

Methods in
Molecular Biology 1580

Springer Protocols

Tamas Dalmay *Editor*



MicroRNA Detection and Target Identification

Methods and Protocols

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:

<http://www.springer.com/series/7651>

MicroRNA Detection and Target Identification

Methods and Protocols

Edited by

Tamas Dalmay

School of Biological Sciences, University of East Anglia, Norwich, UK

 **Humana Press**

Editor

Tamas Dalmay
School of Biological Sciences
University of East Anglia
Norwich, UK

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-6864-0 ISBN 978-1-4939-6866-4 (eBook)
DOI 10.1007/978-1-4939-6866-4

Library of Congress Control Number: 2017937361

© Springer Science+Business Media LLC 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

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

Preface

This book is a follow-up of a previous book in this series; therefore, it is unnecessary to introduce microRNAs (miRNAs) to any reader who is reading this preface. The previous book (MicroRNAs in development; published in 2011) described protocols to detect, profile, and manipulate miRNAs in various organisms, as well as how to validate targets of miRNAs in plants and animals. However, a lot of new techniques have been developed in the last 5–6 years, which warranted a new book.

Some of the new protocols describe slight but important changes to well-established techniques that were described in the previous edition, such as Northern blot (Chapter 1) and preparation of cDNA libraries of small RNAs (Chapter 4). An alternative method to these two approaches to detect miRNAs is RT-qPCR, and there are two protocols for this in the book, one describing high-throughput RT-qPCR (Chapter 2) and the other describing the application of digital PCR for miRNA detection (Chapter 16). In addition, there is a review chapter on the comparison of next-generation sequencing and RT-qPCR platforms (Chapter 3).

MiRNAs have been increasingly used as biomarkers in cell-free body liquids such as serum or urine. The amount of miRNAs in these samples is much lower than in samples containing cells; therefore, there is a need for more sensitive methods. There are a number of protocols for miRNA detection in this book that are based on completely novel approaches. These exciting techniques utilize nanotechnology, microfluidics, or other engineering innovations to lower the detection limit (Chapters 5, 6, 8, 16, 17, 18, and 20).

A very important aspect of miRNA research is to identify and validate their target mRNAs. Identifying targets in plants is relatively straightforward due to the high complementarity between miRNAs and their targets. This near perfect match results in a cleavage at a specific position on the mRNA, and these cleavage fragments can be sequenced and therefore identified. Since that protocol was published in the previous edition, there is no chapter on plant miRNA target identification in this book. However, there are two new experimental approaches for miRNA target identification in animals included in this edition (Chapters 7 and 9).

In addition to wet laboratory protocols, miRNA research hugely relies on bioinformatics approaches, probably more so than most other field of biology. This aspect was completely missing from the previous edition and we now make up for it. There are seven chapters describing either specific programs or entire tool kits or reviewing certain aspects of miRNA bioinformatics. These are chapters 10–15 and 19.

Norwich, UK

Tamas Dalmay

Contents

<i>Preface</i>	V
<i>Contributors</i>	IX
1 Improved Denaturation of Small RNA Duplexes and Its Application for Northern Blotting <i>C. Jake Harris, David C. Baulcombe, and Attila Molnar</i>	1
2 High-Throughput RT-qPCR for the Analysis of Circulating MicroRNAs. <i>Geok Wee Tan and Lu Ping Tan</i>	7
3 Genome-Wide Comparison of Next-Generation Sequencing and qPCR Platforms for microRNA Profiling in Serum <i>Thorarinn Blondal, Maurizia Rossana Brunetto, Daniela Cavallone, Martin Mikkelsen, Michael Thorsen, Yuan Mang, Hazel Pinheiro, Ferruccio Bonino, and Peter Mouritzen</i>	21
4 Small RNA Profiling by Next-Generation Sequencing Using High-Definition Adapters <i>Martina Billmeier and Ping Xu</i>	45
5 Surface Acoustic Wave Lysis and Ion-Exchange Membrane Quantification of Exosomal MicroRNA <i>Katherine E. Richards, David B. Go, and Reginald Hill</i>	59
6 Droplet Microfluidic Device Fabrication and Use for Isothermal Amplification and Detection of MicroRNA <i>Maria Chiara Giuffrida, Roberta D'Agata, and Giuseppe Spoto</i>	71
7 Interrogation of Functional miRNA–Target Interactions by CRISPR/Cas9 Genome Engineering <i>Yale S. Michaels, Qianxin Wu, and Tudor A. Fulga</i>	79
8 Cell-Free Urinary MicroRNAs Expression in Small-Scale Experiments. <i>Ludek Zavesky, Eva Jandakova, Radovan Turyna, Daniela Duskova, Lucie Langmeierova, Vit Weinberger, Lubos Minar, Ales Horinek, and Milada Kohoutova</i>	99
9 Peptide-Based Isolation of Argonaute Protein Complexes Using Ago-APP <i>Judith Hauptmann and Gunter Meister</i>	107
10 Predicting Functional MicroRNA-mRNA Interactions <i>Zixing Wang and Yin Liu</i>	117
11 Computational and Experimental Identification of Tissue-Specific MicroRNA Targets <i>Rabeleh Amirkhah, Hojjat Naderi Meshkin, Ali Farazmand, John E.J. Rasko, and Ulf Schmitz</i>	127

12	sRNAtoolboxVM: Small RNA Analysis in a Virtual Machine	149
	<i>Cristina Gómez-Martín, Ricardo Lebrón, Antonio Rueda, José L. Oliver, and Michael Hackenberg</i>	
13	An Assessment of the Next Generation of Animal miRNA Target Prediction Algorithms	175
	<i>Thomas Bradley and Simon Moxon</i>	
14	The UEA Small RNA Workbench: A Suite of Computational Tools for Small RNA Analysis	193
	<i>Irina Mohorianu, Matthew Benedict Stocks, Christopher Steven Applegate, Leighton Folkes, and Vincent Moulton</i>	
15	Prediction of miRNA–mRNA Interactions Using miRGate	225
	<i>Eduardo Andrés-León, Gonzalo Gómez-López, and David G. Pisano</i>	
16	Detection of microRNAs Using Chip-Based QuantStudio 3D Digital PCR . . .	239
	<i>Cristina Borzi, Linda Calzolari, Davide Conte, Gabriella Sozzi, and Orazio Fortunato</i>	
17	MiRNA Quantitation with Microelectrode Sensors Enabled by Enzymeless Electrochemical Signal Amplification	249
	<i>Tanyu Wang, Gangli Wang, Didier Merlin, and Emilie Viennois</i>	
18	A Robust Protocol to Quantify Circulating Cancer Biomarker MicroRNAs	265
	<i>Emma Bell, Hannah L. Watson, Shivani Bailey, Matthew J. Murray, and Nicholas Coleman</i>	
19	MicroRNAs, Regulatory Networks, and Comorbidities: Decoding Complex Systems	281
	<i>Francesco Russo, Kirstine Belling, Anders Boeck Jensen, Flavia Scoyni, Søren Brunak, and Marco Pellegrini</i>	
20	Label-Free Direct Detection of MiRNAs with Poly-Silicon Nanowire Biosensors	297
	<i>Jing He, Jianjun Zhu, Bin Jiang, and Yulan Zhao</i>	
	Erratum to:	E1
	<i>Index</i>	303

Contributors

- RAHELEH AMIRKHAH • *Reza Institute of Cancer Bioinformatics and Personalized Medicine, Mashhad, Iran*
- EDUARDO ANDRÉS-LEÓN • *Bioinformatics Unit, Instituto de Parasitología y Biomedicina “López Neyra”, Consejo Superior de Investigaciones Científicas (IPBLN-CSIC), PTS Granada, Granada, Spain*
- CHRISTOPHER STEVEN APPLEGATE • *School of Computing Sciences, University of East Anglia, Norwich, UK*
- SHIVANI BAILEY • *Department of Pathology, University of Cambridge, Cambridge, UK*
- DAVID C. BAULCOMBE • *Department of Plant Sciences, University of Cambridge, Cambridge, UK*
- EMMA BELL • *Department of Pathology, University of Cambridge, Cambridge, UK; AstraZeneca, Cambridge Science Park, Cambridge, UK*
- KIRSTINE BELLING • *Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
- MARTINA BILLMEIER • *School of Biological Sciences, University of East Anglia, Norwich, UK*
- THORARINN BLONDAL • *Exiqon A/S, Vedbaek, Denmark*
- FERRUCCIO BONINO • *Hepatology Unit and Laboratory of Molecular Genetics and Pathology of Hepatitis Viruses, Reference Center of the Tuscany Region for Chronic Liver Disease and Cancer, University Hospital of Pisa, Pisa, Italy*
- CRISTINA BORZI • *Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
- THOMAS BRADLEY • *School of Biological Sciences, University of East Anglia, Norwich, UK; Earlham Institute, Norwich Research Park, Norwich, UK*
- SØREN BRUNAK • *Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
- MAURIZIA ROSSANA BRUNETTO • *Hepatology Unit and Laboratory of Molecular Genetics and Pathology of Hepatitis Viruses, Reference Center of the Tuscany Region for Chronic Liver Disease and Cancer, University Hospital of Pisa, Pisa, Italy*
- LINDA CALZOLARI • *Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
- DANIELA CAVALLONE • *Hepatology Unit and Laboratory of Molecular Genetics and Pathology of Hepatitis Viruses, Reference Center of the Tuscany Region for Chronic Liver Disease and Cancer, University Hospital of Pisa, Pisa, Italy*
- NICHOLAS COLEMAN • *Department of Pathology, University of Cambridge, Cambridge, UK; Department of Histopathology, Addenbrooke’s Hospital, Cambridge, UK*
- DAVIDE CONTE • *Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
- ROBERTA D’AGATA • *I.N.B.B. Consortium, Rome, Italy*
- TAMAS DALMAY • *School of Biological Sciences, University of East Anglia, Norwich, UK*

- DANIELA DUSKOVA • *Faculty Transfusion Centre, General University Hospital in Prague, Prague, Czech Republic*
- ALI FARAZMAND • *Department of Cell and Molecular Biology, School of Biology, College of Science, University of Tehran, Tehran, Iran*
- LEIGHTON FOLKES • *The Earlham Institute, Norwich, UK*
- ORAZIO FORTUNATO • *Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
- TUDOR A. FULGA • *Radcliffe Department of Medicine, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK*
- MARIA CHIARA GIUFFRIDA • *I.N.B.B. Consortium, Rome, Italy*
- DAVID B. GO • *Department of Aerospace and Mechanical Engineering, University of Notre Dame, South Bend, IN, USA; Department of Chemical and Biomolecular Engineering, University of Notre Dame, South Bend, IN, USA*
- CRISTINA GÓMEZ-MARTÍN • *Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain*
- GONZALO GÓMEZ-LÓPEZ • *Bioinformatics Unit (UBio), Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
- MICHAEL HACKENBERG • *Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain; Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, Instituto de Biotecnología, Granada, Spain*
- C. JAKE HARRIS • *Department of Plant Sciences, University of Cambridge, Cambridge, UK*
- JUDITH HAUPTMANN • *University of Regensburg, Regensburg, Germany*
- JING HE • *Shanghai Integrated Circuit Research & Development Center, Shanghai, China; School of Life Science, East China Normal University, Shanghai, People's Republic of China*
- REGINALD HILL • *Department of Biological Sciences, Harper Cancer Research Institute, University of Notre Dame, South Bend, IN, USA*
- ALES HORINEK • *First Faculty of Medicine, Institute of Biology and Medical Genetics, Charles University Prague and General University Hospital in Prague, Prague, Czech Republic*
- EVA JANDAKOVA • *Institute of Pathology, University Hospital Brno, Brno, Czech Republic*
- ANDERS BOECK JENSEN • *Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
- BIN JIANG • *Shanghai Integrated Circuit Research & Development Center, Shanghai, China*
- MILADA KOHOUTOVA • *First Faculty of Medicine, Institute of Biology and Medical Genetics, Charles University Prague and General University Hospital in Prague, Prague, Czech Republic*
- LUCIE LANGMEIEROVA • *Faculty Transfusion Centre, General University Hospital in Prague, Prague, Czech Republic*
- RICARDO LEBRÓN • *Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain; Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, Instituto de Biotecnología, Granada, Spain*
- YIN LIU • *Department of Neurobiology and Anatomy, University of Texas Health Science Center at Houston, Houston, TX, USA; University of Texas Graduate School of Biomedical Science, Houston, TX, USA*
- YUAN MANG • *Exiqon A/S, Vedbaek, Denmark*

- GUNTER MEISTER • *University of Regensburg, Regensburg, Germany*
- DIDIER MERLIN • *Department of Chemistry, Georgia State University, Atlanta, GA, USA*
- HOJJAT NADERI MESHKIN • *Stem Cells and Regenerative Medicine Research Group, Academic Center for Education, Culture Research (ACECR), Mashhad, Iran*
- YALE S. MICHAELS • *Radcliffe Department of Medicine, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK*
- MARTIN MIKKELSEN • *Exiqon A/S, Vedbaek, Denmark*
- LUBOS MINAR • *Department of Obstetrics and Gynaecology, University Hospital Brno, Brno, Czech Republic*
- IRINA MOHORIANU • *School of Biological Sciences, University of East Anglia, Norwich, UK; School of Computing Sciences, University of East Anglia, Norwich, UK*
- ATTILA MOLNAR • *School of Biological Sciences, University of Edinburgh, Edinburgh, UK*
- VINCENT MOULTON • *School of Computing Sciences, University of East Anglia, Norwich, UK*
- PETER MOURITZEN • *Exiqon A/S, Vedbaek, Denmark*
- SIMON MOXON • *School of Biological Sciences, University of East Anglia, Norwich, UK; Earlham Institute, Norwich Research Park, Norwich, UK*
- MATTHEW J. MURRAY • *Department of Pathology, University of Cambridge, Cambridge, UK; Department of Paediatrics, Haematology and Oncology, Addenbrooke's Hospital, Cambridge, UK; Department of Paediatrics, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK*
- JOSÉ L. OLIVER • *Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain; Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, Instituto de Biotecnología, Granada, Spain*
- MARCO PELLEGRINI • *Institute of Informatics and Telematics, National Research Council (CNR), Pisa, Italy*
- HAZEL PINHEIRO • *Exiqon A/S, Vedbaek, Denmark*
- DAVID G. PISANO • *Bioinformatics Unit (UBio), Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
- JOHN E.J. RASKO • *Gene & Stem Cell Therapy Program, Centenary Institute, Camperdown, Sydney Medical School, University of Sydney, Camperdown, NSW, Australia*
- KATHERINE E. RICHARDS • *Department of Biological Sciences, Harper Cancer Research Institute, University of Notre Dame, South Bend, IN, USA*
- ANTONIO RUEDA • *Queen Mary University of London, London, UK*
- FRANCESCO RUSSO • *Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
- ULF SCHMITZ • *Gene & Stem Cell Therapy Program, Centenary Institute, Camperdown, Sydney Medical School, University of Sydney, Camperdown, NSW, Australia*
- FLAVIA SCOYNI • *Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark*
- GABRIELLA SOZZI • *Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
- GIUSEPPE SPOTO • *I.N.B.B. Consortium, Rome, Italy; Dipartimento di Scienze Chimiche, Università di Catania, Catania, Italy*
- MATTHEW BENEDICT STOCKS • *School of Computing Sciences, University of East Anglia, Norwich, UK*

- GEOK WEE TAN • *Molecular Pathology Unit, Cancer Research Centre, Institute for Medical Research, Jalan Pahang, Kuala Lumpur, Malaysia*
- LU PING TAN • *Molecular Pathology Unit, Cancer Research Centre, Institute for Medical Research, Jalan Pahang, Kuala Lumpur, Malaysia*
- MICHAEL THORSEN • *Exiqon A/S, Vedbaek, Denmark*
- RADOVAN TURYN • *Institute for the Care of Mother and Child, Prague, Czech Republic*
- EMILIE VIENNOIS • *Department of Chemistry, Georgia State University, Atlanta, GA, USA*
- GANGLI WANG • *Department of Chemistry, Georgia State University, Atlanta, GA, USA*
- TANYU WANG • *Department of Chemistry, Georgia State University, Atlanta, GA, USA*
- ZIXING WANG • *University of Texas M.D. Anderson Cancer Center, Houston, TX, USA*
- HANNAH L. WATSON • *Department of Pathology, University of Cambridge, Cambridge, UK*
- VIT WEINBERGER • *Department of Obstetrics and Gynaecology, University Hospital Brno, Brno, Czech Republic*
- QIANXIN WU • *Radcliffe Department of Medicine, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK*
- PING XU • *School of Biological Sciences, University of East Anglia, Norwich, UK*
- LUDEK ZAVESKY • *First Faculty of Medicine, Institute of Biology and Medical Genetics, Charles University Prague and General University Hospital in Prague, Prague, Czech Republic*
- YULAN ZHAO • *School of Life Science, East China Normal University, Shanghai, China*
- JIANJUN ZHU • *Shanghai Integrated Circuit Research & Development Center, Shanghai, China*

Chapter 1

Improved Denaturation of Small RNA Duplexes and Its Application for Northern Blotting

C. Jake Harris, David C. Baulcombe, and Attila Molnar

Abstract

Small RNAs (sRNAs) are short (18–30 nucleotide) noncoding RNA molecules, which control gene expression and pathogen response in eukaryotes. They are associated with and guide nucleases to target nucleic acids by nucleotide base pairing. We found that current techniques for small RNA detection are adversely affected by the presence of complementary RNA. Thus we established FDF-PAGE (fully denaturing formaldehyde polyacrylamide gel electrophoresis), which dramatically improves denaturation efficiency and subsequently the detection of sequestered sRNAs.

Key words Small RNA, Sequestration, Polyacrylamide gel electrophoresis, Denaturation, Improved detection

1 Introduction

Since their discovery as the mediators of gene silencing [1, 2], small RNAs have been implicated in an ever expanding repertoire of cellular processes, from gene regulation, to the maintenance of genome integrity. The main method of small RNA detection is by small RNA Northern blotting, whereby total RNA is separated on a high percentage polyacrylamide gel, transferred to a membrane and visualized using radioactively labeled complementary probes [1, 3]. This technique is similar to long RNA Northern blotting, used for detection of mRNAs and higher molecular weight transcripts [4]. There are two main differences between long and small RNA Northern blotting techniques. First, small RNA Northern employ 15% polyacrylamide gels for separation of total RNA, while long RNA Northern use agarose gels. This is because agarose gels are insufficient to resolve RNA molecules of 18–30 nt in length. The second difference is that, due to the propensity of long RNA molecules to form secondary structures, long RNA Northern proceed in highly denaturing formaldehyde gels, while small RNA Northern typically employ only 7M urea for denaturation through electrophoresis.

