upload page.
2. *Data upload*. Users need to specify the organism, miRNA ID type, and target(s) type (*see* **Note 2**). Optionally, users can specify tissue types (human only) and include protein–protein interactions (PPI) or transcription factors to gene mapping (tf2gene). Users can select other targets in addition to genes, such as competing endogenous RNAs (lncRNAs, circRNAs, sncRNAs, etc.), which regulate each other by competing for shared miRNAs [7].
3. Enter a list of miRNA IDs (*see* **Note 3**) with one entry per line. In this case, we use the first example dataset. Click "Try Examples" button at the bottom-left corner of the page. In the pop-up dialog, choose "miRNA list 1," and click "Yes" button. The parameters and the list of miRNAs will be automatically entered for the example data (*see* Fig. 2).
4. Click the "Submit" button to upload. An "OK" message will pop up at the top-right corner of the page indicating that the miRNA list has been uploaded successfully. The "Proceed" button at the bottom-right corner is now activated. Click this button to proceed (*see* **Note 4**).
5. *Network building*. The Network Building page will be displayed after a few seconds (*see* Fig. 3). The interaction tables and network summaries will show on the page. In some cases, multiple isolated networks will be generated, with a big "continent" containing most of the queries and several small "islands" containing one or a few queries. These networks will be available for visual exploration in the next step.

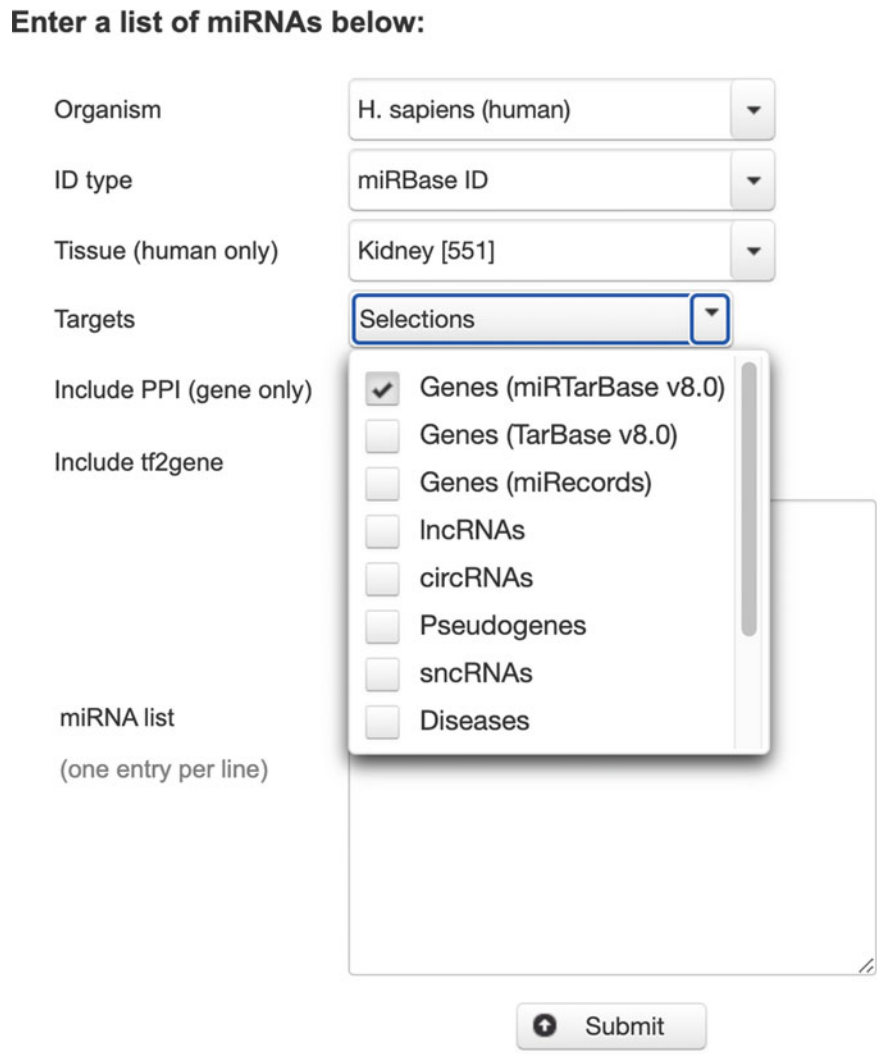


Fig. 2 A screenshot of the miRNA upload page using the example dataset

6. Users can either download/browse the interaction tables by clicking the associated download icon (a downward arrow) or the "Browse" link, or they can explore the results in a network context. Click the "Proceed" button at the bottom to view the network.
7. *(Optional) Network filtering.* When a network is too large (i.e., > 2500 nodes), users can use "Network Tools" in the network builder page to reduce the size of the network based on topological measures (degree and betweenness), shortest paths, to compute minimum network or to manually filter the network based on a given list.

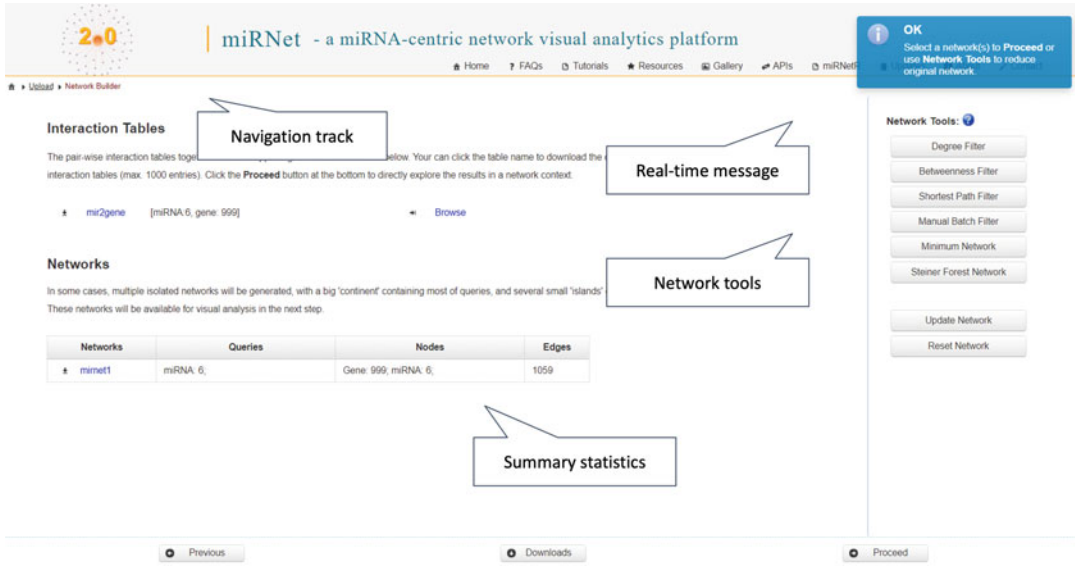


Fig. 3 A screenshot of the "Network Builder" page. The top-right corner shows a real-time message indicating the current status, and it also provides suggestions for the next step. The top-left corner shows the navigation track with the current page highlighted in red. Users can click the links to go back to the corresponding page. The buttons on the bottom allow users to proceed to the next page, return to the previous page, or download the results tables. The "Network Tools" contains various functions for network refinement

A. *Degree filter*. The degree of a node is the number of connections it has to other nodes. Nodes with higher degree values are "hubs" in a network [8]. Click the "Degree Filter" button to bring up the dialog. Users can indicate the node types and enter the degree cutoff value to remove the nodes with a degree value lower than a specified threshold.

B. *Betweenness filter*. The betweenness centrality measures the number of shortest paths going through a node. Nodes with higher betweenness value are considered as "bottlenecks" in a network [8].

C. *Shortest path filter*. The "Shortest Path Filter" is designed to reduce the "hair-ball" effect in network visualization. The goal is to extract a subnetwork by computing pair-wise shortest paths between all major hub nodes and then remove the nodes that are not on the shortest paths.

D. *Compute minimum network*. You can also use "Minimum Network" or "Steiner Forest Network" tool to construct a minimally connected network that contains all query nodes (seeds).

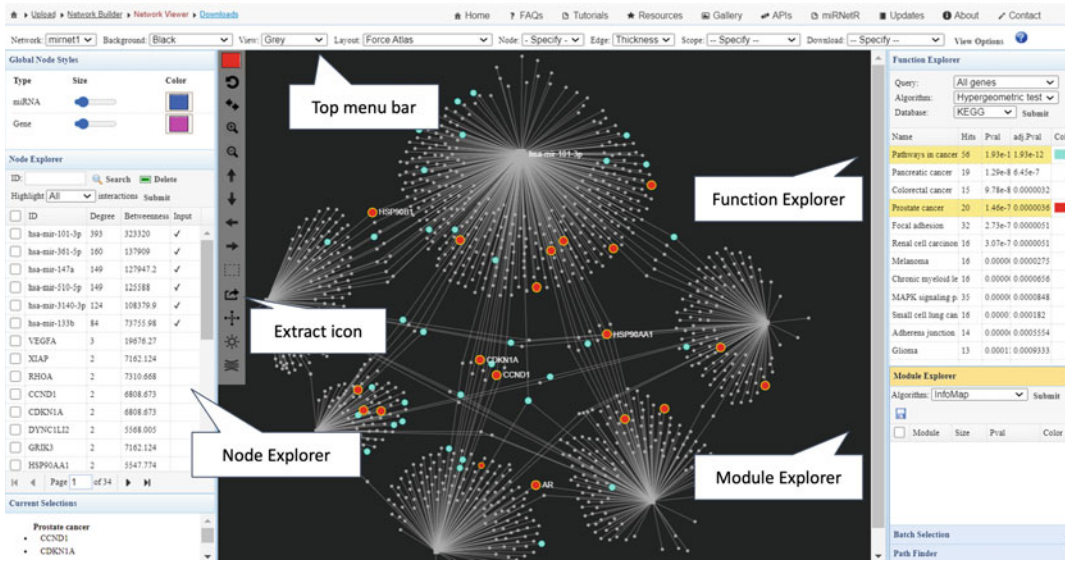


Fig. 4 A screenshot of the "Network Viewer" page demonstrates the main organization of various components. The current network is displayed in the "Grey" view option, with two enriched KEGG pathways (pathways in cancer and prostate cancer pathways) highlighted in blue and red. Clicking the extract icon will extract subnetworks that represent the interaction patterns among highlighted nodes

E. *Manual batch filter*. You can manually filter the network by either excluding or including a given list. Note that after filtering, any resulting orphan nodes will also be removed from the network.

8. *Visual exploration of the network*. On the network view page, the central network viewer displays the largest network by default, surrounded by various panels and tool menus (*see* Fig. 4). Users can zoom, select, or highlight nodes, drag and drop nodes, or extract the highlighted nodes. Mouse events are defined in Table 1.

   A. *Zooming and positioning*. Zoom the network using the mouse scroll function; move the network by clicking on an empty area within the network, and drag it to a new position; click the "Auto fit" icon at the left vertical tool bar (the third position from the top) to fit the entire network to the current window size.

   B. *Node selection and highlighting*. Click the color palette at the top of the left vertical tool bar to bring up the dialog. Select a color and then click the "Choose" button to set the color and close the dialog. Move the cursor over a node, and when the label becomes visible, double-click on the node. The node is highlighted with the selected color.

   C. *Node drag and drop*. Move the cursor over a node, and when the label becomes visible, click and drag the node to a new position.

Table 1
Mouse events in network visualization and exploration

Scope	Purpose	Mouse event
Single node	Display node label	Mouse over the node
	Display node details	Single-click the node
	Adjust color	Set the highlight color and then double-click the node
	Increase size	Repeatedly click the node
	Change position	Mouse over the node until its label appears, and then drag the node
Node-neighbors	Change positions	Mouse over the central node until its label shows up, and then drag the node
	Change colors	Set the highlight color, and then double-click the central node
	Change sizes	Repeatedly click the central node
Highlighted nodes	Change positions	Mouse over any highlighted node until its label appears, and then drag the node
	Change colors	Set the highlight color and then reperform highlighting
	Change sizes	Go to the node tab under view options, and adjust the node size for "highlighted nodes"
Network	Zoom	Mouse over any empty area and then scroll
	Change positions	Mouse over any empty area and then drag
	Change sizes	Go to the node tab under view options, and adjust the node size for "all nodes"

D. *Extract the highlighted nodes*. Click the "Reset" icon at the left vertical tool bar to return to the default view. Set the "Scope" option on the top menu bar to "Node-neighbors"; double click the *hsa-miR-3140-3p* node to highlight the node and its neighbors. Click the "Extract" icon at the left vertical tool bar, and the *hsa-miR-3140-3p* node and its neighbor nodes are now extracted as a new subnetwork ("module1"). Note that you can switch back to the main network ("mirnet1") by selecting from the "Network" dropdown menu at the top menu bar.

9. *Customizing the network*. The top menu bar also provides options to customize the network. Users can change the background color, view option, network layout, node, or edge styles.

A. *Change background color with the "Background" option*. The default background color is black. Users can change the background color to "gradient (light)", "white", or "gradient (dark)" or customize the background to your preferred colors.

B. *Change view style with the "View" option*. The default view is "Topology." Users can switch to "Expression" or "Grey" view.

C. *Change layout with the "Layout" option*. The miRNet 2.0 currently supports >10 types of network layout algorithms, including *Force-Atlas*, *Fruchterman-Reingold*, *Circular*, *Graphopt*, *Large Graph*, *Random*, *Circular Bipartite/Tripartite*, *Linear Bipartite/Tripartite*, *Concentric*, and *Backbone*. The latter four types of layout are designed for complex networks containing multiple node types (miRNAs, genes, TFs, lncRNAs, etc.).

D. *Change node styles with the "Node" option*. Users can customize node label, color, size, and shape. The node labels could be hidden to better visualize the network structure. The node color option allows users to apply predefined color schemes to nodes based on their network topological values. The node sizes and shapes could be adjusted for either all nodes or highlighted nodes. Note the "Global Node Styles" panel at the top left corner can also be used to change the node sizes and colors based on different node types.

E. *Change edge styles with the "Edge" option*. The edge opacity, thickness, and color can be modified. Curved and thin edges are ideal for visualizing larger networks, while straight and thick edges are more suitable for viewing smaller networks.

10. *Node visualization and manipulation*. The "Node Explorer" table on the left panel displays all nodes in the current network, including their IDs, degree, and betweenness values (*see* Fig. 4). For seed nodes, a check mark or their expression values (if provided) will appear under the "Input"/"Expr." column. The node tables could be sorted by clicking on the column headers. Users can also view, search, and delete nodes or highlight hub nodes using the "Node Explorer."

A. *Node viewing*. Click a node ID to view it in the network.

B. *Node search*. Enter a node ID and click "Search" to locate it in the network.

C. *Node deletion*. Select a node to be deleted and then click the "Delete" button. Please note that node deletion is a computationally intensive task. It affects the selected node as well as its connected nodes. After node deletion, the system will reconstruct the network based on the new node list. Alternatively, users can right click a node directly on the network and select the "Delete node" option.

11. *Functional enrichment analysis*. The "Function Explorer" at the top right panel supports functional enrichment analysis. miRNet 2.0 supports four query types (all genes, highlighted genes, all miRNAs, highlighted miRNAs), two enrichment algorithms (hypergeometric tests and empirical sampling), and nine annotation libraries (four gene-set libraries and six miRNA-set libraries) for functional enrichment analysis (*see* **Note 5**). In this example, select "All genes," "Hypergeometric test," and "KEGG" database. Click the "Submit" button to perform the enrichment analysis. The enriched pathways are displayed in the resulting table ranked by their raw *P* values or empirical *P* values from the empirical sampling algorithm (*see* **Note 6**).
12. Use the color palette at the top left tool bar to set a new color (e.g., blue). Click "Pathways in cancer." The corresponding nodes will be highlighted in blue, and their sizes will be slightly increased (*see* Fig. 4).
13. Set a new color (e.g., red) and then select "Prostate cancer" pathway (*see* Fig. 4).
14. To extract the module containing all nodes involved in those two pathways, click the "Extract" icon at the vertical tool bar (*see* Fig. 4). The module is displayed in a default style and is listed as "module1" at the top menu bar. Users can follow **Step 9** to perform customization.
15. *Save the results*. Users can save the current module as portable network graphics (PNG), scalable vector graphics (SVG), or GraphML format. For example, the GraphML file can be imported to Cytoscape for visualization and exploration. The enrichment results can also be saved as a comma-separated value (.csv) file.
16. *Module detection and extraction*. Modules are tightly clustered subnetworks with more internal connections than expected based on random chance. Click on "mirnet1" in the top menu bar to reload the main network.
17. Navigate to the "Module Explorer" at the bottom right. Select "*InfoMap*" algorithm and click "Submit" (*see* **Note 7**). A list of modules will be displayed together with summary statistics about their sizes and *P* values. Click any module to view its nodes highlighted in the currently selected color.
18. The *P* value of a module is calculated using Wilcoxon rank-sum test to compare the number of connections of each node with that of other nodes within the module and with that for nodes outside the module and gives some indication of how significant the connections within a defined module are. We can test whether certain biological functions are enriched in these modules. In this case, select the top two modules.

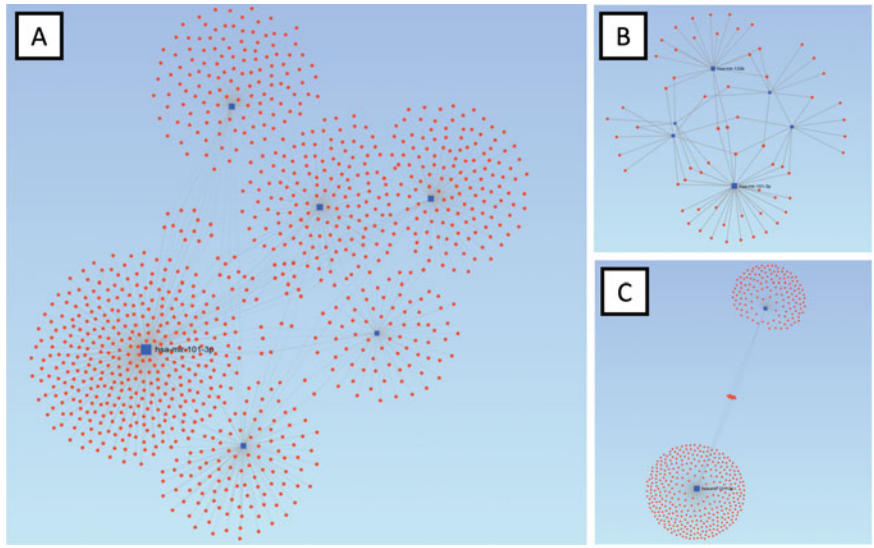


Fig. 5 Network customization and module extraction. (**a**) The same network as in Fig. 4 but using the Large Graph layout in the light gradient background. (**b**) The "Pathways in cancer" module constructed using the Function Explorer. (**c**) The first two modules identified by the *InfoMap* algorithm using the Module Explorer

19. Extract these two modules by clicking the "Extract" icon in the vertical tool bar on the left. Users can perform the functional enrichment analysis on the nodes within the module. It is reassuring that "Pathways in cancer" and "Colorectal cancer" are among the top pathways in the results. The networks can be further customized as described in **Step 9** in Subheading 3.1. An example output is shown in Fig. 5.
20. *Path finder*. Users can use "Path Finder" to view the connections between any two nodes in the network. Click the "Path Finder" at the bottom right to open the sub-panel. You can either right click on the node of interest to select it as "Source" or "Target," or manually enter the node IDs in "From" and "To" textbox. After clicking the "Submit" button, all shortest paths between these two nodes will be displayed in the resulting table. Click any path to highlight it in the network.

3.2 Network Creation and Visualization from Expression Data

1. *Starting up*. Click the "Home" icon to return to the home page. miRNet 2.0 currently supports three types of expression data—microarray, RNASeq, and RT-qPCR. In this case, we will demonstrate the detailed procedures for miRNA RT-qPCR expression data analysis. Click the "RT-qPCR" button to start.
2. *Data upload*. The "Upload" page provides step-by-step procedures for preparing expression data. Click the "Try Examples" button at the bottom-left corner of the page, and click "Yes" to use our example dataset. An "OK" message will appear, indicating that the data have been successfully uploaded (*see* **Note 8**).