However, recent reports suggested that high molecular weight RNA—present in endogenous total RNA samples—might sequester complementary small RNAs from detection, thus implying that methods for small RNA Northern blotting are not fully denaturing [5]. We recently validated this result, finding that both small RNA and high molecular weight RNAs are able to sequester complementary small RNAs of interest, masking them from detection by small RNA Northern blotting [6]. We therefore developed a method that employs formaldehyde to efficiently denature total RNA prior to entry into a 7 M urea polyacrylamide gel, the so-called FDF-PAGE (fully denaturing formaldehyde polyacrylamide gel electrophoresis) [6]. We found that this technique releases small RNAs from sequestration even in the presence of one thousand fold molar excess complementary RNA. FDF-PAGE combines the denaturation efficiency of long RNA Northern blotting with the relative ease of small RNA Northern blots, because we have found that incubation of total RNA in formaldehyde is sufficient for full small RNA denaturation, while long RNA Northern blots require the relatively toxic formaldehyde to be present within the agarose gel itself through electrophoresis. Critically, this improved method for small RNA detection also retains the ability to resolve small RNAs of 18–30 nt in length.

We describe here FDF-PAGE as an improved method for small RNA Northern blotting, which should provide a more accurate representation of small RNA abundance in the sample. To obtain an improved global picture of small RNA levels, FDF-PAGE can also be adapted to generate small RNA libraries. Here, small RNAs are cut out from the polyacrylamide gel after electrophoresis, purified and then used as input for a small RNA library preparation protocols (as described in [6]).

2 Materials

Prepare all solutions using ultrapure autoclaved water (prepared by purifying deionized water to attain a sensitivity of 18 M Ω cm at 25 °C) and analytical grade reagents. Store all reagents at room temperature (unless indicated otherwise). Strictly follow local waste disposal regulations when disposing of waste materials.

2.1 Denaturing Polyacrylamide Gel Components

1. MOPS buffer (10 $\times$): 200 mM MOPS, 50 mM NaOAc, 10 mM EDTA, pH to 7.0. Store at room temperature in a bottle wrapped with aluminum foil.
2. Acrylamide–bis-acrylamide solution: 40% (w/v) 19:1. Store at 4 °C.
3. Ammonium persulfate (SIGMA, A3678): 10% solution in water.
4. *N,N,N',N'*-tetramethyl-ethylenediamine (TEMED). Store at 4 °C.

5. Formaldehyde (37–40%).
6. Formamide (deionized).
7. Running buffer: 0.5× MOPS buffer diluted in ultrapure autoclaved water.
8. 10× loading dye: dissolve 5 mg xylene cyanol FF and 5 mg bromophenol blue in 10 mL water. Store at 4 °C.

2.2 Capillary Blotting Components

1. Transfer buffer (TB): 20× SSC.
2. Membrane: Hybond-N, NX, N+, or ZetaProbe GT.

2.3 Labeling and Hybridization Components

1. 10× Polynucleotide kinase buffer.
2. T4 Polynucleotide kinase, 10 (unit/μL).
3. Hybridization Solution: 0.25 M sodium phosphate buffer pH 7.2 [7] and 7% SDS.
4. Washing buffer (WB): 2× SSC, 0.1% SDS.
5. Microspin G-25 column.
6. 0.5 M EDTA, pH 7.0.
7. 10 μM of ssDNA oligo complementary to small RNA of interest.
8. γ³²P-ATP.

3 Methods

Carry out all procedures at room temperature unless otherwise specified.

3.1 15% Denaturing Polyacrylamide Gel Electrophoresis with 7 M Urea

1. Mix 4.2 g urea, 0.5 mL of 10× MOPS buffer, 3.75 mL of acrylamide:bis-acrylamide solution, and 2.5 mL of ultrapure water with a magnetic stirrer in a 50 mL conical flask until the urea is completely dissolved (*see* **Notes 1** and **2**). Add 70 μL of ammonium persulphate (*see* **Note 3**) and 3.5 μL of TEMED, mix thoroughly and cast gel with a 0.75 mm spacer. Insert a 10-well gel comb immediately without introducing air bubbles and let it polymerize for 30 min (*see* **Note 4**).
2. Assemble electrophoresis equipment and fill with running buffer according to the manufacturer's instruction. Remove gel comb, rinse wells (*see* **Note 5**) with running buffer, and pre-run the gel at 100 Volts for 30 min.
3. In the meantime, prepare RNA samples on ice in the fume hood. For each gel lane add V μL of RNA (*see* **Note 6**), 2.75 μL of formaldehyde, 7.5 μL of formamide, 0.75 μL of MOPS buffer, and 4-V μL of nuclease free water (15 μL volume total) in a 1.5 mL eppendorf tube (*see* **Note 7**).

Mix and incubate the samples at 55 °C for 15 min. Add 2 μ L of 10x loading dye, mix sample, rinse wells, then load immediately into gel.

4. Electrophorese at 50 Volts until the sample has entered the gel and then continue at 150 Volts until the dye front (from the BPB dye in the samples) has reached the bottom of the gel.

3.2 Small RNA Transfer by Capillary Blotting

1. Following electrophoresis, pry the gel plates open with a spatula. Pour 50 mL of TB into a squared petri dish. Transfer the gel to TB and soak for 10 min with gentle agitation.
2. Cut a membrane to the size of the gel and equilibrate it in distilled water in a squared petri dish for 1 min. Pour off water, add 20 mL of TB, and equilibrate for 5 min.
3. Set up the capillary blot (Fig. 1), and transfer the RNA overnight.
4. Dismantle the capillary blot and dry the membrane with the RNA side up for 3 min on Whatman 3MM paper. Cross-link the RNA to the membrane with UV at 120,000 μ J (Stratagene, UV Stratalinker 2400) (*see Note 8*).

3.3 Hybridization with a Radioactively Labeled Oligoprobe

1. Transfer the membrane into a hybridization tube with the RNA side facing the center of tube. Add 5–10 mL of hybridization solution and incubate at 40 °C for 30 min with slow rotation (pre-hybridization step).

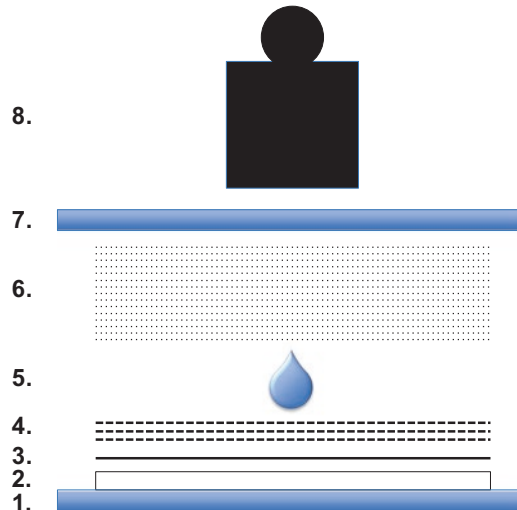


Fig. 1 Capillary blotting with 20 $\times$ SSC without reservoir. Set up the system following the numbers. Transfer the RNA overnight. 1. Clean glass plate; 2. Soaked gel; 3. Pre-wet membrane; 4. Three layers of pre-wet 3MM Whatman paper; 5. 10 mL of 20 $\times$ SSC; 6. Five centimeter thick pile of roll paper; 7. Glass plate; 8. weight (0.5 kg)

2. Mix 2 μL of 10 μM single-stranded DNA (ssDNA) oligo (21–24 nucleotide, reverse complement to the small RNA you would like to detect) with 10 μL of sterile distilled water in an eppendorf tube and incubate at 90 °C for 5 min. Chill the tube on ice for 3 min and add 2 μL of PNK buffer, 5 μL of $\gamma^{32}\text{P}$ -ATP, and 1 μL of T4 PNK.
3. Mix by pipetting 5 times and incubate the reaction mixture at 37 °C for 10–15 min.
4. Separate the labeled DNA from unincorporated nucleotides on a Microspin G-25 column according to the manufacturer's instruction. 1 μL of separated probe should count 500–2000 cps.
5. Add 2 μL of 0.5 M EDTA to the radioactive probe, mix by pipetting 5 times and denature by placing the tube at 90 °C for 5 min, then transfer the tube on ice.
6. Pour off the pre-hybridization solution. Mix the denatured ssDNA probe with 5 mL of fresh hybridization solution in a 15 mL Falcon tube and add to the hybridization tube.
7. Incubate at 40 °C overnight by slowly rotating the hybridization tube.
8. Pour off the hybridization solution and wash the membrane with excess of WB at 40 °C for 10 min.
9. Repeat the washing step for two more times.
10. Wrap the membrane with Saran wrap and expose to Phosphor Image plates. If necessary, increase the stringency of the washes by lowering the salt content of the washing buffer (i.e., 1 $\times$ SSC) or increasing the temperature during the wash.

4 Notes

1. This provides 10 mL volume, sufficient for casting two 0.75 mm gels using BIO-RAD mini-PROTEAN Tetra cell.
2. To speed up the dissolving of urea, place the 10 mL volume in microwave for a *maximum* of 10 s (without magnetic stirrer!).
3. We find that 10% ammonium persulfate performs best when made fresh, but can be kept at 4 °C for a maximum of 2 weeks.
4. Clean all gel loading and running apparatus with soap, rinse thoroughly with deionized ultrapure water, and drip-dry before use. This helps to reduce the possibility of RNase contamination that can degrade the sample.
5. Can use a 10 mL syringe and needle to squirt running buffer directly into wells. This should be performed immediately after comb removal and again just prior to loading the samples.

6. Typically, 5–15 µg of total RNA per lane is sufficient.
7. Formaldehyde, formamide, 10× MOPS, and nuclease-free water can be prepared as a master mix on ice before adding to each RNA sample.
8. After cross-linking, can make a small cut off the top right hand corner to help discern membrane orientation through subsequent steps.

Acknowledgment

C.J.H was supported by a BBSRC PhD Studentship. D.C.B. is the Royal Society Edward Penley Abraham Research Professor. This work was supported by the ERC Advanced Investigator grant ERC-2013-AdG 340642 TRIBE. A.M. is a Chancellor's Fellow at the University of Edinburgh.

References

1. Hamilton AJ, Baulcombe DC (1999) A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* 286:950–952
2. Zamore PD, Tuschl T, Sharp PA, Bartel DP (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 101:25–33
3. Molnár A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC (2007) miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* 447: 1126–1129
4. Terry B, Karol M, Du T (2004) Analysis of RNA by Northern and slot blot. *Curr Protoc Mol Biol* Chapter 4
5. Smith NA, Eamens AL, Wang M-B (2010) The presence of high-molecular-weight viral RNAs interferes with the detection of viral small RNAs. *RNA* 16:1062–1067
6. Harris CJ, Molnar A, Muller SY, Baulcombe DC (2015) FDF-PAGE: a powerful technique revealing previously undetected small RNAs sequestered by complementary transcripts. *Nucleic Acids Res* 43:7590–7599
7. Green MR, Sambrook J (2012) *Molecular cloning: a laboratory manual*, 4th edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor. ISBN: 978-1-936113-42-2

High-Throughput RT-qPCR for the Analysis of Circulating MicroRNAs

Geok Wee Tan and Lu Ping Tan

Abstract

Reverse transcription followed by real-time or quantitative polymerase chain reaction (RT-qPCR) is the gold standard for validation of results from transcriptomic profiling studies such as microarray and RNA sequencing. The current need for most studies, especially biomarker studies, is to evaluate the expression levels or fold changes of many transcripts in a large number of samples. With conventional low to medium throughput qPCR platforms, many qPCR plates would have to be run and a significant amount of RNA input per sample will be required to complete the experiments. This is particularly challenging when the size of study material (small biopsy, laser capture microdissected cells, biofluid, etc.), time, and resources are limited. A sensitive and high-throughput qPCR platform is therefore optimal for the evaluation of many transcripts in a large number of samples because the time needed to complete the entire experiment is shortened and the usage of lab consumables as well as RNA input per sample are low. Here, the methods of high-throughput RT-qPCR for the analysis of circulating microRNAs are described. Two distinctive qPCR chemistries (probe-based and intercalating dye-based) can be applied using the methods described here.

Key words High throughput, qPCR, Preamplification, microRNA, TaqMan, miScript

1 Introduction

The throughput of each qPCR run, as indicated by the amount of transcripts and samples that can be studied in one single qPCR plate/chip, can vary up to 100 times between platforms (Fig. 1). The demand for high-throughput qPCR has become very common, especially in studies which need to analyze a large panel of transcripts in a large number of samples. High-throughput qPCR is an answer to studies that utilize low starting input of samples such as laser capture microdissected tissues, single cells and cell-free nucleic acid in biofluid samples. Also, as compared to qPCR platforms of lower throughput, the time and cost needed to complete the experiments are reduced when high-throughput qPCR is utilized.

In high-throughput microfluidic qPCR platform, each qPCR reaction is carried out in a very small volume (range of nanoliter). Preamplification is necessary to increase the target amount prior to

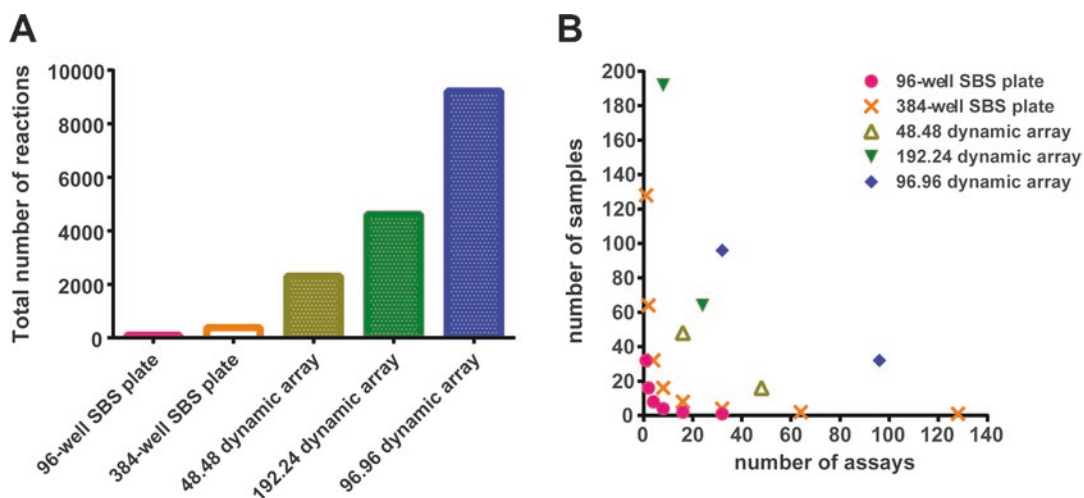


Fig. 1 (a) The numbers of reaction wells/chambers available in each qPCR plate/chip format are different and can vary up to 100 times. (b) The maximum numbers of assays and samples that can be analyzed in different format of plates/chips are compared under the setting of singleplex PCR and triplicate qPCR reactions per sample. Standard SBS plate format has lower throughput compared to high-throughput chips such as Fluidigm's dynamic array

qPCR so that targets/samples can be distributed equally among reaction chambers [1]. In order to prevent garbage in garbage out, cautious steps should be taken to evaluate if nonspecific amplifications and over amplification of targets are introduced during the preamplification step. Stringent quality control has to be applied on each primer assay by analyzing data from positive and negative controls to rule out nonlinear and/or nonspecific amplification.

For microRNA (miRNA) expression studies, different reverse transcription (RT) and qPCR chemistries are available. TaqMan is a probe-based system which is designed to specifically detect miRNAs with reference sequence registered in miRBase. On the other hand, miScript is an intercalating dye-based system which detects miRNAs with reference sequence registered in miRBase as well as all isomirs of the said miRNA. At times, miRNA with reference sequence registered in miRBase may not be the most abundant miRNA expressed in the samples of interest [2]. Therefore, the decision on using primer assay which is specific or generic will have to be based on research needs.

In this chapter, the protocols for RT, preamplification, and qPCR of both TaqMan and miScript systems are described. These protocols are optimized for the use with microfluidic chips in BioMark (Fluidigm) but nonetheless, the quality control (QC) principles developed and described here can be applied to all high-throughput RT-qPCR. Details are given on how to identify and exclude data from downstream analysis based on evidences of

nonspecific and/or nonlinear amplifications. Normalization methods for the analysis of circulating miRNAs data are also explained here. It is foreseeable that the lab methods described here might require modifications in the future due to the changes of reagents or protocols by manufacturers. Nonetheless, principle of data QC detailed here is always valid and should be applied to avoid garbage in, garbage out.

2 Materials

2.1 Positive and Negative Controls

1. RNA from pooled cell lines or pooled synthetic oligonucleotides (*see Note 1*).
2. Nuclease-free water.

2.2 TaqMan System

1. TaqMan MicroRNA Reverse Transcription kit: 100 mM dNTPs, RNase Inhibitor (20 U/ μ l), RT Buffer (10 $\times$), Multiscribe Reverse Transcription (50 U/ μ l).
2. TaqMan MicroRNA Assays: RT primer (5 $\times$) and real time primer (20 $\times$).
3. TaqMan PreAmp Master Mix (2 $\times$).

2.3 miScript PCR System

1. miScript II RT kit: miScript Nucleics Mix (10 $\times$), miScript HiSpec Buffer (5 $\times$), miScript Reverse Transcriptase Mix.
2. miScript Primer Assay (10 $\times$).
3. miScript PreAMP Buffer (5 $\times$).
4. HotStarTaqDNA Polymerase.
5. miScript PreAMP Universal Primer.
6. miScript Universal Primer (10 $\times$).
7. Side Reaction Reducer.

2.4 qPCR

1. Assay Loading Reagent (2 $\times$).
2. GE Sample Loading Reagent (20 $\times$) (for TaqMan system only).
3. DNA Binding Dye Sample Loading Reagent (20 $\times$) (for miScript PCR system only).
4. TaqMan Universal PCR Master Mix, no AmpErase UNG (for TaqMan system only).
5. SsoFast EvaGreen Supermix with Low ROX (for miScript PCR system only).
6. Dynamic Array IFC (integrated fluidic circuit) for gene expression (Fluidigm).