3. *Data annotation*. For the example dataset, the organism is automatically set to *M. musculus* (mouse), and the ID type is set to "miRBase ID." Click "Submit" to perform the annotation.
4. *Data normalization*. "Quantile normalization" is set for the example dataset to enforce the same distributions across all samples. Users can click the "View Data" button to visualize the summary boxplot (*see* **Note 9**).
5. *Differential expression analysis (DEA)*. miRNet 2.0 supports three statistical methods for differential expression analysis, including *limma*, standard t-tests, and non-parametric Mann–Whitney U tests. In addition, users can perform flexible comparisons of interests. In this case, the control group (*Ctrl*) and group A (*GrpA*) are selected for comparison by default. Click "Submit" to perform DEA.
6. *Feature selection*. Users can specify the thresholds for "Adjusted p-value," "Fold change," and "Directions" to select differentially expressed miRNAs/genes. Accept the default and click the "Submit" button to continue (*see* **Note 10**).
7. *Specify a database for network creation*. "Genes (miRTarBase v8.0)" is selected by default. You can choose your target of interests and then click "Submit" to continue.
8. The RT-qPCR expression data has been processed (*see* Fig. 6). Click the "Proceed" button at the bottom of the page to the "Network Builder" page.

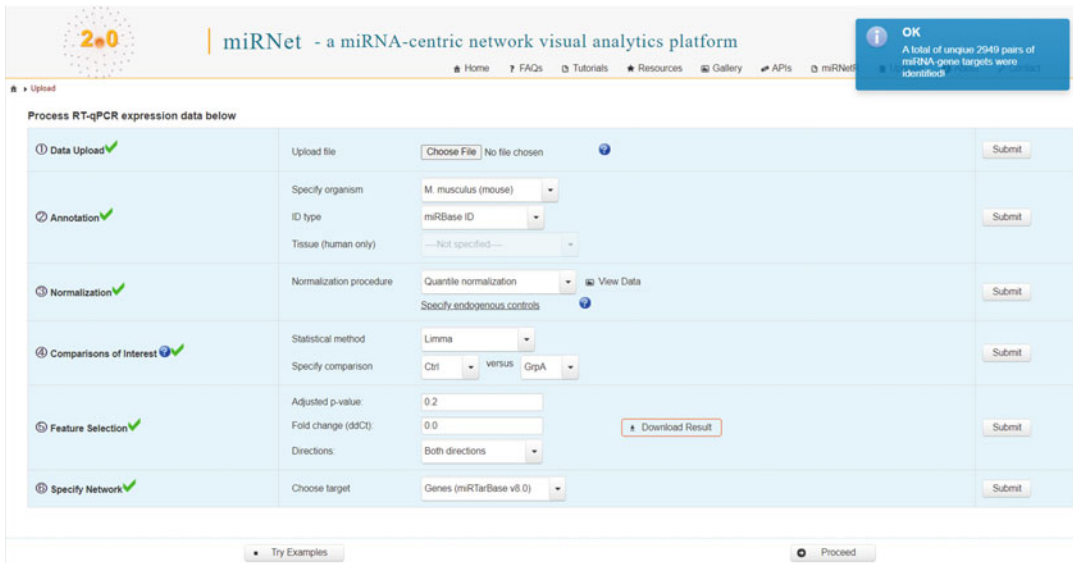


Fig. 6 A screenshot of the RT-qPCR data upload page when all data analysis steps have been completed

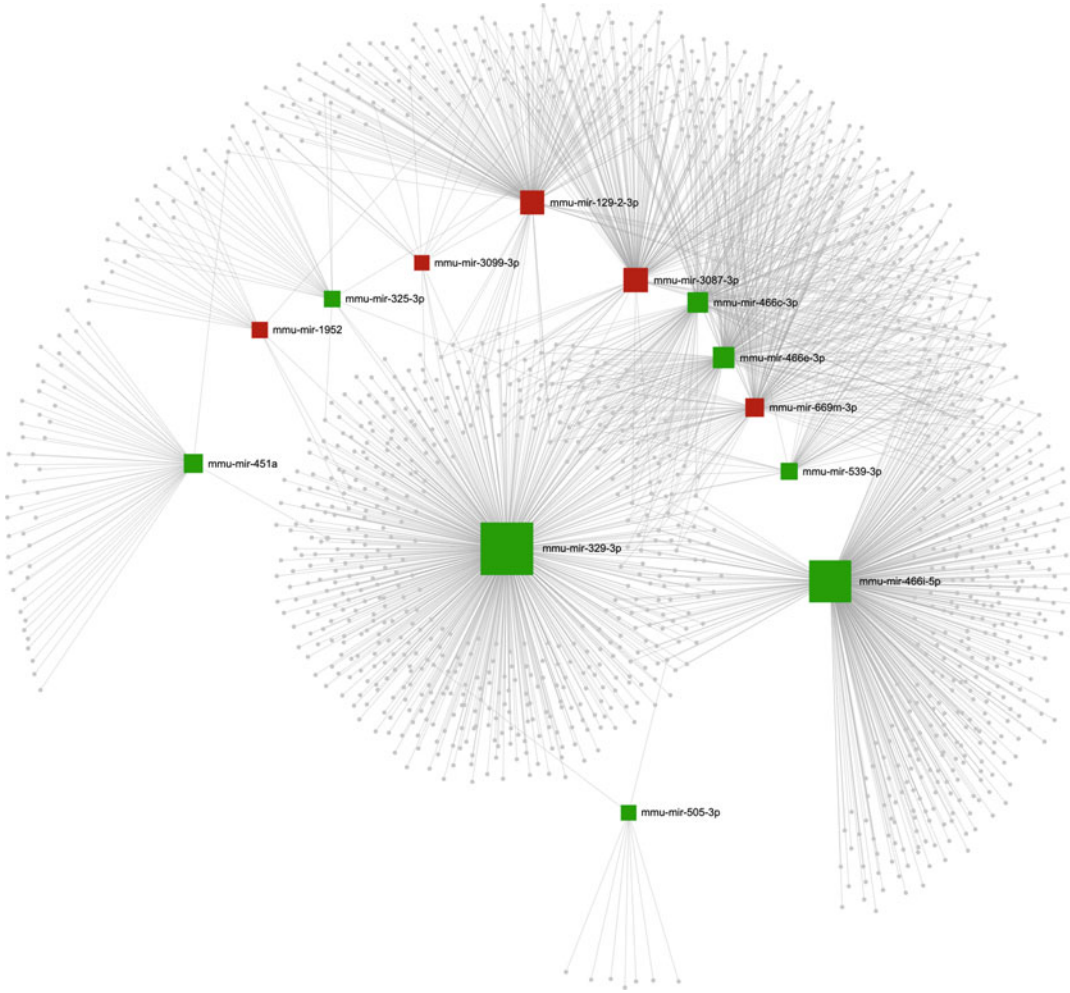


Fig. 7 A miRNA regulatory network in the "Expression" view mode (red for upregulated and green for downregulated miRNAs) based on the concentric layout where nodes are arranged in concentric circles around a focal node

9. Follow **Steps 5**–7 in Subheading 3.1 to build and/or filter the network in the Network Builder.
10. Follow **Steps 8**–**20** in Subheading 3.1 to visually explore and/or customize the network, as well as to perform functional enrichment analysis or module analysis. Figure 7 shows the resulting network in the expression view (red for upregulated and green for downregulated miRNAs) with white background in a concentric layout where nodes are arranged in concentric circles around the focal node. The order of the circles reflects the degree level of their interactions, which facilitates a better understanding of how the focal node relates to the rest of the network. The analysis results showed that there were five upregulated DE-miRNAs and eight downregulated DE-miRNAs. We can clearly see that among the five upregulated DE-miRNAs, mmu-miR-3087-3p has the highest degree

values (degree = 167), whereas among the eight downregulated DE-miRNAs, mmu-miR-329-3p has the highest degree values (degree = 556), suggesting that they might be important regulators for further analysis.

3.3 Creating and Visualizing a Composite miRNA Network

miRNet 2.0 allows users to flexibly integrate different molecular types to create composite networks. There are three typical scenarios: (1) starting from a list of miRNAs with one or multiple targets to build miRNA–target interaction networks as primary networks and then adding PPI networks (i.e., known interactions among target genes) and/or TF–gene networks; (2) starting from a list of genes to build miRNA–gene interaction networks as primary networks and then adding PPI networks and/or TF–miRNA networks; and (3) starting from multiple types of molecules (miRNAs, genes, lncRNAs, circRNAs, pseudogenes, sncRNAs, TFs, small compounds, diseases, or epigenetic modifiers) and connecting them based on known interactions. This section will describe the third scenario on miRNA composite network creation and visualization by building a TF–miRNA co-regulatory network from a list of miRNAs and TFs. Several advanced features will be demonstrated, including network filtering, applying different network layouts, edge bundling, customizing edge color, and network export.

1. *Starting up*. Go to the miRNet 2.0 homepage (www.mirnet.ca), and click the "Multiple query types" button in the top center (*see* Fig. 1). In the pop-up dialog, first specify the organism as "*H. sapiens (human)*" (default), and select "miRNAs" and "Transcription factors" check box (*see* Fig. 8). Optionally, users can specify tissue types (human only). In this protocol, we will leave it as "Not specified" (*see* **Note 11**). Click "OK" to go to the "Upload" page.
2. *Data upload*. The "Upload" page contains two tabs corresponding to the user selections (miRNAs and TFs). Here, we will use a built-in example. Click the "Try Example" link at the miRNA tab. In the pop-up dialog, click the "Yes" button to upload the example miRNA list 2. The example data come from a multiple sclerosis (MS) study aiming to identify the role of miRNA and TF co-regulatory networks in the pathogenesis of MS [9]. miRNet 2.0 will set the parameters for this example miRNA list. Repeat the previous two steps in the transcription factor upload tab using the example TF list 2. Note the corresponding checkbox at the bottom will be checked after the specified data type is uploaded. When all uploaded data types are checked (keep the "Include PPI" option unchecked), users can click "Proceed" to the "Network Builder" page (*see* Fig. 9).