2.5 Others

1. Dilution buffer: TE (10 mM Tris, 0.1 mM EDTA, pH 8.0) or nuclease-free water.

3 Methods

3.1 Positive and Negative Controls

3.1.1 Serial Dilution of Positive Controls

1. Pool RNA from different cell lines or synthetic oligonucleotides into a 1.5 ml tube (*see Note 1*).
2. Prepare a serial dilution from the pooled RNA. Suggestion of serial dilution is shown in Table 1 (*see Note 2*).
3. Positive controls, together with experimental samples should be subjected to RT, preamplification and qPCR using the same reagents and conditions.

3.1.2 Negative Controls (See Note 3)

1. Use nuclease-free water as negative control in RT, preamplification and qPCR steps.
2. These negative controls should be analyzed in the same qPCR run together with the positive controls and experimental samples.

Table 1
Recommendations of serial dilutions for positive controls

Tube ^a	Dilution factor	Concentration for pooled total RNA (g/ μ l)	Concentration for each oligonucleotide (mol/ μ l)
1	1	1.0×10^{-8}	5.0×10^{-16}
2	1/2	5.0×10^{-9}	2.5×10^{-16}
3	1/4	2.5×10^{-9}	1.3×10^{-17}
4	1/8	1.3×10^{-9}	6.3×10^{-18}
5	1/16	6.3×10^{-10}	3.1×10^{-18}
6	1/32	3.1×10^{-10}	1.6×10^{-18}
7	1/64	1.6×10^{-10}	7.8×10^{-19}
8	1/128	7.8×10^{-11}	3.9×10^{-19}
9	1/256	3.9×10^{-11}	2.0×10^{-19}
10	1/512	2.0×10^{-11}	9.8×10^{-20}
11	1/1024	9.8×10^{-12}	4.9×10^{-20}
12	1/2048	4.9×10^{-12}	2.4×10^{-20}
13	1/4096	2.4×10^{-12}	1.2×10^{-21}
14	1/8192	1.2×10^{-13}	6.1×10^{-22}
15	1/16,384	6.1×10^{-14}	3.1×10^{-22}
16	1/32,768	3.1×10^{-14}	1.5×10^{-22}

^aIn the case that limited qPCR chambers/wells are available for positive controls, one can omit several tubes intermittently, e.g., tube no. 2, 3, 5, 6, 8, 9, 11, 13. More titration points in the lower end will allow one to maximize the interpolation range towards detection limit.

3.2 TaqMan System

3.2.1 Preparation of Reverse Transcription Primer Pool (See **Note 4**)

1. Thaw RT primers (5×) on ice, and vortex gently to ensure the content of the tube is well-mixed. Centrifuge briefly to bring down the content.
2. To prepare a 32-plex primer pool, transfer equal volume of TaqMan RT primer (intended to be in the RT primer pool) into a 1.5 ml tube. If less primer multiplexing is needed, adjust final volume with dilution buffer accordingly so that final concentration of 0.16× is achieved for each RT primer.
3. Pipette the primer pool up and down to mix well and keep at −20 °C until further use.

3.2.2 Reverse Transcription

1. Thaw the components of TaqMan MicroRNA Reverse Transcription kit, TaqMan RT primer pool and RNA samples on ice.
2. Mix each tube gently and centrifuge briefly to bring down the content in the tubes.
3. Prepare RT reaction master mix by combining 0.2 µl dNTPs (100 mM), 0.15 µl RNase inhibitor (20 U/µl), 1 µl Reverse Transcription Buffer (10×), 1 µl Multiscribe Reverse Transcriptase (50 U/µl), and 3 µl TaqMan RT primer pool for each RT reaction. Preparation of RT reaction master mix should include an additional 10% in volume to compensate for pipetting losses.
4. For each sample, aliquot 5.35 µl of RT reaction master mix to a PCR tube, add 4.65 µl of RNA sample (*see Note 5*) to each RT reaction and mix well.
5. Mix all components well and incubate on ice for 5 min.
6. Run thermal cycling with the following conditions: 16 °C for 30 min, 42 °C for 30 min, 85 °C for 5 min and hold at 4 °C indefinitely until RT products are retrieved.
7. Dilute the RT products 1:4 with dilution buffer and proceed to the preamplification step. If preamplification cannot be performed immediately, diluted RT products can be kept at −20 °C until further use.

3.2.3 Preparation of Preamplification Primer Pool (See **Note 4**)

1. Thaw real time primers (20×) on ice and vortex gently to ensure the content of the tube is well-mixed. Centrifuge briefly to bring down the content.
2. Add equal volume of each real time primer (20×) into a 1.5 ml tube.
3. Dilute the preamplification primer pool to a final concentration of 0.2× for each primer with dilution buffer.
4. Pipette the primer pool up and down to mix well.
5. Keep the TaqMan preamplification primer pool at −20 °C until further use.

3.2.4 Preamplification

1. Allow diluted RT products and TaqMan preamplification primer pool to thaw on ice. Mix the content in each tube by vortexing gently and centrifuge briefly to bring down the contents of the tubes.
2. Mix the content in TaqMan PreAmp Master Mix (2×) by flicking the tube and centrifuge briefly to bring down the content.
3. Prepare the preamplification reaction master mix by mixing 5 µl TaqMan PreAmp Master Mix (2×) and 2.5 µl TaqMan preamplification primer pool for each preamplification reaction. Preparation of preamplification reaction master mix should include an additional 10% in volume to compensate for pipetting losses.
4. Pipette the preamplification master mix up and down to mix well.
5. For each sample, aliquot 7.5 µl of preamplification reaction master mix to a PCR tube, add 2.5 µl diluted RT products to each preamplification reaction and mix well.
6. Incubate the reaction on ice for 5 min.
7. Run thermal cycling with the following conditions: denaturation at 95 °C for 10 min, followed by 16 cycles of preamplification at 95 °C for 15 s and 60 °C for 4 min (*see Note 6*).
8. Dilute the preamplified products 1:4 with dilution buffer and proceed with qPCR or store the preamplified products at −20 °C until further use.

3.2.5 Pre-qPCR Preparation

1. Inject control line fluid into both accumulators on the dynamic array chip.
2. Remove the protective film before inserting the chip into IFC controller and run the “Prime” script.

3.2.6 qPCR

1. For each assay mix, add 3 µl TaqMan Assay (20×) and 3 µl assay loading reagent (2×) into a well of a 96-well plate, and label this plate as plate A.
2. In a 1.5 ml tube, prepare master mix of sample pre-mix by combining 3 µl TaqMan Universal Master Mix (2×) and 0.3 µl GE sample loading reagent for each sample. Prepare an excess of 10% volume to account for pipetting losses.
3. For each sample mix, transfer 2.7 µl diluted preamplified products (Subheading 3.2.4) and 3.3 µl sample pre-mix from **step 2** into a well of another 96-well plate, and label this plate as plate B.
4. Pipette 5 µl assay mix from each well of plate A into the individual assay inlet and 5 µl sample mix from each well of plate B into the individual sample inlet on the dynamic array chips (*see Note 7*). If 48.48 dynamic array is used, it is advised to pipette negative control into inlet 22 (*see Note 8*).

5. Load the assay mix and sample mix by running the “Load Mix” script in the IFC controller.
6. Run the qPCR in BioMark by using the default protocol. Choose the protocol according to the qPCR chemistry and dynamic array chip that is used, e.g., Protocol GE 48×48 Standard v1.pcl if probe-based chemistry and 48.48 dynamic array is used.

3.3 miScript System

3.3.1 Reverse Transcription

1. Thaw miScript Reverse Transcriptase Mix, miScript Nucleics Mix (10×), miScript HiSpec Buffer (5×), and RNA samples on ice.
2. Mix the content of the tubes well and centrifuge briefly to bring down the content.
3. Prepare RT reaction master mix by combining 2 µl miScript Nucleics Mix (10×), 4 µl miScript HiSpec Buffer (5×), and 2 µl miScript Reverse Transcriptase Mix for each RT reaction. Preparation of RT reaction master mix should include an additional 10% in volume to compensate for pipetting losses.
4. For each sample, aliquot 8 µl of RT reaction master mix to a PCR tube, add 12 µl RNA (*see Note 9*) to each RT reaction and mix well.
5. Mix all the components well and incubate on ice for 5 min.
6. Run thermal cycling with the following conditions: 37 °C for 60 min, 95 °C for 5 min, and hold at 4 °C indefinitely until RT products are retrieved.
7. Dilute the RT products 1:5 with dilution buffer and proceed to the preamplification step. If preamplification cannot be performed immediately, diluted RT products can be kept at −20 °C until further use.

3.3.2 Preparation of Preamplification Primer Pool

1. Thaw miScript Primer Assay (10×) on ice. Mix the content of the tube gently and centrifuge briefly to bring down the content.
2. Add equal volume of each miScript primer assay into a 1.5 ml tube.
3. Pipette the primer pool up and down to mix well.
4. Dilute the preamplification primer pool to a final concentration of 0.4× for each primer with dilution buffer.
5. Keep the miScript preamplification primer pool at −20 °C until further use.

3.3.3 Preamplification

1. Thaw miScript PreAMP Buffer, HotStarTaqDNA Polymerase, miScript PreAMP Primer Mix, miScript preamplification primers pool, and diluted RT products on ice.
2. Mix the content of the tubes and centrifuge briefly to bring down the contents.

3. Prepare the preamplification reaction master mix by mixing 5 μ l miScript PreAMP Buffer (5 $\times$), 2 μ l HotStar Taq DNA Polymerase, 1 μ l miScript PreAMP Universal Primer, 5 μ l miScript preamplification primer pool, and 7 μ l nuclease free water for each preamplification reaction. Preparation of preamplification reaction master mix should include an additional 10% in volume to compensate for pipetting losses.
4. For each sample, aliquot 20 μ l of preamplification reaction master mix to a PCR tube, add 5 μ l diluted RT products to each preamplification reaction and mix well.
5. Incubate the reaction on ice for 5 min.
6. Run thermal cycling with the following conditions: denaturation at 95 $^{\circ}$ C for 15 min, followed by 12 cycles of preamplification at 94 $^{\circ}$ C for 30 s and 60 $^{\circ}$ C for 3 min (*see Note 6*).
7. Remove excess primers by adding 1 μ l side reaction reducer to each reaction and heat up the reactions at 37 $^{\circ}$ C for 15 min and 95 $^{\circ}$ C for 5 min.
8. Dilute the preamplified products 1:5 with dilution buffer and proceed to qPCR or store the preamplified products at -20° C until further use.

3.3.4 Pre-qPCR Preparation

1. Inject control line fluid into both accumulators on the dynamic array chip.
2. Remove the protective film before inserting the chip into IFC controller and run the “Prime” script.

3.3.5 qPCR

1. For each assay mix, add 1.5 μ l miScript Primer Assay (10 $\times$), 1.5 μ l miScript Universal Primer (10 $\times$), and 3 μ l assay loading reagent (2 $\times$) into a well of a 96-well plate, and label this plate as plate A.
2. In a 1.5 ml tube, prepare sample pre-mix by combining 3 μ l SsoFast EvaGreen Supermix with low ROX and 0.3 μ l DNA Binding Dye Sample Loading Reagent for each sample. Prepare an excess volume of 10% to account for pipetting losses.
3. For each sample mix, transfer 2.7 μ l diluted preamplified products and 3.3 μ l sample pre-mix from **step 2** into a well of another 96-well plate, and label this plate as plate B.
4. Pipette 5 μ l assay mix from each well of plate A into the individual assay inlet and 5 μ l sample mix from each well of plate B into the individual sample inlet on the dynamic array chips (*see Note 7*). If 48.48 dynamic array is used, it is advised to pipette negative control into inlet 22 (*see Note 8*).
5. Load the assay mix and sample mix by running the “Load Mix” script on the IFC controller.

6. Run the qPCR in BioMark by using the default protocol. Choose the protocol according to the qPCR chemistry and dynamic array chip that is used, e.g., Protocol GE 48×48 PCR+Melt v1.pcl if intercalating dye-based chemistry and 48.48 dynamic array is used.

3.4 Data Analyses and Quality Control

3.4.1 Analysis Settings

1. Open the .bml file in Real Time PCR Analysis software (Fluidigm).
2. Complete the assay (detector) setup and sample setup based on the mapping of assays and samples in the 96-well plates. Set sample type for positive controls as “Standard”. Specify the relative concentration of each positive control based on the serial dilution factor performed earlier.
3. Click on “Analysis Views”. For analysis settings, use default quality threshold, “Linear (Derivative)” for baseline correction, “User (Detectors)” for Ct threshold method. Click on “Ct Thresholds” tab and check on “Initialize with Auto”. As the term Ct is used in Fluidigm’s software, Ct is used in place of Cq (RDML data standard) throughout this article for easy reference.
4. In melt curve analysis, set the T_m range for each miScript assay as average T_m ± 2 standard deviations (*see Note 10*).

3.4.2 Quality Control and Data Analyses

1. For intercalating dye-based qPCR system, data point with T_m value out of the set range is considered invalid due to nonspecific amplification and should be omitted from analysis.
2. Evaluate the results of negative controls for each assay. If amplification signals are detected in these negative controls, data points with equal or higher Ct values than these negative controls should be omitted from analysis (*see Note 11*).
3. View the standard curves (serial dilution of positive controls) by clicking on the “+” sign next to “Analysis Views” and then click on “Calibration View”.
4. Check the standard curve of each assay to ensure that there is indication of linear amplification (linear regression slope, $m \approx -3.32$ and goodness of fit, $R^2 > 0.9$) (*see Note 12*).
5. Remove any assay from further analyses if there is no indication of linear and specific amplification (*see Note 13*).
6. For each primer assay which passes QC in **step 5**, identify the minimum and maximum Ct values in the interpolation range from the standard curve (Fig. 2 and *see Note 14*). Any data point of unknown samples which is beyond the assay interpolation range should be set to an arbitrary value for undetermined expression, e.g., Ct = 30.

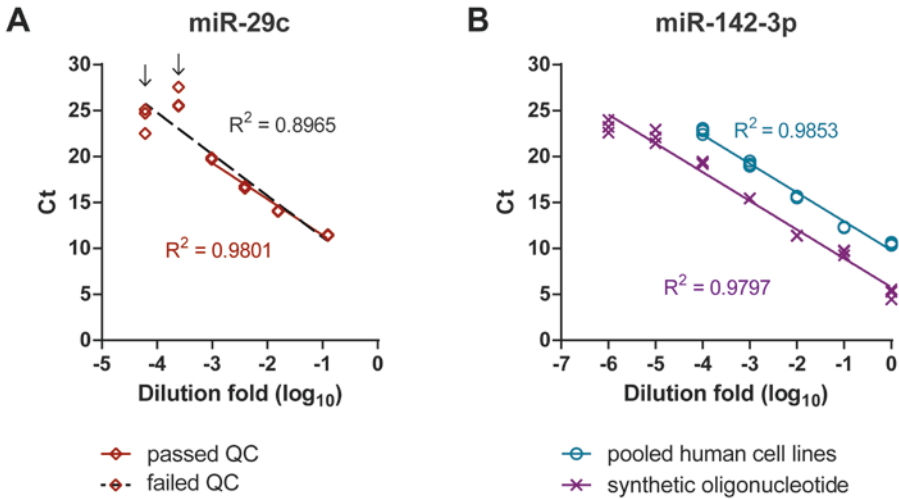


Fig. 2 Linear regression of data points from positive controls. (a) When nonlinear amplification is ignored and an interpolation range of six points from the positive controls are accepted (*dashed line* with $R^2 < 0.9$), fold change between the last two dilution points will be incorrectly calculated based on inaccurate C_t values (as indicated by *arrows*). In this case, for this particular assay miR-29c, the interpolation range for unknown samples should be restricted to the first 4 points of positive controls (*solid line*, $R^2 > 0.9$). (b) Some miRNAs may be less abundant in pooled human cell lines RNA. As a result, the dynamic range of miRNA which can be detected from the serial dilutions of pooled human cell lines RNA is narrower (C_t 10 to 24) than that from the serial dilutions of pooled synthetic oligonucleotides (C_t 5 to 24). Representative graphs shown here are assayed using protocols described in this chapter

7. Technical bias due to extraction can be normalized by using external spike-in controls (*see Note 15*). Normalization can be achieved with these calculations:

Average C_t = average C_t of qPCR replicates

Normalization factor for sample X = median of average $C_{t_{\text{spike-in}}}$ from all samples
 – average $C_{t_{\text{spike-in}}}$ of sample X

Normalized $C_{t_{\text{any assays}}}$ for sample X, dC_t = average $C_{t_{\text{any assays}}}$ for sample X
 + normalization factor for sample X

8. After data normalization, expression of miRNAs can be presented as fold change to reference sample using the relative quantification formula, 2^{-ddC_t} .

4 Notes

1. Positive controls should be used to construct standard curves for each primer assay. This is to ensure that each primer assay can be evaluated for the presence/absence of nonspecific

amplifications as well as over amplification. Positive controls can be created by pooling RNA from different cell lines related to the area of studies. Alternatively, pooled RNA from human cell lines and pooled synthetic oligonucleotides (miRNAs) are commercially available.

2. Typically, a serial dilution needs to show expression with dynamic range of at least five orders of magnitude and the expression levels detected from unknown samples should be within this range. The starting concentration of serial dilution can be as low as 10 ng/ μ l total RNA from pooled human cell lines or 5.0×10^{-16} mol/ μ l of each synthetic oligonucleotide (Table 1). When pooled RNA from human cell lines is used, low abundance transcripts will have a smaller dynamic range (Fig. 2b).
3. Negative controls are needed in all steps of RT-qPCR to rule out cross-contamination. During RT, preamplification and qPCR runs, negative controls should be included in each steps and given specific labels. Negative controls for preamplification (nuclease-free water in preamplification reaction) is especially important, as any positive amplification detected in qPCR from this negative control will indicate nonspecific amplification. Negative control to rule out genomic DNA (no RT enzyme in RT reaction) is not necessary as both TaqMan and miScript systems are not influenced by genomic DNA [3, 4].
4. RT and preamplification primer pools are prepared when custom primer pool is needed. These steps and pools are not required if TaqMan Megaplex primer or the new TaqMan Advanced miRNA Assays are used.
5. Amount of RNA required in the RT reaction is variable. As examples, valid Ct values can be derived from using 4.65 μ l RNA from 25 μ l eluted RNA extracted from 200 μ l plasma/serum [5] or from 10 ng cellular total RNA (data not shown).
6. The recommended amount of preamplification cycle is between 12 and 18 cycles. Preamplification protocols described in this chapter are 16 cycles for TaqMan system and 12 cycles for miScript system. Each primer assay should be evaluated independently during quality control step. Under the preamplification conditions described in this chapter, 1 out of 16 (6.3%) TaqMan assays and 2 out of 16 (12.5%) miScript assays showed nonspecific and/or nonlinear amplification [5].
7. Avoid generating bubbles into the inlets by pressing the plunger of the pipette to the first stop only. Introduction of air bubbles into the inlet will cause part of the reaction chambers to be filled with air instead of qPCR reaction mix. Data acquired from these affected chambers will not be accurate

and can be reflected by negative calls or high CV among qPCR replicates.