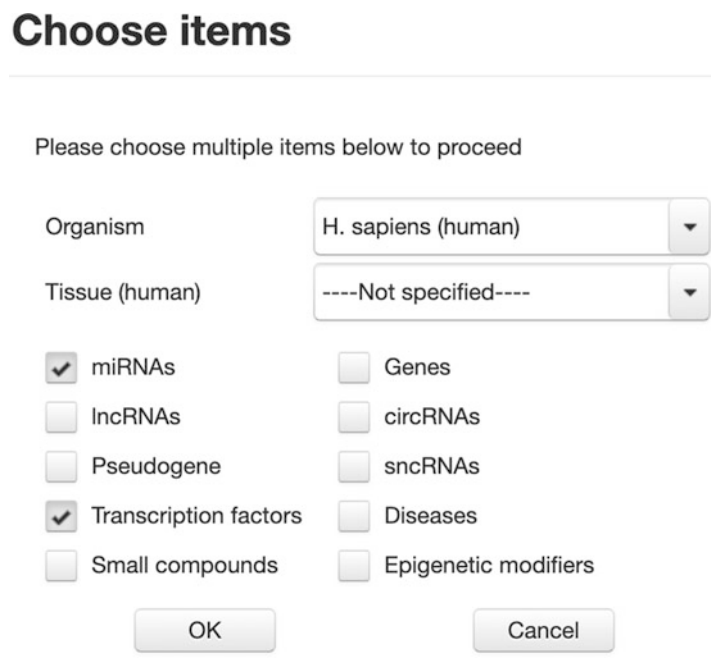


Fig. 8 A screenshot showing the dialog for the "Multiple query types" module. Users can choose multiple items to start their analysis

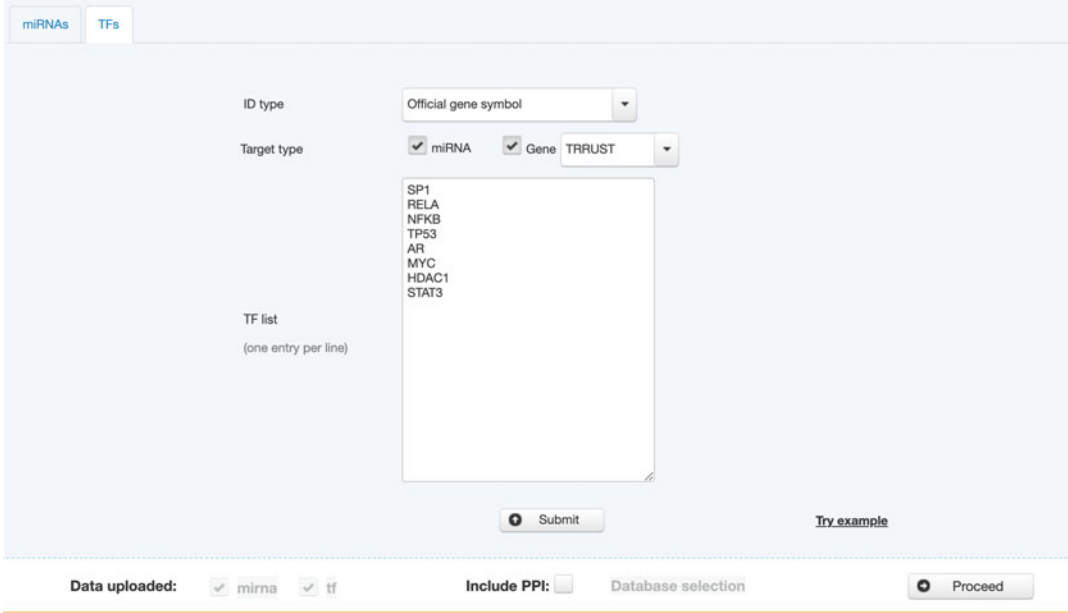


Fig. 9 A screenshot of the upload page in "Multiple query types" module, which contains two tabs corresponding to the user selections (miRNAs and TFs)

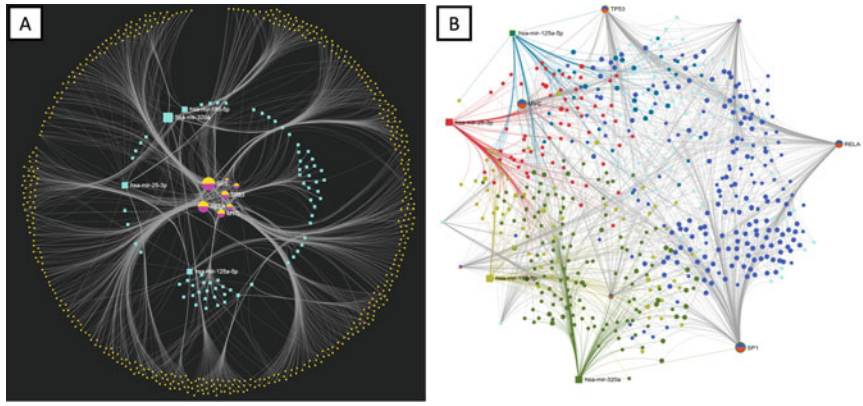


Fig. 10 Two customized miRNA-TF-gene networks with edge bundling in (**a**) circular tripartite and (**b**) Force Atlas layouts. The circular tripartite layout illustrates various interactions between TFs (inner zone), miRNAs (middle layer), and genes (outer layer). In the Force Atlas layout, it is easy to see that a few hub nodes are located at the peripheral part of the network. We can also see their direct target genes of several hub miRNAs. Note that TF nodes are represented as pie charts as they are annotated as both genes and TFs (dual roles) in this graphical representation

3. *Network building and filtering*. After a few seconds, the summary statistics for the resulting interaction tables and networks will be displayed. In this case, three pairwise interaction tables were generated, including mir2gene, tf2gene, and tf2mir. The single network contains 2865 nodes (TF: 7; Gene: 2705; miRNA: 153) and 3712 edges. This network may be too big to make sense of. Therefore, we filter the network by using a degree cutoff as 1.0 (i.e., excluding all terminal nodes) to reduce the network size. The filtered network now contains 575 nodes and 1422 edges. Click the "Proceed" button to view the filtered network.
4. *Network customization*. (1) To change the network layout, click the "Layout" drop-down menu at the top menu bar. Select the "Circular Bipartite/Tripartite" layout to rearrange the nodes in a three-layered layout. It illustrates various interactions between TFs (inner zone), miRNAs (middle layer), as well as genes (outer layer) (*see* Fig. 10a). Alternatively, the "Force Atlas" layout is well suited for visualizing large networks based on a force-directed continuous layout algorithm (*see* Fig. 10b) [10]. (2) To reduce edge crossing in the network, click the "Apply edge bundling" icon at the vertical tool bar (*see* **Note 12**). (3) To clearly view a node of interest and its direct interacting partners, (a) change the scope to "Node-neighbors"; (b) set up the highlighted color by following **Step 8**(b) in Subheading 3.1; (c) double-click on the node. (4) To export the network, users can simply click the option

under the "Download" drop-down menu at the top menu bar. In this case, the resulting miRNA-TF co-regulatory networks depicted a few critical hub nodes (*see* Fig. 10). Among the nodes with the largest degree values, miR-125a-5p has been frequently associated with MS, and *SP1* has been reported to participate in the transcriptional regulations of MS, specifically in modulating the autoimmune response [9]. It is reasonable to hypothesize that these miRNAs and TFs may act cooperatively in the regulatory process of MS.

3.4 Network Creation and Visualization of Other Types of miRNA Networks

Users can also create networks from a list of small compounds, diseases, epigenetic modifiers, or xeno-miRNAs by selecting from the built-in databases available in miRNet 2.0. In this section, we will use "Diseases" as an example to show the procedures.

1. *Starting up*. Go to the miRNet 2.0 home page (www.mirnet.ca), and click the "Diseases" button (*see* Fig. 1).
2. *Data upload*. Users need to manually search and select their diseases of interest from the "Available" panel on the left to the "Selected" panel on the right. In this case, we selected "*Carcinoma, Colon*," "*Carcinoma, Cervical*," and "*Carcinoma, Gastric*" (*see* Fig. 11). After disease selection, click the "Submit" and then "Proceed" button at the bottom of the "Upload" page.

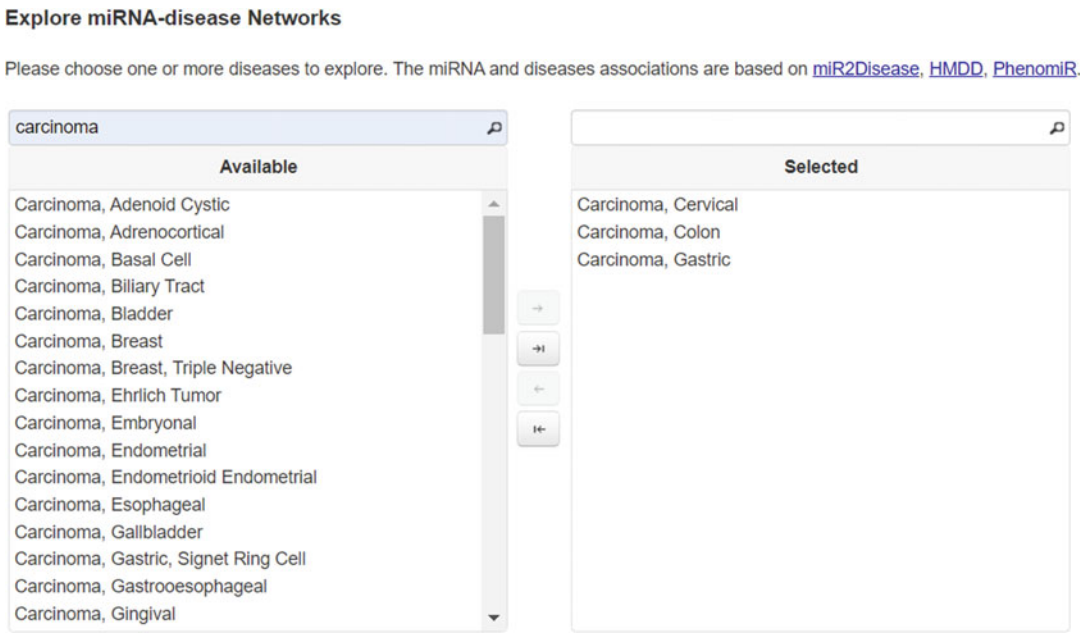


Fig. 11 A screenshot of the upload page for the Disease module. Users can directly drag and drop items of interest from the left "Available" panel to the "Selected" panel on the right

3. Follow **Steps 5**–**7** in Subheading 3.1 to build and/or filter the network in the Network Builder.
4. Follow **Steps 8**–**20** in Subheading 3.1 to visually explore and/or customize the network, as well as to perform functional enrichment analysis or module analysis.

4 Notes

1. Users can (a) upload a list of miRNAs, ncRNAs, genes, TFs, or SNPs; (b) select a list from our built-in databases such as diseases, small compounds, epigenetic modifiers, etc.; (c) upload a miRNA or gene expression table generated from RT-qPCR, microarray, or RNAseq; or (d) upload multiple queries of different input types.
2. miRNet 2.0 currently integrates data from 14 different miRNA databases—TarBase v8.0 [11], miRTarBase v8.0 [12], miRecords [13], and miRanda [14] for miRNA–gene interactions; starBase v2.0 [15] for miRNA–lncRNA, miRNA–circRNA, miRNA–pseudogene, and miRNA–sncRNA interactions; miR2Disease [16], HMDD v3.0 [17], and PhenomiR [18] for miRNA–disease associations; SM2miR [19] and PharmacomiR [20] about the influences of small compounds on miRNA expressions as well as linking miRNAs and drug effects; EpimiR [21] for studying mutual regulation between miRNAs and epigenetic modifiers; TransmiR [22] for miRNA–TF interactions; and ADmiRE [23] and PolymiRTs [24] for SNP data in miRNA genes and miRNA-binding sites. It currently supports ten organisms: *H. sapiens* (human), *M. musculus* (mouse), *R. norvegicus* (rat), *B. taurus* (cattle), *S. scrofa* (pig), *G. Gallus* (chicken), *D. rerio* (zebrafish), *C. elegans* (roundworm), *D. Melanogaster* (fruitfly), and *S. mansoni (blood fluke)*.
3. miRNet 2.0 can automatically recognize different versions of miRBase IDs (v15-v22), as well as convert miRNA precursors to their mature forms based on miRBaseConverter [25].
4. Please note that not all miRNAs will have interaction information in the underlying knowledge base. The data upload will give an error if the database search returns no hits.
5. The gene set libraries include gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and disease ontology databases. The miRNA set libraries are based on TAM 2.0 database [26], which includes miRNA–function, miRNA–disease, miRNA–TF, miRNA–cluster, miRNA–family, and miRNA–tissue set libraries.

6. The empirical sampling method is employed to estimate the null distribution of the target genes as selected based on the input miRNAs against a given miRNA–gene database. The process can be divided into three steps: (1) A list of miRNAs of the same size is arbitrarily chosen from all the miRNAs with known targets from the database. (2) The functional annotations (e.g., GO or KEGG) are then performed for the list. (3) The process is repeated 1000 times (default). (4) Compare the hits in each GO term or KEGG pathway, and the empirical *P* values are calculated as the proportion of overlaps (with GO terms or KEGG pathways) from the 1000 random process that is equal or larger than the original ones. A *P* value <0.001 will be reported if no results from the random process are better than the original miRNA list [27].
7. miRNet 2.0 currently offers three different approaches for module detection—the *InfoMap*, *WalkTrap*, and *Label Propagation* algorithms based on the *igraph* R package [28].
8. The required format is a tab-delimited text (.txt) file. The sample names must be in the first line, followed by the class labels with a new line beginning with "#CLASS:".
9. Several well-established normalization algorithms are provided, including log-transformation, quantile normalization for microarray data based on *limma* [29], trimmed mean of M values (TMM) in combination with a dispersion measure for RNAseq data based on edgeR [30], as well as five widely used normalization methods for RT-qPCR data based on HTqPCR [31].
10. After performing feature selection, the results can be downloaded by clicking the "Download Result" button.
11. The purpose of a tissue filter is to understand the regulatory roles of tissue-specific miRNAs as it has been reported that tissue-specific miRNAs are commonly implicated in diseases associated with specific tissues [32]. miRNet 2.0 will return miRNAs specifically expressed in particular tissues when the tissue type is specified. The experimentally validated tissue-specific miRNA annotations are collected from TSmiR [32] and IMOTA [33], and the exosomal miRNA annotations are based on ExoCarta [34].
12. Edge bundling may take a while to run because it is a computationally intensive task. Node dragging will be disabled after applying edge bundling. Click the icon again to revert to the original network.

Acknowledgments

Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, NSERC-CREATE-MATRIX Scholarship, and Canada Research Chairs (CRC) Program.

References

1. Bracken CP, Scott HS, Goodall GJ (2016) A network-biology perspective of microRNA function and dysfunction in cancer. Nature Reviews Genetics 17. https://doi.org/10.1038/nrg.2016.134
2. Anastasiadou E, Jacob LS, Slack FJ (2017) Non-coding RNA networks in cancer. Nature Reviews Cancer 18. https://doi.org/10.1038/nrc.2017.99
3. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia JJ (2016) miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis. Nucleic Acids Res 44(W1):W135–W141
4. Fan Y, Habib M, Xia J (2018) Xeno-mirNet: a comprehensive database and analytics platform to explore xeno-miRNAs and their potential targets. PeerJ 2018. https://doi.org/10.7717/peerj.5650
5. Fan Y, Xia J (2018) miRNet—functional analysis and visual exploration of miRNA–target interactions in a network context. In: Computational cell biology. Springer, pp 215–233
6. Chang L, Zhou G, Soufan O, Xia J (2020) miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res 48(W1):W244–W251. https://doi.org/10.1093/nar/gkaa467
7. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP (2011) A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? Cell 146(3):353–358
8. Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21(9):1010–1024
9. Nuzziello N, Vilardo L, Pelucchi P, Consiglio A, Liuni S, Trojano M, Liguori MJ (2018) Investigating the role of MicroRNA and transcription factor co-regulatory networks in multiple sclerosis pathogenesis. Ijoms 19(11):3652
10. Jacomy M, Venturini T, Heymann S, Bastian M (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 9(6):e98679. https://doi.org/10.1371/journal.pone.0098679
11. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. Nucleic Acids Res 46(D1):D239–D245
12. Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, Tang Y, Chen Y-G, Jin C-N, Yu Y (2020) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. Nucleic Acids Res 48(D1):D148–D154
13. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T (2009) miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res 37(suppl_1):D105–D110
14. Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11(8):R90
15. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H (2014) starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 42(D1):D92–D97
16. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37(suppl_1):D98–D104
17. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q (2019) HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Res 47(D1):D1013–D1017
18. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol 11(1):R6
19. Liu X, Wang S, Meng F, Wang J, Zhang Y, Dai E, Yu X, Li X, Jiang W (2013) SM2miR:

a database of the experimentally validated small molecules' effects on microRNA expression. Bioinformatics 29(3):409–411

20. Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N (2014) Pharmaco-miR: linking microRNAs and drug effects. Brief Bioinform 15(4):648–659
21. Dai E, Yu X, Zhang Y, Meng F, Wang S, Liu X, Liu D, Wang J, Li X, Jiang W (2014) EpimiR: a database of curated mutual regulation between miRNAs and epigenetic modifications. Database (Oxford) 2014(2014):bau023. https://doi.org/10.1093/database/bau023
22. Tong Z, Cui Q, Wang J, Zhou Y (2019) TransmiR v2. 0: an updated transcription factor-microRNA regulation database. Nucleic Acids Res 47(D1):D253–D258
23. Oak N, Ghosh R, Kl H, Wheeler DA, Ding L, Plon SE (2019) Framework for microRNA variant annotation and prioritization using human population and disease datasets. Hum Mutat 40(1):73–89
24. Bhattacharya A, Ziebarth JD, Cui Y (2014) PolymiRTS database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res 42(D1):D86–D91
25. Xu T, Su N, Liu L, Zhang J, Wang H, Zhang W, Gui J, Yu K, Li J, Le TD (2018) miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. BMC bioinformatics 19(19):179–188
26. Li J, Han X, Wan Y, Zhang S, Zhao Y, Fan R, Cui Q, Zhou Y (2018) TAM 2.0: tool for MicroRNA set analysis. Nucleic Acids Res 46 (W1):W180–W185
27. Bleazard T, Lamb JA, Griffiths-Jones S (2015) Bias in microRNA functional enrichment analysis. Bioinformatics 31(10):1592–1598
28. Csardi G, Nepusz T (2006) The igraph software package for complex network research. Inter J Complex Syst 1695(5):1–9
29. Smyth GK (2005) Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor. Springer, pp 397–420
30. 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
31. Dvinge H, Bertone P (2009) HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. Bioinformatics 25(24):3325–3326
32. Guo Z, Maki M, Ding R, Yang Y, Zhang B, Xiong L (2014) Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. Sci Rep 4: 5150. https://doi.org/10.1038/srep05150
33. Palmieri V, Backes C, Ludwig N, Fehlmann T, Kern F, Meese E, Keller A (2018) IMOTA: an interactive multi-omics tissue atlas for the analysis of human miRNA-target interactions. Nucleic Acids Res 46(D1):D770–d775. https://doi.org/10.1093/nar/gkx701
34. Mathivanan S, Simpson RJ (2009) ExoCarta: a compendium of exosomal proteins and RNA. Proteomics 9:4997–5000. https://doi.org/10.1002/pmic.200900351

Chapter 15

Modeling Plant Transcription Factor Networks Using ConSReg

Qi Song and Song Li

Abstract

Plants have developed complex regulatory programs to respond to various environmental stress such as heat, drought, and cold. Systematic understanding of these biological processes depends on robust construction of regulatory networks which encodes interactions between transcription factors and target genes. In this chapter, we present a computational tool ConSReg, which predicts regulatory interactions using ATAC-seq, DAP-seq, and expression data. By using expression data generated under a specific environmental stress, ConSReg can reconstruct an interpretable, weighted, and stress response-specific regulatory network.