8. Reaction from inlet 22 is channeled to the reaction chambers located at the top edge of the IFC where evaporation can occur during the time course of qPCR. All formats of dynamic arrays have hydration inlets around the edges to reduce the effect of sample evaporation around the edge except for the 48.48 dynamic array. When using 48.48 dynamic array, it is recommended to put negative control instead of actual sample in inlet 22.
9. Amount of RNA required in the RT reaction is variable. As an example, valid Ct values can be derived from using 12 μ l RNA from 25 μ l eluted RNA extracted from 200 μ l plasma/serum [5].
10. In qPCR system which utilizes intercalating dye as detection chemistry, melt curve analysis is required for each primer assay in each sample. Any double peaks or unexpected T_m seen in the melt curve analysis is an indication of nonspecific amplification and/or noise from primer-dimer formation. The T_m range is meant to exclude data point with nonspecific amplification from further analysis. If isomirs are the subjects of interest, T_m range can be modified according to the calculation of expected T_m values for all isomirs.
11. When Ct values detected in experimental samples are equal or higher than those detected in negative control samples, one cannot distinguish whether these are real signals or merely noise from false positive.
12. Auto threshold setting may not be optimal for each primer assay. Under auto threshold setting, if linear amplification (linear regression slope between -3.10 and -3.58 and goodness of fit, $R^2 > 0.9$) is not seen in standard curve, one can adjust the threshold line manually in “Ct Thresholds” setting or consider reducing the interpolation range by omitting positive control data points at the end (solid line in Fig. 2a). Data points should not be omitted intermittently. If there is no indication of linear amplification after all these attempts, one needs to omit this primer assay from further analysis. For qPCR studies, calculating fold change between samples is the ultimate goal. If nonlinear amplification is ignored (dash line in Fig. 2a), fold change between samples can be wrongly calculated based on inaccurate Ct values (data points indicated by arrows in Fig. 2a).
13. Primer assays that have been optimized under standard RT-qPCR conditions may not be optimal for RT-preamp-qPCR.
14. In BioMark, low expression is represented by $Ct > 25$. Due to the variation in primer assay efficiencies, the cutoff (minimum and maximum Ct values) for each primer assay has to be based on the linear range of individual standard curve.

15. Commonly used external spike-in controls include synthetic *C. elegans* miRNAs (cel-miR-39 and/or cel-miR-54). During RNA extraction, 500 amol of synthetic oligonucleotide can be spiked-in after the addition of lysis buffer. Exogenous control is advised to be added after sample lysis to avoid degradation by endogenous RNases from plasma/serum.

Acknowledgment

We thank the Director General of Health Malaysia for his approval to publish this article. The work described here is supported by the Ministry of Health Malaysia (NMRR-11-597-9667). We also acknowledge the support of the Director of Institute for Medical Research (IMR) Malaysia and colleagues at the Molecular Pathology Unit in IMR.

References

1. Svec D, Rusnakova V, Korenkova V et al (2013) Dye-based high-throughput qPCR in microfluidic platform BioMark, PCR technology: current innovations. CRC Press, Boca Raton, pp. 323–339
2. Lee LW, Zhang S, Etheridge A et al (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. RNA 16:2170–2180
3. Chen C, Ridzon DA, Broomer AJ et al (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33:e179
4. QIAGEN (2011) miScript PCR system handbook
5. Tan GW, Khoo ASB, Tan LP (2015) Evaluation of extraction kits and RT-qPCR systems adapted to high-throughput platform for circulating miRNAs. Sci Rep 5:9430

Genome-Wide Comparison of Next-Generation Sequencing and qPCR Platforms for microRNA Profiling in Serum

Thorarinn Blondal, Maurizia Rossana Brunetto, Daniela Cavallone, Martin Mikkelsen, Michael Thorsen, Yuan Mang, Hazel Pinheiro, Ferruccio Bonino, and Peter Mouritzen

Abstract

This study compares next-generation sequencing (NGS) technologies that have been optimized specifically for biofluid samples, with more established qPCR-based methods for profiling microRNAs in biofluids. The same patient serum samples were analyzed by NGS and qPCR, and differences in the serum microRNA profile between HBV and HCV infected patients were investigated. While there was overall good agreement between NGS and qPCR, there were some differences between the platforms, highlighting the importance of validation.

Key words microRNA, miRNA, Next-generation sequencing, NGS, qPCR, Real-time PCR, Biofluids, Serum, Plasma, Platform comparison, HBV, HCV, Liquid biopsy, Profiling, Validation

1 Introduction

Profiling microRNAs in biofluids is challenging due to the limited amount of RNA present in biofluids, as well as presence of inhibitory compounds which have the potential to inhibit downstream enzymatic processes. In addition, the presence of cellular components may lead to contamination of the “cell-free” biofluid microRNA profile, e.g., through hemolysis (lysis of red blood cells).

It is important to standardize sample collection protocols and to monitor any potential sources of pre-analytical variability through rigorous Quality Control (QC) procedures [1]. Each step of the workflow from RNA isolation to RNA QC and NGS library preparation or RT-qPCR needs to be optimized for challenging samples like biofluids with limited RNA content.

We have optimized protocols to maximize detection of microRNA while minimizing carryover of any compounds from biofluid samples, which may inhibit downstream enzymatic processes. Quality checks have been implemented in every step of the protocols to monitor performance and ensure high-quality data.

In order to compare the results obtained using NGS and qPCR platforms for microRNA profiling in serum, genome-wide microRNA profiling was performed at Exiqon Services on the same serum samples from ten different individuals by both NGS (library preparation using the NEBNext® Small RNA Library Prep kit followed by sequencing on the Illumina NextSeq500) and qPCR (miRCURY LNA™ Universal RT microRNA PCR System).

We selected five patients with hepatitis B virus (HBV) and five with hepatitis C virus (HCV) infections, which are major causes of chronic hepatitis worldwide [2]. Improved noninvasive biomarkers are needed to manage these patients, and serum microRNAs may represent promising candidates as this new class of biomarkers has an important role in the interaction between virus and host [3]. In a previous qPCR microRNA profiling study, liver-derived microRNAs were found to be detected at high levels in the sera of HBV patients, and a microRNA signature associated was discovered that could help identify patients with both natural and therapy induced immune control of chronic HBV infection [4].

2 Materials

1. Clinical serum samples (*see Note 1*).
2. Serum and plasma pools.
3. miRCURY™ RNA Isolation Kit—Biofluids.
4. MS2 carrier RNA.
5. miRCURY™ RNA Spike-Ins.
6. miRCURY™ RNA Spike-In Kit.
7. miRCURY LNA™ qPCR Assays (*see Note 2*).
8. NEBNext Small RNA Library Prep kit.
9. Universal cDNA Synthesis Kit II, Human miRNome PCR Panels I + II V3.
10. ExiLENT SYBR® Green Master Mix Kit.
11. LightCycler® 480 Real-Time PCR System.
12. QiaQuick columns.
13. Bioanalyzer 2100.
14. LabChip XT.

3 Methods

All protocols from RNA isolation to RNA QC and NGS library preparation or RT-qPCR are specifically optimized for analysis of microRNAs in challenging biofluid samples with limited RNA content (Table 1). The optimized protocols are designed to maximize reads/signals from microRNAs while minimizing carryover of any compounds from the biofluid samples which may inhibit downstream enzymatic processes.

3.1 Optimized Protocols for Biofluids microRNA NGS

1. Centrifuge 500 µl human serum to remove cells and debris prior to freezing. Spin for 5 min at 3000 × g, then remove 450 µl supernatant and either isolate RNA immediately or store the supernatant in RNase-free tubes (e.g., cryo-tubes) at −80 °C.

Table 1
Optimized NGS and qPCR methods for biofluids microRNA profiling. The optimized protocols for NGS microRNA Biofluids are available in Exiqon Services. The optimized protocols for qPCR analysis of microRNAs in biofluids are available in the instruction manuals for the relevant Exiqon products (see Subheading 2). The serum samples were first centrifuged to remove cells and debris, and RNA isolation was performed using 450 µl supernatant (for NGS) or 200 µl supernatant (for qPCR). For details of the qPCR-based quality control of Biofluid RNA samples see Note 2

NGS	qPCR
<i>RNA isolation from 500 µl serum</i>	<i>RNA isolation from 250 µl serum</i>
• Proprietary protocol optimized for NGS microRNA biofluids • Ultra-low elution volumes • Include 52 RNA spike-ins for RNA QC and sequencing QC	• miRCURY™ RNA isolation Kit—Biofluids • Include RNA spike-ins for RNA QC (miRCURY™ RNA Spike-In Kit) • Use carrier RNA (MS2)
<i>RNA QC</i>	<i>RNA QC</i>
• qPCR-based QC to monitor RNA isolation efficiency, inhibition and detect outliers • Hemolysis indicator • Spike-in controls • Endogenous microRNA controls	• qPCR-based QC to monitor RNA isolation efficiency, inhibition and detect outliers • Hemolysis indicator • Spike-in controls • Endogenous microRNA controls
<i>Library preparation</i>	<i>RT-qPCR</i>
• Proprietary protocol optimized for biofluids with low concentration of starting material (based on NEBNext® Small RNA Library Prep Kit) • Size selection to maximize microRNA reads • QC of library by Bioanalyzer and qPCR	• miRCURY LNA™ Universal RT microRNA PCR System • Human miRNome Panel + II (742 microRNAs were analyzed) • Protocol optimized for biofluid samples
<i>microRNA sequencing</i>	
• Illumina platform (NextSeq500) • 1 × 50 bp reads, 10 M raw reads per sample	

2. Isolate RNA from 450 μ l supernatant including the addition of 52 RNA spike-ins (synthetic microRNAs of plant origin) for the purposes of RNA QC and Sequencing QC.
3. Perform RNA QC using a qPCR-based method for assessment of RNA quality [1] using RNA spike-ins and endogenous microRNAs to monitor RNA isolation efficiency, inhibition and detect outliers and signs of hemolysis. *See Note 2.*
4. Ligate adaptors to the 3' and 5' ends of small RNAs in each individual RNA sample, and convert into cDNA. Pre-amplify the cDNA with primers containing sample specific indexes.
5. After 18 cycles of pre-PCR, purify the libraries on QiaQuick columns and evaluate the insert efficiency using the Bioanalyzer 2100 instrument with the high sensitivity DNA chip. Calculate concentrations using area under the peak, and pool the samples together in equimolar concentrations prior to library size selection. This minimizes any technical variation introduced during the library size selection step. *See Note 3.*
6. Size-select the microRNA cDNA libraries using a LabChip XT and excise a band representing adaptors plus 15–30 bp insert, following the manufacturer's instructions. This step is crucial to maximize the number of microRNA reads. *See Note 3.*
7. Quantify the library pool(s) using qPCR and determine the optimal concentration of the library pool to be used to generate the clusters on the surface of a flowcell.
8. Perform sequencing on the Illumina NextSeq500 using v2 High Output sequencing methodology according to the manufacturer's instructions, 1 $\times$ 50 bp reads, 10 M raw reads per sample.

3.2 Optimized Protocols for Biofluids microRNA qPCR

1. Centrifuge 250 μ l human serum to remove cells and debris prior to freezing. Spin for 5 min at 3000 $\times g$, then remove 200 μ l supernatant and either isolate RNA immediately or store the supernatant in RNase-free tubes at -80°C .
2. Isolate RNA from 200 μ l supernatant using the miRCURYTM RNA Isolation Kit—Biofluids with the addition of 1 μ g MS2 carrier RNA, and miRCURYTM RNA Spike-Ins for RNA QC.
3. Perform RNA QC using a qPCR-based method for assessment of RNA quality [1] using RNA spike-ins and endogenous microRNAs to monitor RNA isolation efficiency, inhibition and detect outliers and signs of hemolysis. *See Note 2.*
4. Reverse transcribe 16 μ l RNA in 80 μ l reactions using the miRCURY LNATM Universal cDNA Synthesis Kit II. Dilute the cDNA 50 $\times$ and assay in 10 μ l PCR reactions according to the biofluids protocol for miRCURY LNATM Universal RT

microRNA PCR using the microRNA Ready-to-Use PCR, Human miRNome Panels I + II and ExiLent SYBR® Green Master Mix kit. Perform the real time PCR amplification in a LightCycler® 480 Real-Time PCR System in 384 well plates. Reverse-transcribe a blank purification negative control (water instead of serum in the RNA isolation) and profile it alongside the samples.

3.3 Analysis Pipelines for Biofluids NGS and qPCR

1. Different data analysis pipelines have been developed for NGS and qPCR data from biofluid samples (Table 2). First perform quality control checks on the raw data. In the case of qPCR analysis, remove from the analysis any amplifications that fail to meet the defined acceptance criteria (Table 2). Remove sample Cq values that are less than 5 Cq below the negative control (blank purification) in order to ensure that the microRNA signals are sufficiently different to any background signal. For Biofluids NGS (in addition to standard NGS data QC) analyze 52 RNA spike-ins added during the RNA isolation step, to monitor the reproducibility and linearity of the library preparations and sequencing reactions.
2. Map the NGS reads to miRBase and to the appropriate reference genome (in this case the human reference genome, assembly GRCh37) using Bowtie2 genome mapper, and identify novel microRNAs and isomiRs. Novel microRNAs are predicted using algorithms based on miRPara [5].
3. For both NGS and qPCR data, apply a threshold to the data prior to normalization and differential expression analysis. The appropriate threshold will depend on the project, but the threshold is designed to focus the analysis on the most reliable microRNA reads/signals (based on TPM or Cq value, or on detection in a group or percentage of samples). Methods for normalization and differential expression analysis also differ between NGS and qPCR (*see* Table 2). One important difference is that the EdgeR package used for NGS differential expression analysis is able to include microRNAs where the read numbers are zero in several samples or even a whole group; however, this is not the case with the methods used for qPCR differential expression analysis.

3.4 Reproducibility of Biofluids NGS and qPCR

The technical reproducibility of the entire NGS and qPCR workflow from RNA isolation to microRNA profiling can be assessed by comparing results from independent RNA isolations performed on different days, using the same biofluid sample. The Biofluids microRNA NGS workflow shows excellent reproducibility when comparing the TPM normalized counts obtained from endogenous microRNAs between replicate RNA isolations (Pearson $R^2 = 0.9494$, Fig. 1). Similarly, the Biofluids microRNA qPCR

Table 2
Analysis pipelines for NGS and qPCR. Data quality control and processing applied for NGS and qPCR data from serum samples. NGS data quality assessment includes reports on GC content, Kmer content (stretches of identical bases or overrepresentation of certain sequence motifs), per base N content (undetermined bases), and per base sequence content. The samples in this study were mapped to the human reference genome, assembly GRCh37. The thresholds applied to the data are selected depending on the particular project

NGS	qPCR
<i>RNA sequencing data</i> (FASTQ files)	<i>RNA qPCR data</i> (Cq values)
<i>Data QC and filtering</i>	<i>Data QC</i>
• Base and read quality • Data quality assessment • Adapter trimming • Identify, remove, and analyze 52 RNA spike-ins (sequencing linearity and reproducibility)	• T_m and melting curve analysis • PCR efficiency • Comparison with negative control (blank purification) • RNA spike-ins and interplate calibrators
<i>Mapping and counting</i>	<i>Filtering</i>
• miRBase • Reference genome • Other reference sources (if applicable), e.g., Rfam, SmallRNA • Abundant sequences (outmapped) <i>Assembly of novel data</i> • Prediction of novel microRNAs (miRPara) • Mapping to other species in miRBase • Identification of isomiRs	• Flag and remove amplifications that fail to meet acceptance criteria • Dissociation curve with single clean peak • Amplicon melting temperature consistent for the same assay between samples • PCR efficiency shows no sign of inhibition • Sample Cq at least 5 Cq below negative control • Remove assays flagged in > $x\%$ of sample^a • Remove samples flagged in > $x\%$ of assays^a
<i>Apply threshold</i>	<i>Apply threshold</i>
• TPM threshold • Detected in a group or $x\%$ of samples^a	• Cq value threshold • Detected in a group or $x\%$ of samples^a
<i>Normalization</i>	<i>Normalization</i>
• Trimmed Mean of M values (EdgeR package)	• Global mean (mean expression value of all microRNAs expressed in all samples after filtering)
<i>Differential expression analysis</i>	<i>Differential expression analysis</i>
• Negative binomial Exact test (EdgeR package)	• ANOVA and Wilcoxon tests

^aCustom analysis offered in Exiqon Services. In this study, microRNA qPCR data was normalized to the mean expression value of all microRNAs detected in all samples after filtering ($n = 127$ assays). *TPM* Tags Per Million mapped reads

workflow shows excellent reproducibility between replicate RNA isolations ($R^2 = 0.9637$, Fig. 2). This confirms that the RNA isolation and profiling platforms themselves do not introduce any significant source of variability. The microRNA profile of an individual also appears to be relatively stable over time (*see* **Note 4**).