Key words Machine learning, Regulatory network, Plant environment stress

1 Introduction

Plants have developed complex regulatory programs to cope with adverse environmental conditions. For example, abscisic acid (ABA) signaling pathway is involved in multiple environmental stress such as cold [1], heat [2], drought [3], and salinity [4]. Understanding the stress response network, which consists of interactions between transcription factors (TFs) and target genes, would facilitate selection of candidate genes for reengineering plants to combat various environmental stress. With the accumulation of abundant genomic data sets from model plant species *Arabidopsis thaliana*, reconstruction of such networks has become feasible. In a recent publication [5], we presented a computational framework ConSReg (condition-specific regulation) that integrates ATAC-seq (assay for transposase-accessible chromatin using sequencing), DAP-seq (DNA affinity purification sequencing), and RNA-seq data to reconstruct *Arabidopsis* stress response networks under different environmental conditions. Other existing methods either require multiple RNA-seq samples to infer

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_15,

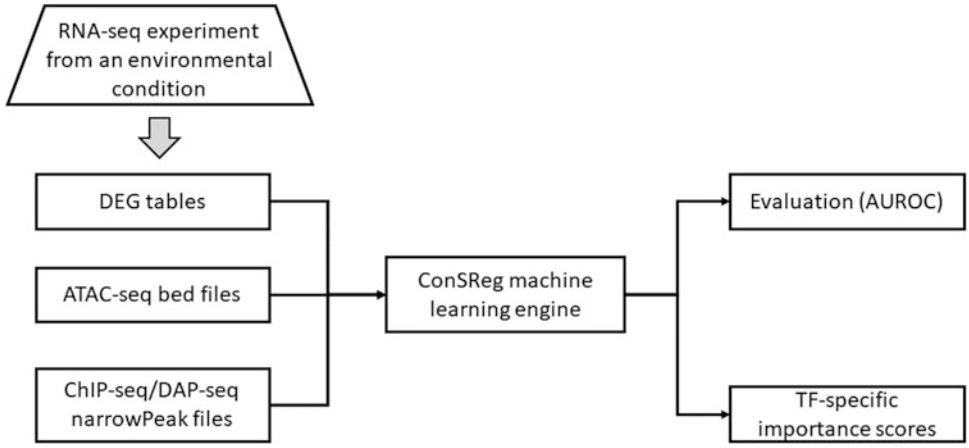


Fig. 1 Schematic flowchart of ConSReg pipeline. ConSReg takes DEG tables, ATAC-seq bed files, and ChIP-seq/DAP-seq narrowPeak files as inputs. ConSReg will map ChIP-seq/DAP-seq TF binding sites to the nearest genes and identify the overlapped regions between the binding sites and the ATAC-seq regions, which produces activity score for each TF. Activity scores are further used with DEG information to perform machine learning classification task. Based on the trained machine models, users may perform evaluation using AUROC values (upper right box) or rank all TFs for a given condition (lower right box)

regulatory associations, or they only produce regulatory interactions independent of environmental conditions. In contrast, ConSReg can infer condition-specific regulatory associations with only two RNA-seq samples: one generated under a certain condition (e.g., environmental stress as treatment) and one generated as a control. This may greatly facilitate the users because RNA-seq experiment with small sample size is not uncommon (*see* Fig. 1). Additionally, the results generated from ConSReg are succinct and interpretable, which only consist of a three-column table with TF name column, target gene name column, and importance score column that represents importance of each TF with respect to the condition. Each output table indicates the ranking of TFs in each treatment VS control from the differential expression analysis. While the previous publication mainly focused on presenting algorithm details, evaluations, and case studies for ConSReg framework, this chapter aims at providing a step-by-step user protocol to guide readers interested in using ConSReg. ConSReg is available as a Python package on GitHub: https://github.com/LiLabAtVT/ConSReg.

2 Materials

2.1 Computer and Operating System Environment

HPC or PC, Linux preferred**,** Python (version ≥3.6) and R. In this protocol, we will only demonstrate the use of ConSReg on Linux (CentOS 8.2.2004). Other distributions of Linux systems could work in a similar manner.

2.2 Installation

ConSReg depends on multiple Python and R packages. To simplify installation steps, we recommend using package manager to install all dependencies in one shot. Anaconda is a package management system primarily designed to manage dependencies for Python. With recent updates, Anaconda also has enabled package management for R. In this tutorial, we will use the steps below to install all dependencies with Anaconda.

1. Retrieve Anaconda from the official website (https://www.Anaconda.com/), and install the version for your operating system (OS). Anaconda is free for individual users with non-commercial purposes.
2. Once Anaconda is installed, open the Linux terminal (*see* **Note 1**), and use the following commands to install ConSReg. In this example, "conda create" is the command to create a new environment named "consreg," in which the ConSReg package and all its dependencies will be hosted. You may also use other names.

```
conda create -y -n consreg python = 3.6
conda activate consreg
conda install -y -c bioconda --no-channel-priority bioconductor-chipseeker
conda install -y --no-channel-priority r-base r-essentials
conda install -y --no-channel-priority -c Conda-forge r-gglasso r-rrf r-devtools
pip install ConSReg
```

3. After ConSReg is installed into the new environment, use the following commands to activate/deactivate the new environment and Python (*see* **Note 2**).

```
conda activate consreg # activate the environment
conda deactivate # deactivate the environment
```

2.3 Prepare Input Data for ConSReg

ConSReg can take three types of genomic data as inputs:

1. *Open chromatin position information from ATAC-seq data*, represented as a bed file, which should include three columns: column 1 represents chromosome name (e.g., chr1, chr2); column 2 represents start position of an open chromatin region; and column 3 represents end position of an open chromatin region (see also the sample data files in the ConSReg GitHub repository: https://github.com/LiLabAtVT/ConSReg/tree/master/data/atac_seq_all_peaks).
2. *TF binding location information from DAP-seq/ChIP-seq data*, represented as narrowPeak files, which should contain the same types of columns stated for ATAC-seq data (see also the sample

data files in the ConSReg GitHub repository: https://github.com/LiLabAtVT/ConSReg/tree/master/data/dap_seq_all_peaks). Note that the binding information for different TFs should be in separate files and each file is named by the name of the corresponding TF. For example, file "AT1G01060.narrowPeak" contains only the TF binding locations for the TF "AT1G01060."

3. *Differentially expressed gene (DEG) information from RNA-seq/microarray data*, represented as tables, which should contain four columns: column 1 represents gene name; a column named "baseMean" that represents mean expression for the gene; a column named "log2FoldChange" that represents log2-scaled fold change; and a column named "padj" that represents the adjusted p-values from statistical test for differential expressions. This type of DEG information usually could be generated from DESeq2 package (see DESeq2 page for more information: https://bioconductor.org/packages/release/bioc/html/DESeq2.html). ConSReg can take multiple tables as inputs, and each table corresponds to a DEG analysis from one experiment. In this tutorial, we will use processed sample data deposited in the ConSReg GitHub repository (https://github.com/LiLabAtVT/ConSReg/tree/master/data).

3 Methods

3.1 Preprocessing of the Input Data

In this section, we will start the analyses by downloading sample data and the repository. Download the ConSReg GitHub repository using the following commands (*see* **Note 3**):

```
 cd your_path # your_folder represents the path that hosts ConSReg package.
 git clone https://github.com/LiLabAtVT/ConSReg.git # download the repository and data
```

Change the current directory to the root folder of the downloaded repository.

```
cd your_path/ConSReg
```

Activate the ConSReg environment and Python:

```
conda activate ConSReg
python
```

This will start Python console, where the users can run code for the analysis. You may also use Jupyter notebook to run Python code in a graphical user interface (see Jupyter page for more information: https://jupyter.org/). For this tutorial, we will only focus on using the basic Python console. After Python console is successfully launched, you will see the version information of Python, as well as a line starting with ">>>" symbol where you can type in the code.

The first step of analysis starts from preprocessing the input data as specified in Subheading 2.3. Use the following code to import ConSReg package and other necessary packages:

```
import pandas as pd.
import os
import re
from ConSReg.Main import ConSReg
from ConSReg.Main import load_obj
```

Then specify some file names for the input data (*see* **Note 4**).

```
# dap-seq narrow peak files
dap_file_list = os.Listdir("data/dap_seq_all_peaks/")
dap_files = ["data/dap_seq_all_peaks/" + file for file in
dap_file_list if re.Match(".*narrowPeak",file) is not none]

# ATAC-seq peak file
atac_file = "data/atac_seq_all_peaks/all_merged.Bed"

# Arabidopsis genome annotation file
GFF_file = "data/GFF/TAIR10_GFF3_genes.gff"

# differential contrast result generated by DESeq2
diff_file_list = os.Listdir("data/diff_evalB/")
diff_files = ["data/diff_evalB/" + file for file in diff_fi-
le_list if re.Match(".*csv",file) is not none]
```

Next, we will first create a new analysis object using the function "ConSReg()" and then specify the parameters as follows for running the preprocessing step (*see* **Note 5**).

```
analysis = ConSReg()
params = {
'dap_files':dap_files,
'diff_files':diff_files,
'atac_file':atac_file,
'gff_file':gff_file,
'dap_chr_col':0,
'dap_chr_start_col':1,
```

```
'dap_chr_end_col':2,
'dap_strand_col':None,
'dap_signal_col':None,
'atac_chr_col':0,
'atac_chr_start_col':1,
'atac_chr_end_col':2,
'atac_signal_col':None,
'up_tss':3000,
'down_tss':500,
'up_type':'all',
'down_type':'all',
'use_peak_signal':False,
'n_jobs':2,
'Verbose':True
}
analysis.Preprocess(**params)
```

After preprocessing is finished, run the following code to generate feature matrices. Feature matrices will be used as inputs for the machine learning engines of ConSReg package (*see* **Note 6**).

```
analysis.gen_feature_mat(neg_type='udg',verbose = true)
```

3.2 Building and Evaluating Models

Before applying ConSReg to the data, it is important to examine the accuracy of ConSReg model for the input data. This will showcase whether ConSReg model can explain the environmental response well. Users may also evaluate the model performance by running cross-validation (CV) using the following code (*see* **Note 7**):

```
Analysis.eval_by_cv(ml_engine = 'lrlasso',rep = 5, n_jobs =
2)
```

Once CV evaluation is completed, you may view the CV result table by typing:

```
analysis.Auroc
```

You will see a table formatted as below:

diff_name	auroc_mean_UR	auroc_std_UR	auroc_mean_DR	auroc_std_DR
PRJEB10930_3_ T-PRJEB10930_3_C.csv	0.840888	0.020679	0.799490	0.017722
GSE81202_14-GSE81202_12. Csv	0.839107	0.008541	0.844182	0.011545

(continued)

diff_name	auroc_mean_UR	auroc_std_UR	auroc_mean_DR	auroc_std_DR
GSE81202_18-GSE81202_16.Csv	0.838261	0.010441	0.834370	0.007388
GSE63406_6-GSE63406_4.Csv	0.812779	0.013311	0.857815	0.005773

The column "diff_name" refers to the name of differential expression comparison, and the file name corresponds to the table file specified in "diff_file_list" (*see* Subheading **3.1**). The evaluation metric used here is area under the receiver operating characteristic curve (AUROC). The value of AUROC ranges between 0 and 1, and 1 means the best performance. ConSReg evaluates the model performance separately for comparison generated from upregulated DEG VS negative control genes (shown in column "auroc_mean_UR") and from downregulated DEG VS negative control genes (shown in column "auroc_mean_DR"). In the above table, ConSReg has achieved a good performance for these comparisons.