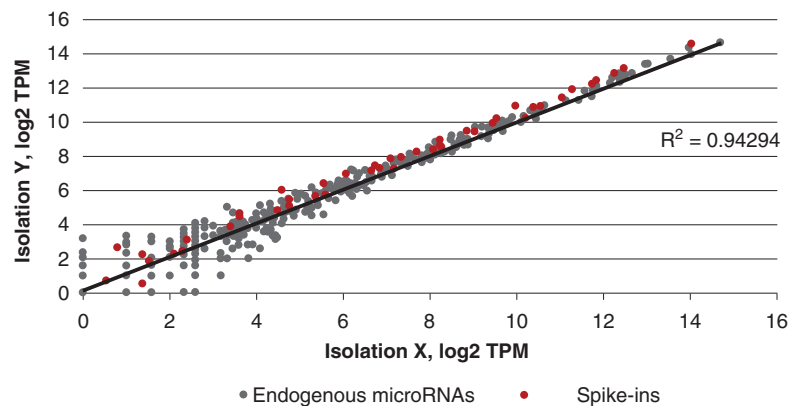


Fig. 1 Excellent technical reproducibility of microRNA sequencing from biofluids. Two independent RNA isolations were performed using the same pool of plasma on different days, followed by two independent library preparations and sequencing runs on a NextSeq500 instrument. Correlation performed using filtered data: only endogenous microRNAs detected >1 TPM. Endogenous microRNA TPMs are Tags Per Million mapped reads and Spike-in TPMs are Tags Per Million reads. Correlation displayed is for the endogenous microRNAs ($R^2 = 0.9494$). Correlation for Spike-ins is $R^2 = 0.9875$

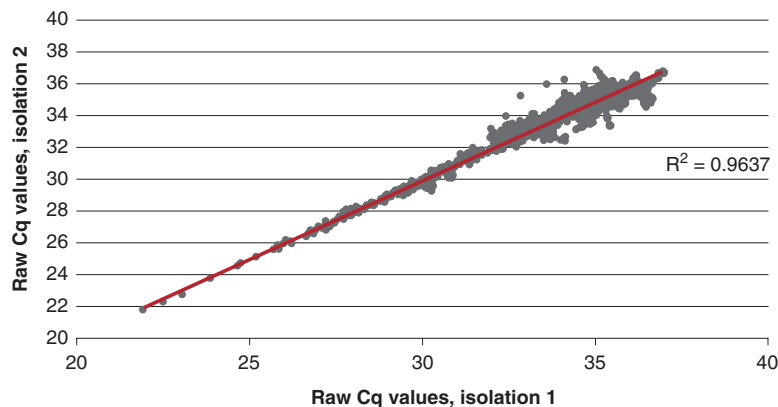


Fig. 2 Excellent technical reproducibility of microRNA qPCR from biofluids. Two independent RNA isolations were performed using the same plasma sample on different days, followed by two RT-qPCR microRNA profiling experiments using Human Panels I + II. Correlation performed using filtered data

The inclusion of controls such as the 52 RNA spike-ins (for NGS) and the miRCURY™ RNA Spike-ins and interplate calibrators (for qPCR) is important to monitor the reproducibility between samples in a profiling project. The 52 RNA spike-ins added during RNA isolation (prior to NGS library preparation) span the dynamic range of most endogenous microRNAs on the sequencing platform, and show excellent reproducibility between

replicate RNA isolations and sequencing runs (R^2 correlation of NGS spike-ins in Fig. 1 is 0.9875). The average NGS spike-in correlation (R^2) across the ten HBV and HCV serum samples was 0.976. The 52 RNA spike-ins used for NGS data QC represent typically 0.1–1% of the total number of reads per sample.

3.5 Analysis of Biofluids microRNA NGS Reads

A high percentage of the reads obtained from the NGS analysis could be mapped to miRBase or to the human genome (on average 80% mappable reads for the HBV group and 74% for the HCV group). This suggests that the samples and libraries prepared were of good quality, also indicated by the high quality of the sequencing data obtained (base and read Q-scores were >30 which is equivalent to >99.9% accuracy).

The NGS libraries were size-selected in order to maximize the reads from the relevant microRNA fraction. However, there was some sample-to-sample variation observed in the percentage of microRNA reads. The HBV serum samples contained on average a higher percentage of microRNA reads than the HCV serum samples (Fig. 3). This difference is attributed to the large amount of liver-derived microRNA present in the serum of HBV patients, and liver-derived microRNA have been shown to be exported and circulated in the blood within hepatitis B surface antigen (HBsAg) particles [6].

It is notable that just a few microRNAs make up a sizeable fraction of the NGS reads from serum (Fig. 4). The top ten microRNAs with the highest average number of counts across the HCV

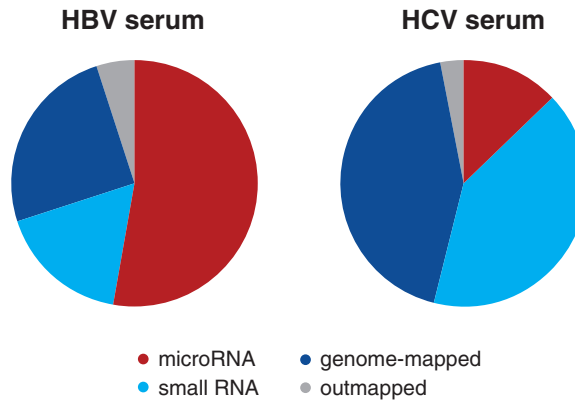


Fig. 3 The percentage of microRNA reads in serum varies between different sample types. microRNAs in HBV serum samples represented on average 53% of mappable reads, whereas in the case of HCV serum samples the percentage was lower (13%). This is attributed to the large amount of liver-derived microRNA present in the serum of HBV patients. The remaining mappable reads can be categorized as small RNA (of which fragments of tRNAs, Y RNAs and snRNAs are the most prevalent), other genome-mapped RNAs (degradation products of longer RNAs including mRNA or lncRNA), and outmapped (abundant sequences like mitochondrial and ribosomal RNA as well as homopolymers)

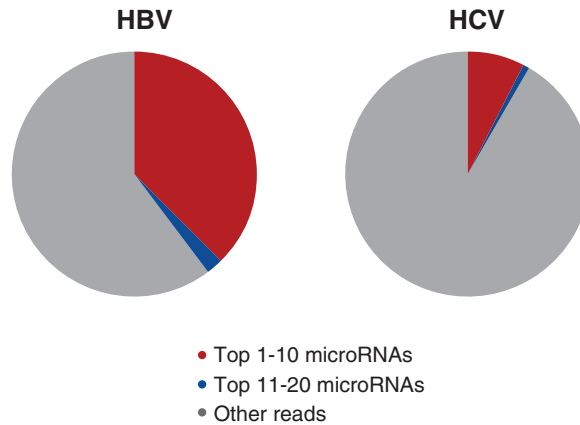


Fig. 4 A few microRNA make up a sizeable fraction of the NGS reads from serum. The percentage of total reads accounted for by the top 1–10 and top 11–20 microRNAs (top 1–10 microRNAs are those with the highest average number of counts across the HBV or HCV groups). The top 10 microRNAs account for 37.5% of the total reads from HBV serum samples, or 7.6% of the total reads from HCV serum samples. This difference is largely due to the high levels of miR-122-5p in the serum of HBV patients, which alone accounts for 31.6% of the total reads in the HBV serum samples

group of serum samples account for 7.6% of the total reads. In the case of the HBV serum samples, the top ten microRNAs account for 37.5% of the total reads. This difference is largely due to the dramatically increased levels of miR-122-5p in the serum of HBV patients, which alone accounts for 31.6% of the total reads in the HBV group.

This illustrates that measurement of a particular microRNA by NGS is not independent of other microRNAs, so if a particular microRNA is highly abundant or if certain samples contain a large amount of microRNA, this can reduce the number of reads available for other microRNAs. Therefore care must be taken to ensure that the sequencing depth is sufficient to allow accurate analysis even of microRNAs found at low levels in biofluids.

3.6 Number of microRNAs Detected in Serum

The number of microRNAs detected did vary between the different serum samples, and was also influenced by the threshold applied to the dataset (Fig. 5). The average number of microRNAs detected per sample in the HBV and HCV groups was fairly consistent with results from other projects conducted by Exiqon Services involving serum/plasma samples from patients with non-liver diseases.

Up to around 900 different microRNAs may be detected by NGS in human serum/plasma when no threshold is applied to the data. However, many of these microRNAs are detected at very low read numbers (e.g., below 10 absolute counts). Low read numbers in NGS data may represent accurate reads from microRNAs

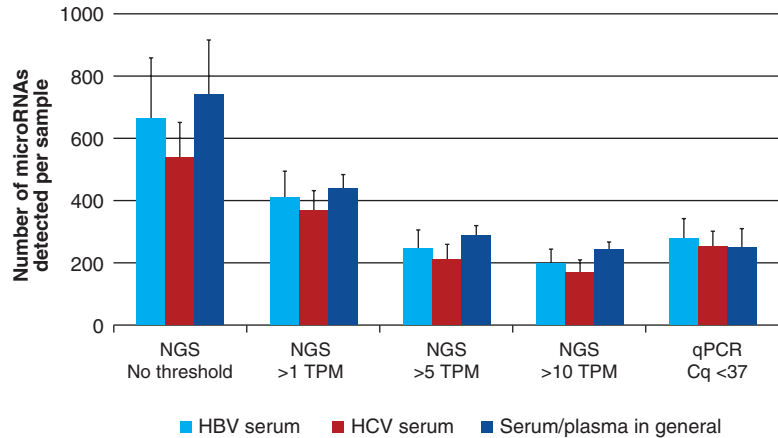


Fig. 5 The number of microRNAs detected in serum depends on the threshold applied. Average number of microRNAs detected per sample by NGS or qPCR. Results are shown for the HBV ($n = 5$) and HCV ($n = 5$) serum samples, as well as human serum/plasma samples (from patients with non-liver diseases) profiled in a variety of different projects by Exiqon Services using NGS ($n = 165$ samples) or qPCR ($n = 1264$ samples). Thresholds were applied to the data: values >1 , >5 , or >10 *TPM* in the case of NGS data, or values $<Cq$ 37 passing filtering in the case of qPCR data. *TPM* Tags Per Million mapped reads. Error bars represent the standard deviation

present at low levels, however small RNA library ligation bias can lead to overrepresentation of rare sequences, as well as underrepresentation of abundant sequences [7]. Low read numbers are sometimes difficult to validate due to inherent variability. Increasing the sequencing depth does increase the number of microRNAs detected, but does not alter the number of microRNAs detected when applying a *TPM* threshold (*see Note 3*).

The qPCR panels contained assays to detect a maximum of 742 microRNAs. The qPCR data was also subjected to rigorous data QC and filtering (Table 2), followed by application of a threshold. Applying a threshold based on Tags per Million mapped reads (*TPM*) in the case of NGS data, or *Cq* value in the case of qPCR data, can focus the analysis on the most reliable microRNA reads/signals. Overall, the threshold applied has the largest impact on the number of microRNAs detected per sample.

3.7 Overlap in microRNAs Detected by NGS and qPCR

When comparing the full list of microRNAs detected in any of the five serum samples within each group by NGS or qPCR, there is significant overlap: 323 microRNAs were detected by both platforms in the group of HBV serum samples (out of 684 microRNAs detected by NGS and 408 microRNAs detected by qPCR) (Fig. 6). Similar results were obtained for the group of HCV serum samples (Fig. 6).

Several microRNAs were detected by one platform but not the other; different RNA isolation methods, small RNA library

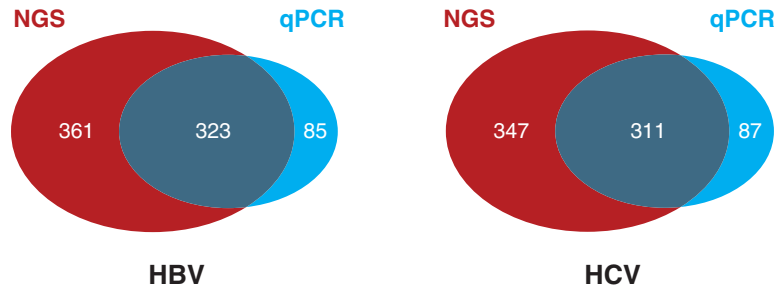


Fig. 6 There is significant overlap in the microRNAs detected in serum by NGS and qPCR in the HBV and HCV groups. microRNAs detected in at least 1 out of 5 samples per group, with >1 TPM or Cq value <37 after filtering, were included. In the case of the HBV serum samples, 323 microRNAs were detected by both NGS and qPCR and 85 microRNAs were detected only by qPCR. 361 microRNAs were detected in the HBV serum samples only by NGS. Note that the qPCR panels contain assays to detect 742 microRNAs. See text for further explanation

ligation bias [7], as well as the presence of isomiRs which mean the qPCR assays may detect only a fraction of the microRNA sequences (Subheading 3.12) all contribute to the differences observed between the NGS and qPCR platforms.

Interestingly, several microRNAs were detected in serum only by qPCR: 85 microRNAs in the HBV group and 87 microRNAs in the HCV group. The absence of these microRNAs from the NGS data may be due to small RNA library ligation bias resulting in underrepresentation of some sequences [7]. Alternatively, it may reflect higher sensitivity of qPCR compared to NGS. The miR-CURY LNA™ Universal RT microRNA PCR System has been demonstrated to the platform with highest specificity [8], and all qPCR assays have been rigorously wet-lab validated. Thus, we are confident that these are valid signals, and not false positives arising from cross-reactivity of the qPCR primers.

NGS is a hypothesis-free approach, with the possibility to detect any microRNA sequence (within the sensitivity range of the platform), whereas qPCR can detect only those 742 microRNAs included on the qPCR panels used in this study. Consistent with this, a substantial number of microRNAs were detected in serum only by NGS: 361 microRNAs in the case of the HBV group and 347 in the case of the HCV group. The use of custom qPCR assays would enable analysis of any microRNAs including novel microRNAs or isomiRs by qPCR following identification by NGS screening (*see* Subheading 3.13).

The majority of the “NGS only” microRNAs are not present on the qPCR panels, which contained assays to detect 742 microRNAs. However, we noted that qPCR assays are present for 83 of the “NGS only” microRNAs in the HBV group, yet these microRNAs were not detected by qPCR (similarly, qPCR assays are present for

88 of the “NGS only” microRNAs in the HCV group). Inspection of the NGS data revealed that the majority of these NGS reads resulted from isomiRs of the particular microRNA. Indeed, in several cases, the miRBase sequence was entirely absent from the NGS data. In addition, these microRNAs tended to be present at lower levels than microRNAs that were consistently detected by both platforms. The combination of a microRNA being at a lower level and comprising mainly of isomiRs differing from the miRBase sequence may make it challenging to detect using a qPCR assay designed to the miRBase sequence. For validation purposes, it may be relevant to design custom qPCR assays to detect the particular isomiR sequence (*see* Subheading 3.12).

3.8 microRNA Levels Measured by NGS and qPCR

The number of reads measured by NGS for a particular microRNA sequence is not directly related to its abundance, i.e., NGS is not used for absolute quantification. Therefore NGS data, like microarray data, should be normalized and analyzed to identify relative differential expression between samples. In this study, both NGS and qPCR platforms were used to perform relative quantification. Bearing this in mind, we investigated the correlation between microRNA levels measured by NGS and qPCR for the HBV and HCV groups of serum samples. As expected the correlation between microRNA levels measured by NGS (average log2TPM) and qPCR (average delta Cq) was low ($R^2 = 0.3918$ for the HBV group of serum samples), however there was an overall tendency for microRNAs with high detection levels measured by NGS to also show high detection levels by qPCR (Fig. 7).

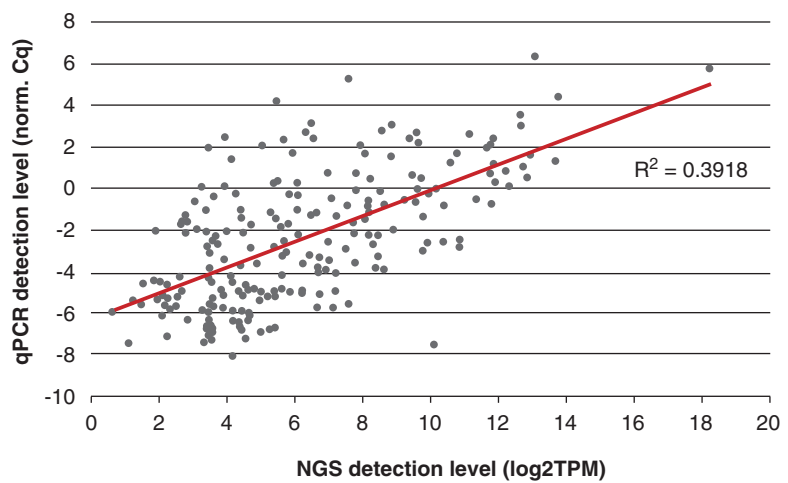


Fig. 7 Correlation in detection levels of serum microRNAs measured by NGS and qPCR. Data from 209 microRNAs are presented. Two criteria for inclusion were applied: (1) microRNAs that are detected with sum of TPM > 10 across all HBV and HCV samples, and (2) microRNAs that are detected in at least one out of five HBV serum samples with >1 TPM or Cq value <37 after filtering

Notably, microRNAs that were detected at lower levels by NGS (with <100 TPM on average across all HBV and HCV samples) displayed a poorer correlation in detection levels between platforms ($R^2 = 0.1201$ for the HBV group of serum samples, data not shown). This may reflect the fact that small RNA library ligation bias can result in overrepresentation or underrepresentation of sequences [7].

3.9 Correlation in Log Fold Changes Identified by NGS and qPCR

The purpose of an expression analysis study is to identify microRNAs which are differentially regulated between sample groups. In this study we looked for microRNAs differentially expressed between the groups of HBV and HCV serum samples. Overall, log-fold changes in microRNA expression levels between HCV and HBV serum samples showed a broad agreement between NGS and qPCR ($R^2 = 0.6054$, Fig. 8a). MicroRNAs that were detected at higher expression levels (with >100 TPM on average across all HBV and HCV samples) displayed a greater correlation in log fold changes between platforms ($R^2 = 0.8212$, Fig. 8b).

The direction of change (up or downregulation) was usually consistent between NGS and qPCR, however, a few microRNAs did show opposite trends on each platform, particularly if the microRNAs were present at <10 TPM (Fig. 9) or if the log fold changes were small (less than 2 log fold change between sample groups) (Fig. 10). Interestingly, certain liver-derived microRNAs (miR-122-5p, miR-192-5p, miR-99a-5p), were found to be detected at lower levels in serum from HCV patients compared to HBV patients. The high level of liver-derived microRNAs in HBV patients is consistent with published results, where the same microRNAs were found at higher levels in the sera of chronic HBV patients compared with inactive carriers [4].

These results clearly demonstrate the value of independent validation of NGS results. We strongly believe that microRNAs identified as differentially expressed by NGS must be validated using an independent, unbiased and highly accurate method like qPCR. In particular, NGS reads in the lower TPM range, or small fold changes between sample groups, require more careful validation.

3.10 Overlap in microRNAs with Significant Differences in Abundance by NGS and qPCR

While the log fold change and direction of change did show broad agreement between the NGS and qPCR platforms, there was generally a poor correlation between the p -values obtained by NGS and qPCR (data not shown). This is to be expected as the NGS and qPCR platforms use different methods for data normalization and statistical analysis. This will result in different false discovery rates, making p -values difficult to compare between platforms.

Nevertheless, we compared the microRNAs identified by NGS and qPCR as displaying significant differences in abundance between the HBV and HCV groups of serum samples. In order to include as many microRNAs as possible in the comparison, microRNAs with significant p -values ($p < 0.05$) before correction for

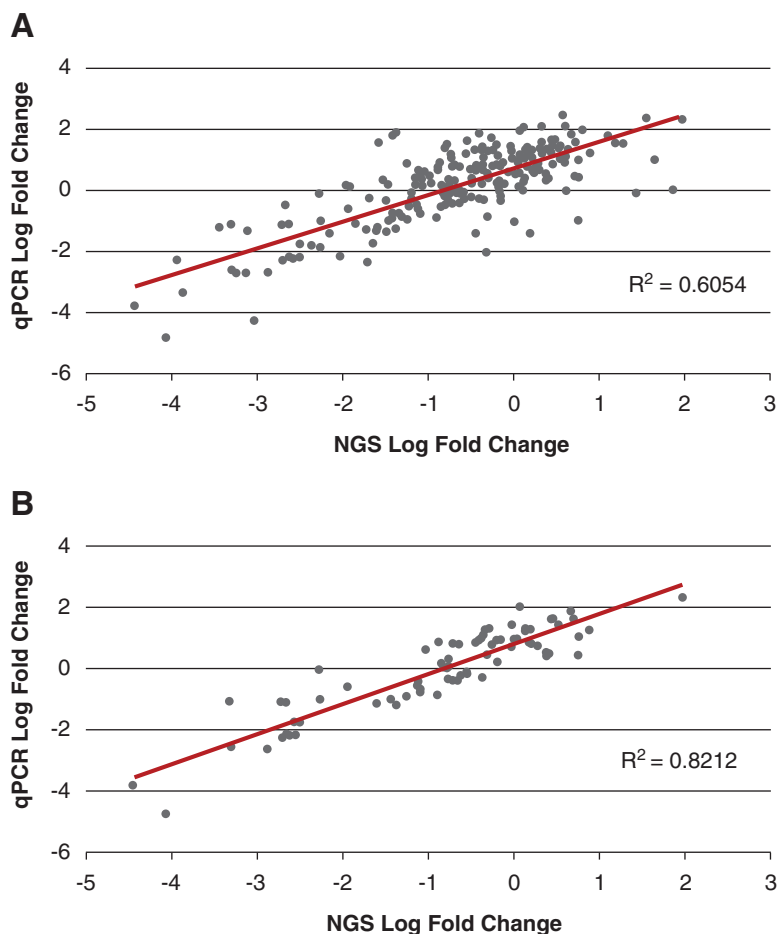


Fig. 8 Correlation in log fold changes of serum microRNAs identified between HBV and HCV groups by NGS and qPCR. Criteria for inclusion are the same as described in Fig. 7. (a) Correlation for 209 microRNAs. (b) Correlation for a subset of the more abundant microRNAs (75 microRNAs each with an average TPM >100 across all HBV and HCV samples)

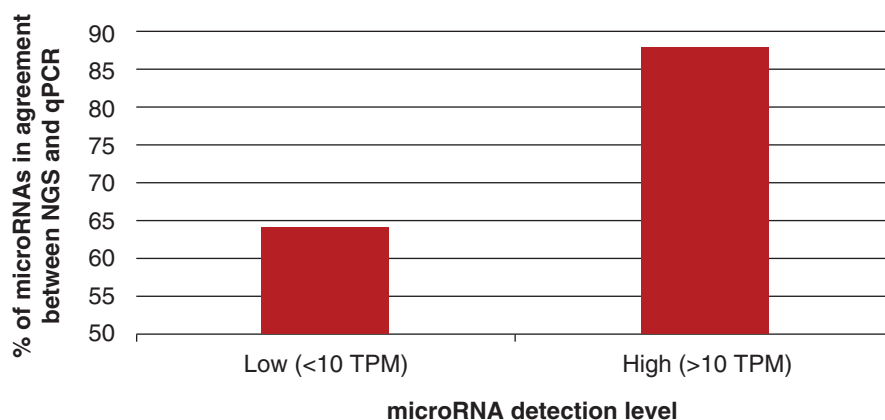


Fig. 9 Percentage of microRNAs in agreement between NGS and qPCR regarding direction of fold change. The agreement between NGS and qPCR is higher among microRNAs with a higher average detection level in NGS (>10 TPM). Results from 56 microRNAs are presented. Criteria for inclusion are the same as described in Fig. 7. In addition, difference was >1 log fold change between HBV and HCV groups in NGS

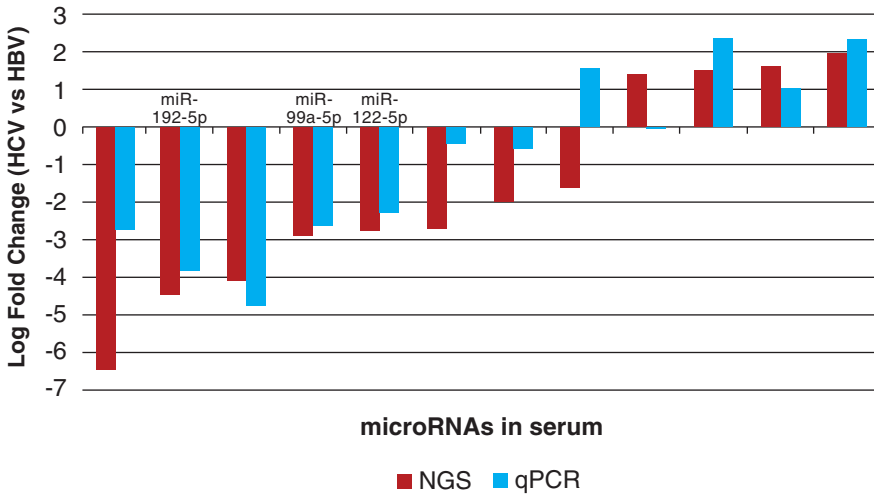


Fig. 10 Comparison of serum microRNA log fold changes identified by NGS and qPCR. Log fold change in microRNA detection levels between HCV and HBV serum samples show a broad agreement between NGS and qPCR. microRNAs previously identified as differentially detected in the sera of chronic HBV patients compared with inactive carriers [4] are indicated by name. Criteria for inclusion are the same as described in Fig. 7. In addition, microRNAs display significantly different levels HCV vs. HBV by NGS (p -value < 0.05). A subset of the microRNAs changed between HCV and HBV are shown

multiple testing were included in the analysis (p -values are before correction for multiple testing unless otherwise stated). 47 microRNAs were identified as displaying significant differences in abundance by NGS, and of these, 19 microRNAs were confirmed to have significantly different levels by qPCR (Fig. 11). 30 microRNAs were identified as significantly different only by qPCR.

This study included only five patients per group, and it is likely that a larger number of individuals per group would enhance the statistical power of the study, and increase the number of microRNAs identified by both platforms with statistically significant differences in abundance.

3.11 Novel microRNAs

The NGS analysis (miRPara prediction algorithm) was also able to identify 11 predicted novel microRNAs present in the serum samples, which would require validation using custom miRCURY LNA™ microRNA qPCR assays. This is consistent with our expectations that serum would not contain a high number of novel microRNAs, as the serum microRNA profile is dominated by microRNAs which are highly expressed in the larger tissues and organs in the body, which have been largely well characterized. Seven of the 11 novel microRNAs detected in this study were detected at levels >10 TPM (average across all ten samples), but none of the novel microRNAs were detected at significantly different levels between HBV and HCV groups.

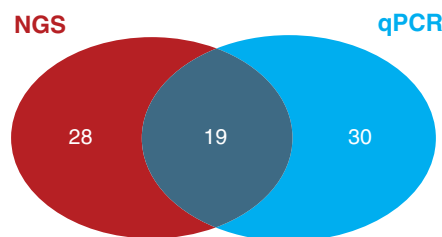


Fig. 11 Overlap in significantly differentially detected serum microRNAs identified by NGS and qPCR. See text for further explanation. Criteria for inclusion are the same as described in Fig. 7. In addition, microRNAs display significantly different levels HCV vs. HBV by NGS and qPCR (p -value < 0.05)

3.12 IsomiRs

IsomiRs differ in their sequence or length compared to the sequence listed in miRBase, but are derived from the same precursor microRNA. IsomiRs are detected by NGS, but they are poor targets for qPCR assays which are designed to detect the miRBase sequence. In the NGS data analysis, isomiR reads are mapped to the related miRBase sequence, and the combined reads from all isomiRs for a particular microRNA are included in the differential expression analysis.

Where differences are found in the NGS and qPCR results for a particular microRNA, it can be relevant to examine the NGS reads for the individual isomiRs, to evaluate whether the isomiRs present in a particular sample differ significantly from the miRBase sequence. If this is the case, validation can be performed using custom miRCURY LNA™ microRNA qPCR assays designed to a particular isomiR sequence.

The impact of isomiRs on the analysis is potentially significant. In one HBV serum sample, isomiRs represented 29% of the microRNA counts. We noted that there were differences between the HBV and HCV groups in terms of the distribution of miR-122-5p isomiRs: the sequence listed in miRBase appears to be less abundant in HCV (9.6% of total reads) compared to HBV (26.9% of total reads), whereas an isomiR lacking 1 nt at the 3'-end appears to be predominant in HCV patients (64.3% of total reads) compared to HBV patients (32.1% of total reads).

We have also noted that biofluid samples tend to contain more isomiRs, particularly isomiRs truncated at the 3' end. However it remains to be investigated whether the ends of biofluid microRNAs could be partially subjected to nuclease activity, even though microRNAs are known to be protected in association with Ago2 complexes, HDL proteins as well as within exosomes [9–11]. MiR-122 is the most abundant liver microRNA and the differential expression of its isomiRs in HBV versus HCV infected patients is consistent with the evidence that liver derived microRNA are carried in the blood within HBsAg particles [6]; in addition the HBsAg envelope may protect Ago2-microRNA complexes from

nuclease activity, reducing the relative number of 3' truncated isomeric microRNAs. HBsAg was found to co-localize with Ago2 and several liver microRNA including miR-122 and miR-99 in multiple subcellular compartments [12].

Lowly abundant microRNAs that consist mainly of isomiRs differing from the miRBase sequence can be challenging to detect using a qPCR assay designed to the miRBase sequence (*see* Subheading 3.7). More highly abundant microRNAs appear to be composed of a greater number of different isomiRs, which may mean they are detected by qPCR, even if the miRBase sequence is not the most abundant in the sample.

The ability of a standard qPCR assay (designed to the miRBase sequence) to detect isomiR sequences will depend on several factors: the difference in microRNA sequence length, the number and location of sequence differences compared to the miRBase sequence, and the qPCR primer design. Exiqon's miRCURY LNA™ microRNA qPCR System has been demonstrated to display very high specificity [8], hence isomiRs with sequence differences compared to the miRBase sequence will be poorly detected by the Exiqon qPCR System, but may still be amplified by less specific qPCR systems. IsomiRs with deletions at the 3'-end of the mature microRNA sequence are particularly poor targets for stem-loop primer based qPCR [13].

3.13 Choosing the Most Optimal Platform (NGS vs. qPCR)

The optimal choice of platform for a particular study will depend on several factors, including the goals of the study. One consideration is whether the discovery of novel microRNAs and identification of isomiRs is important for the study. NGS takes a hypothesis-free approach and is able to detect "any microRNA" present at sufficiently high levels, including any predicted novel microRNA or isomiR. On the other hand, qPCR enables very accurate and sensitive analysis of differential expression between sample groups, using panels containing a preselected set of microRNA assays. The standard qPCR assays are designed to a specific microRNA sequence (usually the one listed in miRBase), however custom qPCR assays can be designed to amplify isomiRs that differ from the sequence in miRBase.

Another consideration is the amount of biofluid sample available. The NGS and qPCR protocols described here are optimized for microRNA profiling from biofluids and recommend a larger starting amount of biofluid for NGS (500 µl serum/plasma) compared to qPCR (250 µl serum/plasma).

We find that the combination of NGS and qPCR together is a powerful approach. An initial screening can be performed by NGS on a smaller number of the samples, in order to identify a subset of microRNAs that are relevant for profiling in a larger number of samples. Custom qPCR panels are ideal for profiling a selected subset of microRNAs in a larger number of samples in order to

identify differentially regulated microRNAs. The candidate microRNAs can then be validated, ideally in an independent set of samples, again using custom qPCR panels. Crucially, for a novel microRNA or isomiR is identified by the NGS screening, a custom qPCR assay can be designed to detect the new microRNA or isomiR, and this new assay can also be included on the custom qPCR panels. In this way, the study benefits from the advantage of NGS for discovery purposes, while at the same time taking advantage of the accuracy of qPCR for differential expression analysis. This approach is also highly cost-effective, enabling many more samples to be included in the study, and thereby increasing statistical confidence in the results.

3.14 Discussion

There are few other studies that have compared NGS and qPCR for microRNA profiling in biofluids. The miRQC study also found an overall good concordance of differential expression status between NGS and Exiqon's qPCR platform: 85–90% of microRNAs differentially expressed by NGS could be correctly validated with Exiqon's qPCR system [8].

The miRQC study found sequencing platforms to be less accurate than qPCR platforms for the analysis of low copy number microRNAs in samples such as biofluids [8]. This is consistent with our observations that microRNAs at the lower TPM range show inferior correlation with qPCR, in terms of detection level, log fold change and direction of change.

While platform sensitivity has not been directly compared in this study, the observation that several microRNAs were detected in serum only by qPCR (85 microRNAs for HBV serum samples and 87 microRNAs for HCV serum samples) is in agreement with the miRQC study finding that sequencing platforms display low sensitivity for samples such as biofluids with low RNA input amounts [8]. A study comparing NGS and qPCR for analysis of microRNAs in urine (a biofluid with even lower RNA content than serum) identified 111 microRNAs that could only be detected by qPCR [14], indicating higher sensitivity of qPCR compared to NGS as a possible reason for this.

Nassirpour et al. [14] also raised concerns about the suitability of qPCR for microRNA analysis due to the presence of isomiRs (40% of microRNAs identified in urine by NGS were present only in isomeric forms). In our study, log fold changes correlated well between NGS and qPCR for the more highly abundant microRNAs, in spite of the presence of isomiRs and the fact that standard qPCR assays were used (designed to the miRBase sequence). It is possible that the presence of isomiRs may be one reason for the poorer correlation between NGS and qPCR for more lowly abundant microRNAs; however, custom-designed qPCR assays can be included in a validation study to investigate the impact of isomiRs.

This study demonstrated that the HBV serum samples contained on average a higher percentage of microRNA reads than the HCV serum samples, and identified several microRNAs (including liver-derived microRNAs) present in higher levels in the HBV serum samples. Recent evidence suggests that the liver-derived microRNAs may be exported into the bloodstream within defective hepatitis B surface antigen (HBsAg) particles [6, 12], opening the possibility that around 300 million HBsAg carriers worldwide could be ideal subjects for the noninvasive study of the dynamics of liver epigenetics under different pathophysiological conditions.

3.15 Conclusions

Reproducible, high-quality microRNA profiling data can be obtained from biofluids by both NGS and qPCR, provided that protocols and analysis pipelines specifically optimized for biofluids are used. The number of microRNAs detected in serum by NGS and qPCR was similar, after filtering the data and applying thresholds, which are important to focus the analysis on the most reliable microRNA reads/signals.

There was good agreement between NGS and qPCR in terms of the microRNAs detected in serum. Detection levels between the NGS and qPCR platforms did correlate to some degree, even though both platforms are used for relative quantification. There was overlap in the microRNAs identified by NGS and qPCR with significant differences between groups, and a broad correlation in log fold changes identified by the two platforms, in spite of the presence of isomiRs. There were some differences between platforms, clearly indicating that validation of results using an independent method is always required, particularly for microRNAs with lower detection levels or smaller fold changes.

Combining NGS and qPCR in the study design enables the discovery of novel microRNAs and isomiRs by NGS, followed by accurate differential expression analysis using custom qPCR panels. This approach is also highly cost-effective, enabling many more samples to be included in the study, thereby increasing statistical confidence in the results.

4 Notes

1. It is important to standardize sample collection protocols and to monitor any potential sources of pre-analytical variability through rigorous QC procedures. The presence of cellular components may lead to contamination of the “cell-free” biofluid microRNA profile, e.g., through hemolysis (lysis of red blood cells). Hemolysis can occur at the time of phlebotomy, so it is important to ensure that blood is collected by staff highly experienced in venepuncture, using a high quality col-

lection tube and consistent procedures for handling samples. It is recommended to process the whole blood immediately into either serum or plasma. If the whole blood is not centrifuged immediately after collection (or after the clotting time, in the case of serum), the samples should be stored at room temperature for no longer than 4 h. Storage at 4 °C may result in lysis of thrombocytes and contamination of the cell-free microRNA profile. In our experiments serum samples were collected from five untreated patients with chronic hepatitis B and five untreated patients with chronic hepatitis C following standard hospital protocols. The two groups of patients were matched for sex, age, grade and stage of liver disease, demographic features, and liver disease cofactors and without comorbidities.