3.3 Predict Stress- or Condition-Specific Transcription Factors

The main analysis for ConSReg is to prioritize and rank TFs for the given environmental condition. This can be easily done with one line of code:

```
analysis.compute_imp_score(n_resampling = 200, n_jobs =
2, verbose = true)
```

"n_resampling" is the only parameter to be specified for this analysis, which represents number of replicates for subsampling and model fitting. The final importance score for each TF is computed as the number of times the TF gets selected divided by the total number of replicates. TFs were then ranked by these scores by descending order. Users may view the importance score results by the following code:

```
analysis.imp_scores_UR
analysis.imp_scores_DR
```

The first line shows the importance scores for upregulated DEG VS negative control genes, and the second line shows the importance scores for negative-regulated DEG VS negative control genes. An example of result table is shown below, in which the rows represent TFs and columns represent each differential expression comparison performed:

	PRJEB10930_3_T-PRJEB10930_3_C.csv	GSE81202_14-GSE81202_12.csv	GSE81202_18-GSE81202_16.csv
AT1G01060	0.080	0.000	0.000
AT1G01250	0.000	0.000	0.000
AT1G01720	1.000	0.000	0.000
AT1G02230	0.000	0.000	0.000

Since the result table is a Pandas data frame, we can sort the TFs for each condition of interest. For example, column "GSE81202_14-GSE81202_12.csv" represents the differential expression comparison for light stress treatment. We can sort the TFs by importance scores for this condition:

```
analysis.imp_scores_UR["GSE81202_14-GSE81202_12.Csv"].sort_-
values(ascending = false)
```

The resulted ranking shows the most important TFs (top 20 below):

```
AT1G27730 1.000
AT1G69570 1.000
AT5G05550 1.000
AT3G46590 1.000
AT5G62940 1.000
AT1G08010 1.000
AT2G18380 1.000
AT4G31800 1.000
AT2G40620 1.000
AT4G16750 1.000
AT2G45410 1.000
AT1G72740 1.000
AT5G19790 1.000
AT2G31230 1.000
AT3G22760 0.995
AT2G23290 0.995
AT1G06850 0.995
AT5G67300 0.995
AT5G16820 0.995
AT5G45580 0.990
AT1G28370 0.990
```

4 Notes

1. In Linux OS (e.g., Ubuntu), you can use the shortcut key Ctrl + Alt + T to open a terminal. Other distributions of Linux may use different keys.
2. A newly created conda environment is a folder that hosts all packages separated from the OS environment. Any changes made to this environment would not affect the OS. When you activate a conda environment, environment name will appear at the leftmost position in the terminal which indicates only packages installed for the current active environment can be imported by Python.
3. This repository also contains the sample data sets. The sample data sets will be downloaded along with the repository.
4. Inputs from DAP-seq/ChIP-seq data are separate files, and each corresponds to the genomic binding locations for a single TF. Input from ATAC-seq file is a merged single file which contains all open chromatin regions on different chromosomes. Inputs from differential expression comparisons (DEG tables) are separate files, and each corresponds to DEGs from a single environmental condition.
5. The details of these parameters are described below:

dap_file	A list. File names of DAP-seq peak files (bed format)
diff_file	A list. File names of differential contrasts, in the format of DESeq2 output file
atac_file	String. File name of atac peak files (bed format). Specify none if no atac-seq file is available
gff_file	String. File name of genome annotation gff file
dap_chr_col	Int, column number for dap-seq chromosome information, 0 indexed
dap_chr_start_col	Int, column number for dap-seq peak start position, 0 indexed
dap_chr_end_col	Int, column number for dap-seq peak end position, 0 indexed
dap_strand_col	Int or none, column number for dap-seq peak strand information, 0 indexed
dap_signal_col	Int or none, column number for dap-seq peak signal value, 0 indexed
atac_chr_col	Column number for atac-seq chromosome information, 0 indexed

(continued)

atac_chr_start_col	Column number for atac-seq peak start position, 0 indexed
atac_chr_end_col	Column number for atac-seq peak end position, 0 indexed
atac_signal_col	Column number for atac-seq peak signal value, 0 indexed
up_tss	Positions relative to upstream region of TSS. This is used for finding the nearest gene for each binding site
down_tss	Positions relative to downstream region of TSS. This is used for finding the nearest gene for each binding site
up_type	Type of binding sites. "All" or "intergenic"
down_type	Type of binding sites. "All" or "intron" or "non_intron"
use_peak_signal	True/false. Whether to use peak signal for ATAC-seq and DAP-seq?
use_atac_peak_signal	True/false

6. Here, the parameter "neg_type" indicates the negative control genes (the negative class) for machine learning classification task. This tutorial uses its default value "udg," which means "undetected genes." See original ConSReg publication [5] for more details.
7. The parameter "ml_engine" specifies the machine learning classifier to be used for ConSReg analysis. In our previous study [5], we have shown that logistic regression + LASSO can achieve very good performance at low computational cost. Therefore, in this tutorial we use "lrlasso" for our analysis. The parameter "n_jobs" specifies how many CPU cores to be used for running the analysis in parallel. While this can greatly speed up the computation, "n_jobs" should be at least one less than all available CPU cores to avoid system crash. For example, if you run the analysis on a four-core computer, you should specify a n_job less than or equal to 3. Parameter "rep" specifies the fold of cross-validation. Typically, a k fold cross-validation indicates that a data set is split into k partitions and a machine learning classifier will iteratively take k-1 partitions for model training and use the remaining one for testing the model. The final performance metric will be averaged over k rounds of tests.

Acknowledgments

We thank Jeffress Trust Awards Program in Interdisciplinary Research, US Department of Energy Funding [DE-SC0020358], and Hatch Programs from US Department of Agriculture for providing funding support for the development of ConSReg package.

References

1. Huang X, Shi H, Hu Z, Liu A, Amombo E, Chen L et al (2017) ABA is involved in regulation of cold stress response in Bermuda grass. Front Plant Sci 8:1613
2. Li N, Euring D, Cha JY, Lin Z, Lu M, Huang LJ et al (2021) Plant hormone-mediated regulation of heat tolerance in response to global climate change. Front Plant Sci 11:627969
3. Zhou Y, He R, Guo Y, Liu K, Huang G, Peng C et al (2019) A novel ABA functional analogue B2 enhances drought tolerance in wheat. Sci Rep 9:2887
4. Waśkiewicz A, Beszterda M, Goliński P. (2013) ABA: Role in plant signaling under salt stress. Salt Stress Plants Signalling Omi Adapt 1:175–196
5. Song Q, Lee J, Akter S, Rogers M, Grene R, Li S (2021) Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Res 48:e62

Chapter 16

Identification of Plant Co-regulatory Modules Using CoReg

Qi Song and Song Li

Abstract

Regulatory network is often characterized by complex interactions between transcription factors (TFs) and target genes. The synergistic regulations among multiple TFs may co-induce/co-suppress the expressions of the similar target genes. Such information is important for understanding stress response signaling pathways in plants. In this chapter, we present a computational tool, CoReg, for mining co-regulatory gene modules from network topology. The analysis results can be used to interpret co-regulation effects in regulatory networks generated by high-through TF–DNA interaction screenings such as yeast-one-hybrid, ChIP-seq, and DAP-seq.

Key words Regulatory network, Co-regulation

1 Introduction

Transcription of a gene is usually initiated by transcription factors (TFs) that bind to the promoter region of the gene. This process can regulate expressions of the target genes which may further activate cascades of signaling pathways. It is important to investigate interactions between TFs and target genes to identify potential regulatory programs from the network topology. Currently, there are multiple high-throughput assays that can screen for interactions between TFs and target genes, including enhanced yeast-one-hybrid (eY1H) [1], protein binding microarray [2], chromatin immunoprecipitation sequencing (ChIP-seq) [3], and DNA affinity purification sequencing (DAP-seq) [4]. With the accumulation of many large data sets from these platforms, a computational tool is needed to characterize the roles of TFs from the network data. It has been reported that TFs can form protein complex and bind to promoter region to co-regulate the downstream genes [5]. This may underlie the biological meaning of observing a gene co-targeted by multiple TFs in a regulatory network. Furthermore, a TF can be functionally redundant to other TFs and may be activated when its counterpart fails to function [6]. To identify

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7_16,

such co-regulation events, we have previously developed a computational tool CoReg (**co**-**reg**ulation), which takes a regulatory network as input and identifies co-regulatory gene modules from the network [7]. A co-regulatory module is defined as a set of genes which share similar target genes and regulators. In this chapter, we focus on presenting a step-by-step protocol of identifying co-regulatory gene modules from a regulatory network using CoReg package. A short tutorial is also available at the GitHub repository: https://lilabatvt.github.io/CoReg/.

2 Materials

2.1 Computer and Operating System Environment

HPC or PC, Linux preferred**,** Python (version ≥3.6) and R. In this protocol, we will only demonstrate the use of CoReg on Linux (Ubuntu 20.04) operating system (OS). Other distributions of Linux systems, Windows, and macOS would work in a similar manner.

2.2 Installation

Installation of CoReg package is very straight-forward and easy. CoReg package is only dependent on R language platform and some other R packages.

1. Retrieve R from https://www.r-project.org/ if there isn't any version of R installed in your OS.
2. Once R is installed, in R console, type in the following code to install CoReg package. This will install CoReg directly from the GitHub repository and will automatically synchronize the current installation with the remote repository.

```
install.Packages("devtools")
library(devtools)
install_github("CoReg")
```

3. Start CoReg by type in the following code:

```
library(CoReg)
```

2.3 Prepare Input Data for ConSReg

In CoReg package, we provide an *Arabidopsis* sample network data set generated from a published research [8]. Users may load the sample data by:

```
data(athNet)
```

The loaded network is an igraph object. To better understand the inputs, we may convert it to an edge list and then view this edge list in R. To do this, load igraph package, and use "as_edgelist()" function to convert "athNet" to two-column edge list.

```
library(igraph)
athNet <- as_edgelist(athNet)
```

If we view the converted edge list by "head()" function, we can see a two-column edge list with the first column representing IDs of TFs and second column resenting IDs of target genes.

```
head(athNet,5)
"AT5G61590" "AT1G09610"
"AT5G61590" "AT1G71930"
"AT5G61590" "AT4G39070"
"AT5G61590" "AT5G15630"
"AT5G61590" "AT1G04240"
```

Alternatively, users may load their own network using the function "networkFromFile()." Below is an example of loading a CSV file named "araNet.csv." The input file needs to be formatted as a two-column CSV file as described above. A sample CSV file can be downloaded from the GitHub repository: https://raw.githubusercontent.com/LiLabAtVT/CoReg/master/data/athNet.csv

```
athNet<-networkFromFile("araNet.Csv",",")
```

3 Methods

3.1 Identification of Co-regulatory Modules

The main and most important functionality for CoReg package is the identification of co-regulatory modules from gene regulatory network. A co-regulatory module is defined as a set of genes sharing similar targets and similar TFs that co-regulate them (*see* Fig. 1). CoReg identifies such module by computing the number of shared TFs and the number of shared targets for each pair of genes. This would yield a similarity matrix among all genes, which will be used as inputs for hierarchical clustering followed by dynamic tree cut. See the original CoReg publication for more details [7].