2. Standard methods for measurement of RNA yield and quality, e.g., OD260 measurement, are not reliable for biofluid RNA due to the RNA content being below the reliable range of the instrument, and the presence of contaminants which can confound the readings. Therefore we have developed a qPCR-based method for assessment of RNA quality ([1] and Table 3)

Table 3
Spike-ins and endogenous microRNAs used for qPCR-based QC of Biofluid samples. The RNA spike-ins and corresponding qPCR assays (Exiqon A/S) are used to monitor the efficiency and yield of the RNA isolation, the performance of the cDNA synthesis and PCR reactions as well as hemolysis and general sample quality

	Recommended use	Application for Biofluid QC
<i>Spike-ins</i>		
UniSp2, UniSP4 and UniSp5	Add RNA spike-ins to lysis buffer during RNA isolation (three different concentrations)*	Monitor RNA isolation efficiency
UniSp6 and cel-miR-39-3p	Add RNA spike-ins to cDNA synthesis reaction (different concentrations)*	Check for RT and PCR inhibitors
52 RNA spike-ins (nonhuman)	Add RNA spike-ins to lysis buffer during RNA isolation (range of concentrations)**	Monitor RNA isolation efficiency and sequencing QC
UniSp3	Template and primers present in PCR panels	Independent PCR monitoring, inter-plate calibration
<i>Endogenous microRNAs</i>		
miR-451a and miR-23a-3p	Check ΔCq (miR-23a-3p-miR-451a)	Hemolysis indicator
miR-30c-5p, miR-103a-3p, miR-124-3p and miR-191-5p	Measure levels of endogenous microRNAs expected to be present in biofluids (serum/plasma, urine, CSF) by RT-qPCR	General sample quality check

*: applied when RNA is applied in qPCR

** : applied when RNA is applied in NGS

for which RNA spike-ins are added during the RNA isolation and cDNA synthesis steps, and measured by RT-qPCR alongside a range of endogenous microRNAs.

- Library insert efficiency is assessed using the Agilent Bioanalyzer. The peak of fluorescence at 127 nt represents adaptors without inserts (Fig. 12a). A peak around 148 nt represents the microRNA population ligated between adaptor sequences which will be selected when the library is size fractionated. It is important to see at least a shoulder of a peak representing the microRNA population. An additional peak is also often present at approx 160 nt, which represents other degraded RNA molecules (including rRNA and other small RNAs like Y RNAs and tRNAs). The read length distribution (Fig. 12b) after library size selection and adaptor trimming usually shows three main peaks: a classical microRNA peak around 20–24 nt, a peak around 30–34 nt which represents degraded Y RNAs, tRNAs and other small RNAs (may include piRNA, however piRNAs are rarely detected in biofluids other than seminal fluid or CSF), rRNAs

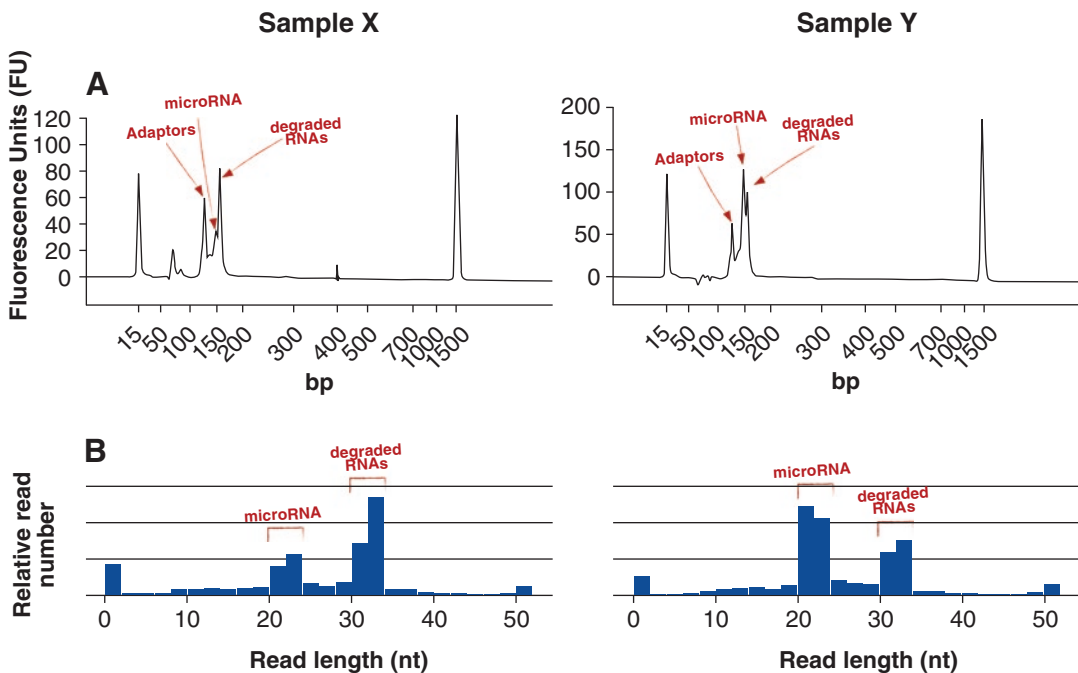


Fig. 12 NGS Biofluids microRNA library QC. Two samples illustrating low and high microRNA content (Samples X and Y respectively) are shown. (a) Insert efficiency is assessed using the Agilent Bioanalyzer before library size selection. The peak of fluorescence at around 127 nt represents adaptors without inserts. A peak of fluorescence around 148 nt represents the microRNA population including adaptor sequences which will be selected when the library is size fractionated. An additional peak is also present at 157 nt, which represents other degraded RNA molecules (including rRNA and other small RNAs like tRNAs). (b) Read length distribution of the samples shown in (a), after library size selection and adaptor trimming. Three main peaks are usually seen: a classical microRNA peak around 20–24 nt, a peak around 30–34 nt which represents degraded tRNA, and a peak around 50 nt which represents longer degraded RNAs

and other longer RNA species, and a peak around 50 nt which represents longer degraded RNAs of unknown size from various origins (these appear as 50 nt since the sequencing specifications limit the reads to 50 bp). The relative abundance of the peaks observed on the Agilent Bioanalyzer trace may vary from sample to sample. For example, Sample Y shows a much higher microRNA peak compared to the degraded RNA peak, whereas in Sample X the opposite is true (Fig. 12a). This difference is reflected in the size distribution of the reads obtained: Sample Y with the higher microRNA peak also shows a higher relative number of reads in the microRNA size range (20–24 nt, Fig. 12b). Note that the degraded RNA peak (around 157 nt) is often dominant in biofluid samples and difficult to size-exclude. This can result in technical variation between samples, due to variability introduced in the library size selection step. Therefore we pool samples to be analyzed together prior to library size selection. The samples are pooled together in equimolar concentrations based on DNA quantification using “area under the peak” from 140 to 160 nt size range. All samples should be pooled and excised in the same manner, preferably as one pool. When gel-excising the microRNA band, try to avoid the empty adaptors and the degraded RNA peak, in order to maximize the microRNA content.

4. Variations in the microRNA profile of biofluids from the same individual may be expected over time due to a variety of factors, e.g., circadian rhythm, diet etc. We compared the microRNA profile of the five HCV baseline serum samples with serum samples taken from the same individuals 6 months after a course of antiviral interferon therapy to which they were classified as nonresponders (data not shown). Only two out of 394 microRNAs showed significantly different levels (with a p -value < 0.05 after Benjamini–Hochberg false discovery rate correction) between the baseline and post-treatment serum samples, and both of these microRNAs were detected at low levels (average TPM of three or below in both groups). One of these microRNAs was present on the qPCR panels, but there was no significant difference in detection level when analyzed by qPCR. This suggests that the microRNA profile of serum collected from the same individual even months apart is relatively stable. However it must be noted that these samples were paired pre- and post-treatment serum samples, and hence the treatment may have some influence on the microRNA profile, even if the individuals did not respond to the therapy. Increasing sequencing depth does not increase the number of microRNAs reliably detected in serum. A pool of human serum and plasma was prepared prior to RNA isolation, and the same pool was sequenced with either 10 M or 60 M raw

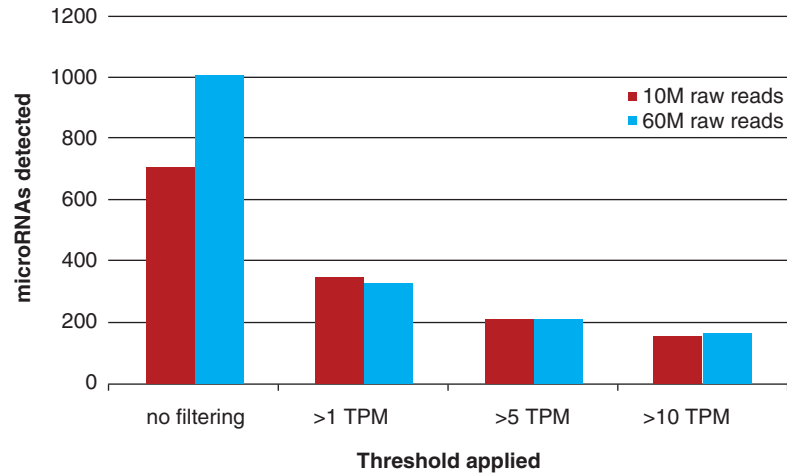


Fig. 13 Increasing sequencing depth does not increase the number of microRNAs detected after TPM normalization. A pool of human serum and plasma was prepared, and the same pool was sequenced with either 10 M or 60 M raw reads per sample. Sequencing with 60 M raw reads does result in detection of more microRNAs, but after normalization to the total number of reads per sample, the number of microRNAs detected per sample at given TPM thresholds is very similar to the that obtained with 10 M raw reads

reads per sample. Sequencing with 60 M raw reads does result in detection of more microRNAs, but after normalization to the total number of reads per sample, the number of microRNAs detected per sample at given TPM thresholds is very similar to the that obtained with 10 M raw reads (Fig. 13).

Acknowledgment

We thank Jannie Rosendahl Christensen, Tine Charlotte Sandberg, and Efat Sadat for excellent technical assistance.

References

1. Blondal T, Jensby Nielsen S, Baker A, Andreasen D, Mouritzen P, Wrang Teilm M, Dahlsveen IK (2013) Assessing sample and miRNA profile quality in serum and plasma or other biofluids. *Methods* 59(1):S1–S6
2. Lavanchy D (2008) Chronic viral hepatitis as a public health issue in the world. *Best Pract Res Clin Gastroenterol* 22(6):991–1008
3. Fan HX, Tang H (2014) Complex interactions between microRNAs and hepatitis B/C viruses. *World J Gastroenterol* 20(37):13477–13492
4. Brunetto MR, Cavallone D, Oliveri F, Moriconi F, Colombatto P, Coco B, Ciccorossi P, Rastelli C, Romagnoli V, Cherubini B, Teilm MW, Blondal T, Bonino F (2014) A serum microRNA signature is associated with the immune control of chronic hepatitis B virus infection. *PLoS One* 9(10):e110782
5. Wu Y, Wei B, Liu H, Li T, Rayner S (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding

- regions in genome scale sequences. *BMC Bioinformatics* 12:107
6. Novellino L, Rossi RL, Bonino F, Cavallone D, Abrignani S, Pagani M, Brunetto MR (2012) Circulating hepatitis B surface antigen particles carry hepatocellular microRNAs. *PLoS One* 7(3):e31952
 7. Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, Ojo T, Luo S, Schroth G, Tuschl T (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* 17(9):1697–1712
 8. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L, Derveaux S, Feng Y, Fulmer-Smentek S, Gerstmayer B, Gouffon J, Grimley C, Lader E, Lee KY, Luo S, Mouritzen P, Narayanan A, Patel S, Peiffer S, Rüberg S, Schroth G, Schuster D, Shaffer JM, Shelton EJ, Silveria S, Ulmanella U, Veeramachaneni V, Staedtler F, Peters T, Guettouche T, Wong L, Vandesompele J (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods* 11(8):809–815
 9. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL, Stirewalt DL, Tait JF, Tewari M (2011) Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc Natl Acad Sci U S A* 108(12):5003–5008
 10. Vickers KC, Palmisano BT, Shoucri BM, Shamburek RD, Remaley AT (2011) MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat Cell Biol* 13(4):423–433
 11. Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO (2007) Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol* 9(6):654–659
 12. Hayes CN, Akamatsu S, Tsuge M, Miki D, Akiyama R, Abe H, Ochi H, Hiraga N, Imamura M, Takahashi S, Aikata H, Kawaoka T, Kawakami Y, Ohishi W, Chayama K (2012) Hepatitis B virus-specific miRNAs and Argonaute2 play a role in the viral life cycle. *PLoS One* 7(10):e47490
 13. Wu H, Neilson JR, Kumar P, Manocha M, Shankar P, Sharp PA, Manjunath N (2007) miRNA profiling of naïve, effector and memory CD8 T cells. *PLoS One* 2(10):e1020
 14. Nassirpour R, Mathur S, Gosink MM, Li Y, Shoieb AM, Wood J, O'Neil SP, Homer BL, Whiteley LO (2014) Identification of tubular injury microRNA biomarkers in urine: comparison of next-generation sequencing and qPCR-based profiling platforms. *BMC Genomics* 15:485

Small RNA Profiling by Next-Generation Sequencing Using High-Definition Adapters

Martina Billmeier and Ping Xu

Abstract

Small RNAs (sRNAs) as key regulators of gene expression play fundamental roles in many biological processes. Next-generation sequencing (NGS) has become an important tool for sRNA discovery and profiling. However, NGS data often show bias for or against certain sequences which is mainly caused by adapter oligonucleotides that are ligated to sRNAs more or less efficiently by RNA ligases. In order to reduce ligation bias, High-definition (HD) adapters for the Illumina sequencing platform were developed. However, a large amount of direct 5' and 3' adapter ligation products are often produced when the current commercially available kits are used for cloning with HD adapters. In this chapter we describe a protocol for sRNA library construction using HD adapters with drastically reduced direct 5' adapter–3' adapter ligation product. The protocol can be used for sRNA library preparation from total RNA or sRNA of various plant, animal, insect, or fungal samples. The protocol includes total RNA extraction from plant leaf tissue and cultured mammalian cells and sRNA library construction using HD adapters.

Key words Small RNA, Next-generation sequencing, Small RNA profiling, Library construction, Reduction of ligation bias

1 Introduction

During the last decade several classes of regulatory small noncoding RNAs (sRNAs) with a length ~20–30 nucleotides (nt) have been identified [1, 2]. They play important roles in regulating gene expression, chromatin structure and chromosome segregation in sequence-specific manner, and are involved in many biological processes such as development, defense, stress response, disease, and maintenance of genome stability [3, 4]. Thus, sRNA discovery and profiling are important approaches for understanding the molecular regulation networks underlying diverse biological processes.

There are three major approaches to sRNA profiling: hybridization-based methods such as Northern blot and microarray analysis, quantitative reverse transcription-PCR (qRT-PCR), and sequencing of cDNA libraries of sRNAs. The advantages and disadvantages of these methods have been well reviewed by

Pritchard et al. [5]. Among all three types of methods, only sRNA cloning and sequencing allows the discovery of new diverse sRNA sequences of different sizes. In particular, the development of next-generation sequencing (NGS) technology has enabled high throughput sRNA profiling (sRNA-Seq) over a wide range of expression levels. Several NGS platforms are available for massive parallel sequencing of sRNAs, including Illumina, Applied Biosystems (ABI) SoLiD and 454 Life Sciences pyrosequencing (Roche), while the Illumina platform is the most popular one for sRNA analysis due to its advantage in short nucleic acid fragment sequencing.

NGS analysis of sRNA profiling requires construction of corresponding cDNA libraries, which can be done using sRNA cloning kits that are commercially available. These protocols typically involve ligating customized oligonucleotides (adapters) to both ends of sRNAs for further reverse transcription (RT) and PCR amplification. The ligation reactions are catalyzed by T4 RNA ligases including T4 RNA ligase 1 and 2 (T4Rnl1 and T4Rnl2) and more recently the truncated mutant T4Rnl2 for reducing undesired side reactions [6–8]. However, the ligation efficiency of T4 RNA ligases are affected by the sequence and structure of RNA substrates and reaction environments, and the effects are different among different T4 RNA ligases [8–13]. Particularly, the intermolecular co-fold structures of the adapters and sRNA have a strong influence on the ligation efficiency of T4 RNA ligases [9–11, 13]. Thus, different sRNA sequences are ligated with certain adapters at different efficiencies. This is the major contributor to the strong bias for or against certain sRNAs observed in all sRNA libraries regardless of sequencing platforms and to the discrepancies observed when different library construction methods are used for the same RNA samples [9–11, 14, 15]. In many cases, some sRNAs can be up to a thousand times overrepresented or underrepresented in NGS data. In order to reduce the ligation bias, mixed adapter pools and improved reaction conditions have been used for library construction, where mixed adapters increase the chance of diverse sRNAs and adapters to form the preferred co-fold structure for RNA ligases [9–12, 16].

Sorefan et al. developed High Definition (HD) adapters for the Illumina platform [11]. HD adapters contain four degenerate nucleotides at the ligation ends of the commercial Illumina adapters (Fig. 1). However, when the currently available commercial kits for sRNA library construction for Illumina sequencing platform were used along with HD adapters, a large amount of 5' adapter–3' adapter ligation products was produced. In order to overcome this problem, some published protocols included a purification step for the sRNA ligation products through gel-extraction, which was time-consuming and often resulted in a substantial loss of the ligated products. In 2011, Epicentre Biotechnologies introduced a degradase, RecJ exonuclease to remove the excessive 3' adapter used

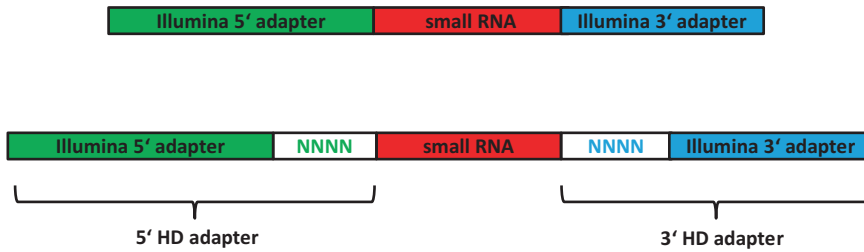


Fig. 1 Small RNAs ligated to the Illumina adapters and HD adapters. The HD adapters include four degenerate nucleotides at the ligation ends of the commercial Illumina 5' adapters (green box) and 3' (blue box) adapters

in sRNA library construction, which largely reduced 5' adapter–3' adapter ligating products [17, 18]. Based on this method, we developed a protocol for sRNA library construction using HD adapters [19, 20]. In this protocol, the reaction conditions for sRNA-adapter ligation and the removal of excessive adapters are optimized. High-quality sRNA libraries can be constructed from 1 µg of total RNA and 300 ng of sRNA.