Here, this chapter aims at introducing a step-by-step protocol for CoReg package. In the previous section, we have walked through how to load the sample data set and how to format the data to be used as inputs. Once we have loaded the data (in the previous section, this is loaded into an R object named "athNet"), identification of co-regulatory module can be conveniently done with just one step:

```
athRes<-CoReg(athNet)
```

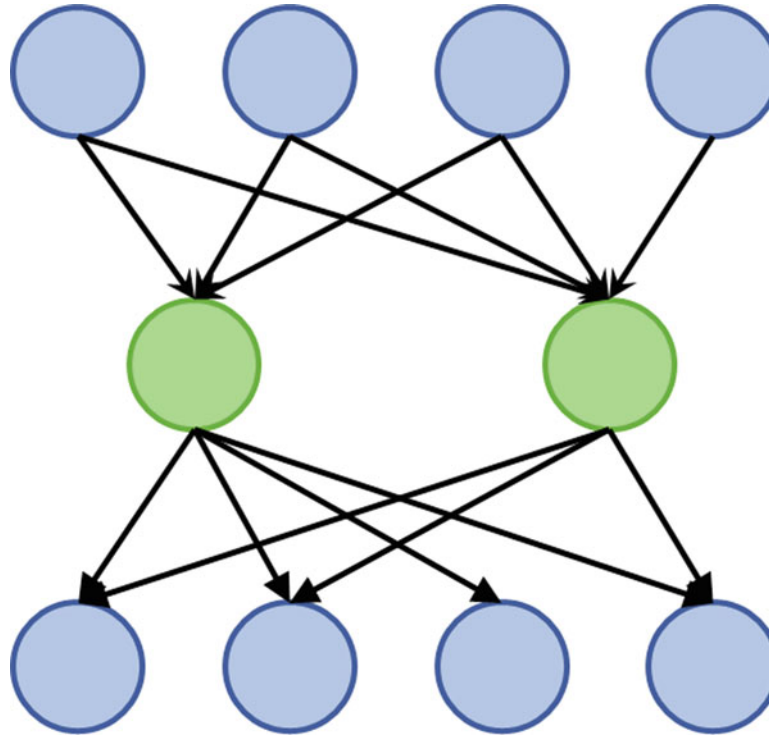


Fig. 1 Schematic illustration of a co-regulatory module. The two green circles in the middle represent TFs co-targeted by many common regulators (blue circles on the top) and co-target many common target genes (blue circles at the bottom). This type of gene module is defined as a co-regulatory module. Note that a co-regulatory module may contain more than two genes

Users can then access module finding results through the "module" attribute of the returned object:

```
head(athRes$module,5)
ID module
1 AT5G61590 8
2 AT1G09610 5
3 AT5G44210 1
4 AT3G27010 13
5 AT1G71130 6
```

"module" attribute is a two-column table with gene ID in the first column and module ID in the second column. Users may save this table using R built-in function "write.csv()."

3.2 Evaluation of the Modules Using Rewiring Simulations

In CoReg package, network rewiring simulation is provided as a method to evaluate the robustness of the identified modules. Simulation is performed by first duplicating some TFs and their connections to the target genes and then randomly rewiring these

connections. CoReg computes a rewiring recall score to assess how well the algorithm can assign the duplicated and the original TFs to the same module. To perform this simulation, we may use "rewSim ()" function in CoReg package. Below is the code for performing a simulation and comparing different module finding algorithms on this simulated network:

```
simRes <- rewSim(athNet, nDup = 50, dDup = 10, rewProb = c
(0.1,0.3,0.5), methods = c("coregJac","lp","wt","eb"),nRep =
5)
```

See **Note 1** for more details about the arguments.

Users may view the simulation results by accessing "evalResult" attribute (*see* **Note 2**):

```
> simRes$evalResult
coregJac-score-mean coregJac-score-sd lp-score-mean
0.6783810 0.05910124 0.006738950
0.5260488 0.02481673 0.004440511
0.4453706 0.06063716 0.004444444
Lp-score-sd wt-score-mean wt-score-sd eb-score-mean
3.152014e-03 0.07181961 0.017050826 0.04908824
8.794761e-06 0.05231855 0.010154373 0.04368346
0.000000e+00 0.04861389 0.004677145 0.03799002
Eb-score-sd
0.002338189
0.007732074
0.005569694
```

Each row in this table summarizes the mean and standard deviation of rewiring recall scores from five runs of the simulation (as specified by "nRep" argument). For the above output results, we can see that CoReg has achieved the highest score among all methods.

Additionally, CoReg can compute a simulation-based AUC value (*see* **Note 3**) to evaluate CoReg module finding results. The simulation process is similar to the previously described procedures. We can compute the AUC values using "computeAuROC()" function:

```
auROCres <- computeAuROC(athNet,nDup = 50,dDup = 10,rewProb =
0.3,
simMethods = c("jaccard","geometric","nvlogweighted","wt"))
```

See **Note 4** for more details about the arguments.

Finally, we can view AUC values by accessing the "AUC" attribute of the returned object:

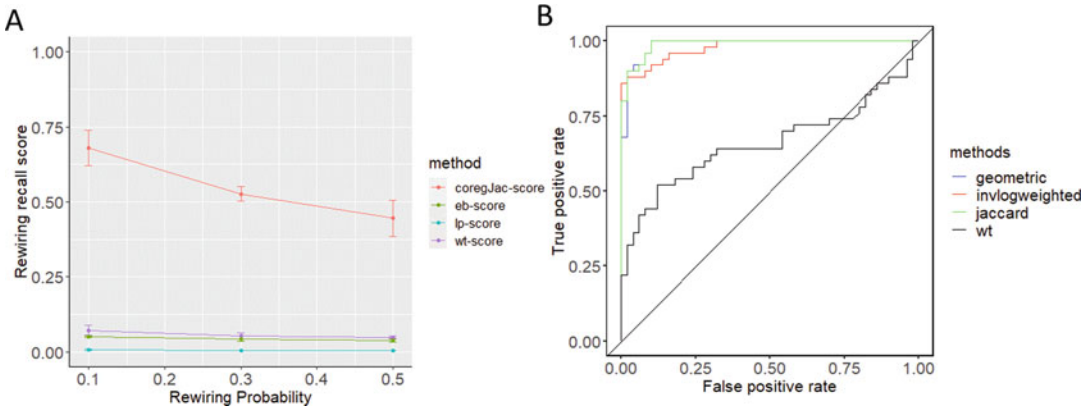


Fig. 2 The rewiring recall score plot and ROC plot. (**a**) The rewiring recall score plot for all methods tested. (**b**) The ROC curves for all methods tested. The closer the curve is to the upper left corner, the better performance (and higher AUC value) it indicates

```
auROCres$AUC
Jaccard geometric invlogweighted wt
0.9896 0.9876 0.9780 0.6594
```

This will give a quick evaluation on different similarity index used by CoReg (*see* **Note 4**). Users may choose the similarity index that achieves the best performance. For the two simulations performed above, we can plot the results using "plot()" function in R (*see* Fig. 2).

```
plot(simRes)
plot(auROCres)
```

4 Notes

1. "nDup" represents number of the TFs to be duplicated. The connections of these TFs to other genes will be rewired. "dDup" is the minimum degree for the TFs to be duplicated. "rewProb" is a set of probability values specifying the probability of rewiring each edge coming from the duplicated TFs. "methods" specify what methods to be compared with. Currently, CoReg supports using different three types of similarity indices to perform clustering: "coregJac" is CoReg + Jaccard index; "coregGeo" is CoReg + geometric index; "coregInv" is CoReg + inverse log-weighted index). CoReg package also includes other clustering methods: "lp" is label propagation; "wt" is walk trap, "eb" is edge betweenness. "nRep" specifies how many replicates of simulation to be performed. CoReg will report the mean and standard deviations of scores from all replicates.

2. The abbreviated names in the header of the output are explained in **Note 1**.
3. AUC stands for area under the receiver operating characteristic curve (ROC curve). AUC value ranges between 0.5 and 1, with 0.5 indicating random predictions and 1 the perfect performance.
4. All arguments in "computeAuROC()" have the same meanings as specified in ***Note 1***; "computeAuROC()" has one additional argument named "simMethods" that specifies the clustering methods to be compared: "jaccard" stands for CoReg + jaccard similarity index; "geometric" stands for CoReg + geometric similarity index; "invlogweighted" stands for CoReg + inverse log weighted similarity index; "wt" stands for CoReg + walktrap-based similarity index.

Acknowledgments

We thank Virginia Agricultural Experiment Station (Blacksburg), USADA National Institute of Food and Agriculture, US Department of Agriculture (Washington, DC) for supporting the development of CoReg.

References

1. Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S et al (2011) Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nat Methods 8:1059–1064
2. Bulyk ML (2006) Protein binding microarrays for the characterization of DNA-protein interactions. Adv Biochem Eng Biotechnol 104:65–68
3. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502
4. O'Malley RC, Huang SSC, Song L, Lewsey MG, Bartlett A, Nery JR et al (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. Cell 165:1280–1292
5. Liu W, Stewart CN (2016) Plant synthetic promoters and transcription factors. Curr Opin Biotechnol 27:36–44
6. Wu WS, Lai FJ (2015) Functional redundancy of transcription factors explains why most binding targets of a transcription factor are not affected when the transcription factor is knocked out. BMC Syst Biol 9(Suppl 6): S2
7. Song Q, Grene R, Heath LS, Li S (2017) Identification of regulatory modules in genome scale transcription regulatory networks. BMC Syst Biol 11:140
8. Taylor-Teeples M, Lin L, De Lucas M, Turco G, Toal TW, Gaudinier A et al (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature 517: 571–575

INDEX

Qi Song and Zhipeng Tao (eds.), *Transcription Factor Regulatory Networks*,
Methods in Molecular Biology, vol. 2594, https://doi.org/10.1007/978-1-0716-2815-7,

N

P

R

S

T

U

W

Y