2 Materials

2.1 RNA Isolation

The total RNA extraction is based on the method published by Chomczynski and Sacchi [21], where a monophasic solution of phenol and guanidine isothiocyanate was used in the extraction buffer. All the solutions should be prepared with nuclease-free water.

1. TRI Reagent® solution

The TRI Reagent® solution should be stored at 4 °C in dark containers. Phenol is corrosive, and guanidine isothiocyanate is an irritant. Gloves and personal protection should be worn while working with TRI Reagent®.

2. Chloroform.

3. Isopropanol.

4. Ethanol.

5. 80% ethanol.

6. Nuclease-free water.

7. 3 M sodium acetate, pH 5.0.

8. 5 mg/ml glycogen.

9. mirVana miRNA isolation kit™.

2.2 Adenylation of 3' HD Adapter

1. 3' HD adapter oligo

The oligo sequence is 5'-NNNNTGGAATTCTCGGGTGCCAAGG(2'3'ddC)-3'. Its 3' end is modified with dideoxycytidine (ddC) which prevents 3' extension by DNA polymerases.

Store the adapter in 100 μ M and 10 μ M stock solutions at -20°C .

2. T4 polynucleotide kinase (10 U/ μ L) and 10 $\times$ buffer.
3. 10 mM ATP.
4. 5' DNA adenylation kit.
5. 15% urea denaturing polyacrylamide gel
To prepare 15 mL polyacrylamide gel, weigh 6.3 g urea. Dissolve the urea in 3.5 mL nuclease-free water and 1.5 mL 5 $\times$ TBE in a 50 mL falcon tube at 37–42 $^{\circ}\text{C}$, followed by mixing it with 5.5 mL 40% acrylamide/bis solution (19:1), 7.5 μ L N,N,N',N'-tetramethyl-ethylenediamine (TEMED) and 150 μ L 10% ammonium persulfate (APS). Adjust the final volume to 15 mL with nuclease-free water. Precast mini-gels for nucleic acid can also be used.
6. 2 $\times$ loading dye: 5 mM EDTA, 0.1% bromophenol blue, 0.1% xylene cyanol, and 95% formamide.
7. SYBR[®] Gold nucleic acid stain.

2.3 3' HD Adapter Ligation

1. 50% PEG 8000 solution.
2. RNaseOut (40 U/ μ L).
3. T4Rnl2, truncated K227Q (200 U/ μ L) and 10 $\times$ buffer.
4. RNA clean and concentrator kit[™].

2.4 Removal of Excessive 3' HD Adapter

1. 5' deadenylase (10 U/ μ L) and 10 $\times$ buffer.
2. RNaseOut (40 U/ μ L).
3. 100 mM DTT.
4. 25 mM EDTA.
5. 500 mM Tris-HCl, pH 9.0.
6. 50 mM MgCl₂.
7. RecJ exonuclease (10 U/ μ L) or RecJ_f (30 U/ μ L).

2.5 5' HD Adapter Ligation

1. 5' HD adapter
The oligo sequence is 5'-/5AmMC6/GUUCAGAGUUCU ACAGUCCGACGAUCNNNN-3', where its 5' end is modified with amino modifier C6. For our experiments this RNA oligo was synthesized by Integrated DNA technologies (IDT). The oligo should be stored in 100 μ M and 20 μ M stock solutions at -80°C .
2. 10 mM ATP.
3. T4Rnl1 (5 U/ μ L) and 10 $\times$ buffer.
4. 50% PEG 8000.

2.6 cDNA Synthesis

1. High performance MMLV Reverse Transcriptase (200 U/ μ L) and 10 $\times$ reaction buffer.
2. 10 mM dNTPs (dATP, dCTP, dGTP, dTTP).
3. 100 mM DTT.
4. RT primer
The oligo sequence is: 5'-GCCTTGGCACCCGAGA ATTCCA-3'. This oligo should be stored in 100 μ M and 20 μ M stock solutions at -20°C .

**2.7 PCR
Amplification of cDNA**

1. 10 mM dNTPs.
2. Phusion DNA polymerase (2 U/ μ L) and 5 $\times$ high fidelity Phusion buffer.
3. Illumina RP-1 forward primer and reverse index primers.
The oligos can be synthesized through Sigma-Aldrich and IDT. They are stored in 10 μ M stock solutions in -20°C .

**2.8 Gel Extraction
of PCR Products**

1. 8% native polyacrylamide gel
To prepare 15 mL 8% native polyacrylamide gel, mix 10 mL nuclease-free water, 1.5 mL 5 $\times$ TBE, 3 mL 40% acrylamide-bis solution (19:1), 7.5 μ L TEMED, and 150 μ L 10% APS. The final volume of the mixture is adjusted with nuclease-free water to 15 mL.
2. 1 $\times$ NEB restriction enzyme buffer 2.
3. 20 bp DNA ladder.
4. Spin-X column.
5. 5 mg/mL glycoblue.

3 Methods

Most of the functional sRNAs are characterized by 5' phosphate and 3' hydroxyl groups. For sRNA library construction, the sRNAs are directly ligated to oligonucleotide adapters at the 3' and 5' end through T4Rnl2 and T4Rnl1. In order to reduce intramolecular circularization and concatemerization, truncated T4Rnl2 K227Q mutant is used for 3' end ligation [8], where preadenylated adapter is used as one substrate. For more efficient ligation, excessive 3' HD adapter is used. The reaction mixture which contains sRNAs tagged with 3' adapter is cleaned and concentrated. Before 5' end ligation, the remaining 3' HD adapter is degraded by the RecJ exonuclease which is a magnesium-dependent 5' to 3' exonuclease and specific to single-stranded DNA [22]. The sRNAs with the 3' tag are further ligated to the 5' end HD adapter by T4Rnl1. The sRNAs that have been ligated to both adapters are then reverse-transcribed with the RT primer and the cDNA is amplified using high fidelity DNA

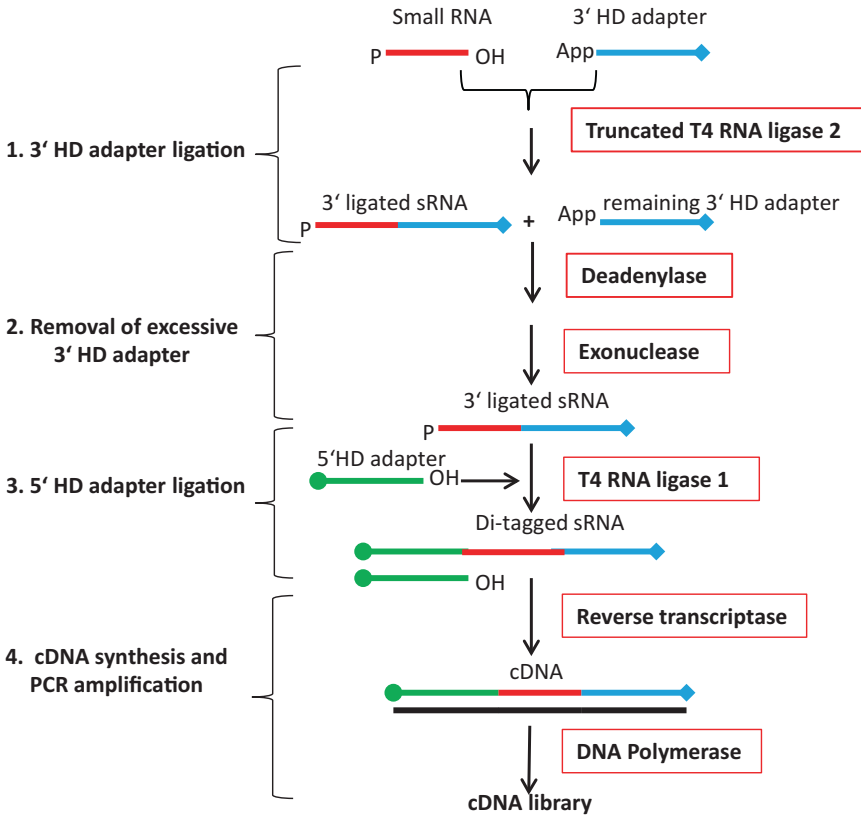


Fig. 2 Overview of sRNA library construction using HD adapters

polymerase (*see Note 1*). The PCR products are separated by size through polyacrylamide gel electrophoresis (PAGE) and the band containing desired cDNA insert is extracted from the gel for sequencing. A schematic overview of the workflow is shown in Fig. 2.

3.1 RNA Isolation

The total RNA or sRNA can be extracted through customized methods or purified through column-based kits. The procedure described here for RNA extraction is based on the manual provided by TRI Reagent® manufacturer. For liquid materials such as body fluid and cell suspensions, TRI Reagent® RT-Liquid solution can be used (*see Note 2*).

1. Homogenization

Tissue samples: Grind 100 mg tissue in liquid nitrogen until a fine powder is obtained. Mix 1 mL TRI Reagent® solution with the powder in a microcentrifuge tube.

Cultured cells or body fluids: Homogenize and lyse the cells directly by adding 1 mL TRI Reagent® solution per $5\text{--}10 \times 10^6$ cells or per 10 cm² culture dish area and pipette up and down.

2. Incubate the mixture for 5 min at room temperature to allow complete dissociation of nucleoprotein complexes.

3. Add 200 μL chloroform, mix vigorously for 15 s and incubate at room temperature for 5 min.
4. Centrifuge at $16,000 \times g$ at 4°C for 15 min and transfer the aqueous phase to a fresh centrifuge tube. Remember the volume of transferred solution.
5. Add the same volume of isopropanol into the aqueous RNA extract, mix well and incubate at room temperature for 10 min or on ice or -80°C for 30 min. The co-precipitation time and temperature may vary with different samples.
6. Centrifuge the above mixture at $10,000 \times g$ at 4°C for 10 min and remove the supernatant. The centrifugation time and speed can be adjusted based on tissues.
7. Wash the RNA pellets with 1 mL 80% ethanol, vortex and centrifuge at $7500 \times g$ for 5 min at 4°C . Repeat once.
8. Air-dry the RNA pellet for 5–10 min, resuspend the RNA in nuclease-free water, and store at -80°C .
9. Quantify and measure the RNA concentration using a NanoDrop.

3.2 sRNA Purification

We recommend to further purify the total RNA or small RNA fraction before sRNA library construction using the mirVana miRNA isolation kit (Ambion) following the manufacturer's protocol. If needed, the sRNA fraction can be concentrated through ethanol precipitation. Make a note of the volume of the RNA eluate, mix the RNA elute with 3 volumes of absolute ethanol, 0.1 volume of sodium acetate (pH 5.0) and 1 μL of glycogen (5 mg/mL), and store it at -20°C or -80°C overnight. The next day, centrifuge the samples at $20,000 \times g$ for 20 min at 4°C , wash the RNA pellet twice with 80% ethanol, air-dry the RNA for 5 min and resuspend it in proper volume of nuclease-free water. Determine the RNA concentration using a NanoDrop.

3.3 Adenylation of 3' HD Adapter

The 3' HD adapters are first phosphorylated with T4 polynucleotide kinase, and then adenylated using Mth RNA ligase from NEB.

1. Mix well the following reagents in a microcentrifuge tube and incubate the tube at 37°C for 30 min.

Components for the reaction mix	Volume (μL)
100 μM 3' HD adapter	6
10 $\times$ T4 polynucleotide kinase buffer	10
10 mM ATP	10
Nuclease-free water	72
T4 polynucleotide kinase (10 U/ μL)	2

2. Mix the above phosphorylated 3' HD adapter with 10 μL of 3 M NaOAc (pH 5.0), 250 μL of absolute ethanol, and 2 μL of glycogen (5 mg/mL). Store the mixture at $-20\text{ }^{\circ}\text{C}$ or $-80\text{ }^{\circ}\text{C}$ overnight.
3. Centrifuge the above precipitate at $20,000 \times g$ for 10 min at $4\text{ }^{\circ}\text{C}$ and discard the supernatant.
4. Wash the pellet with 250 μL 80% ethanol.
5. Air-dry for 5–10 min and resuspend in 12 μL nuclease-free water.
6. Follow the manual for 5' DNA adenylation kit to adenylate the phosphorylated 3' HD adapter. The reaction mix includes the following components:

Components for the reaction mix	Volume (μL)
Phosphorylated 3' HD adapter	4.5
10 $\times$ 5' DNA Adenylation reaction buffer	4
1 mM ATP	4
Mth RNA ligase	4
Nuclease-free water	23.5

7. Incubate the reaction mix at $65\text{ }^{\circ}\text{C}$ for 1 h.
8. Incubate the reaction mix at $85\text{ }^{\circ}\text{C}$ for 5 min to inactivate the enzyme, and chill it on ice.
9. Add 160 μL nuclease-free water to the inactivated reaction followed by adding 200 μL phenol–chloroform. Mix them well.
10. Centrifuge the mixture at tube at top speed for 10 min. Transfer the aqueous phase into a fresh tube and make a note of its volume.
11. Add 2.5 volume of absolute ethanol, 0.1 volume of 3 M NaOAc, and 1 μL of glycogen (5 mg/mL).
12. Centrifuge at $20,000 \times g$ at $4\text{ }^{\circ}\text{C}$ for 15 min and discard the supernatant.
13. Wash the pellet twice with 80% ethanol.
14. Air-dry the pellet for 5–10 min and resuspend it in 12 μL of nuclease-free water.
15. Determine the concentration of the adenylated adapter by NanoDrop and adjust it to a final concentration of 10 μM .
16. Observe the adenylated adapter and non-adenylated adapter in the same 15% denaturing gel. For our experiments the gel electrophoresis was performed using the mini PROTEAN III system (Biorad) and the gel was stained with SYBR[®] Gold.

3.4 Small RNA Library Construction Using HD Adapters

3.4.1 3' HD Adapter Ligation

1. Thaw 50% PEG in a 37 °C heat block and keep the PEG at room temperature.
2. Set up heat blocks at 26 °C and 70 °C.
3. Mix 11.25 μL of the RNA sample with 1 μL of 10 μM preadenylated 3' HD adapter.
4. Denature the RNA and adapter mixture at 70 °C for 2 min and chill it on ice.
5. Add the following ligation reaction mix to the denatured RNA and adapter mixture. Please note that 50% PEG solution is viscous and be sure to pipette the reaction mixture up and down several times.

Components for the reaction mix	Volume (μL)
10 $\times$ T4 Rnl2 buffer	2
RNaseOUT (40 U/ μL)	0.75
Truncated T4 Rnl2 K227Q (200 U/ μL)	1
Nuclease-free water	3.25
50% PEG 8000	4

6. Incubate the reaction mix at 26 °C for 2–3 h.
7. Clean up the 3' end ligation reaction using the RNA clean and concentrator kit (Zymo Research) following the protocol provided by the manufacturer. Elute the RNA in 12.1 μL nuclease-free water and keep it on ice.

3.4.2 Removal of Excessive 3' HD Adapter

1. Deadenylate the remaining 3' HD adapter using deadenylase. Pipette the following reagents to the above cleaned sRNA-adapter ligating product.

Components for the reaction mix	Volume (μL)
10 $\times$ deadenylase buffer	1.6
100 mM DTT	0.8
RNaseOUT (40 U/ μL)	0.5
5' deadenylase (10 U/ μL)	1.0

2. Incubate the above reaction mix at 30 °C for 30 min.
3. Stop the reaction by adding of 4 μL of 25 mM EDTA and place the tube on ice immediately.
4. Add 2 μL of 0.5 M Tris-HCl (pH 9.0), 7 μL of 50 mM MgCl_2 , and 1 μL of RecJ Exonuclease or 1 μL of RecJ_f Exonuclease to the above tube. Mix them well.
5. Incubate the exonuclease reaction at 37 °C for 30 min.

**3.4.3 5' HD Adapter
Ligation**

1. Denature 1 μL of 20 μM 5' HD adapter at 70 °C for 2 min and chill it on ice.
2. Add it into the tube for the exonuclease reaction mix from Subheading 3.4.2.
3. Add the following reagents into the above tube sequentially, and pipette the reaction mix up and down.

Components for the reaction mix	Volume (μL)
10 $\times$ T4 Rnl1 buffer	1
10 mM ATP	1
5' adapter (20 μM)	1
T4 Rnl1	1
50% PEG 8000	7

4. Incubate the 5' ligation reaction at 26 °C for 2–3 h.
5. Clean up the 5' ligation reaction using the RNA clean and concentrator kit following the manual provided by the manufacturer.
6. Elute the ligation product with nuclease-free water in a total volume of 30 μL and place the reaction on ice.

**3.4.4 cDNA Synthesis
and Amplification**

1. Set up the following cDNA synthesis reaction and mix it well. Incubate the reaction at 37 °C for 20 min.

Components for the RT reaction mix	Volume (μL)
sRNAs ligated to both 5' and 3' adapter (from Subheading 3.4.3)	30
10 $\times$ MMLV Reverse transcription buffer	4
10 mM dNTP	2
100 mM DTT	2
20 μM RT Primer	1
High performance MMLV reverse transcriptase	1

2. Inactivate the reaction at 85 °C for 15 min and keep it on ice. The cDNA can also be stored at –20 °C or –80 °C.

3. Set up the following PCR reaction mix for cDNA amplification

Components for the PCR reaction mix	Volume (μL)
cDNA	4
Nuclease-free water	9.3
10 mM dNTPs	0.5
5 $\times$ high fidelity Phusion buffer	4
Illumina RP-1 primer (10 μM)	1
Illumina index primer (10 μM)	1
Phusion DNA polymerase (2 U/ μL)	0.2

4. Run the PCR reaction at following conditions. The PCR products can be stored at $-20\text{ }^{\circ}\text{C}$ (*see* **Note 3**).

1. 98 $^{\circ}\text{C}$ initial denaturation	30 s	
2. 98 $^{\circ}\text{C}$ denaturation	10 s	} 10–15 cycles
3. 55 $^{\circ}\text{C}$ annealing	30 s	
4. 72 $^{\circ}\text{C}$ elongation	15 s	
5. 72 $^{\circ}\text{C}$ final extension	10 min	
6. 4 $^{\circ}\text{C}$ hold	Forever	

3.4.5 Gel Extraction of PCR Products

1. Separate the PCR products on a native 8% PAGE gel (*see* Subheading 2.8) and load a 20 bp DNA ladder for size determination. We used the mini PROTEAN III system for the electrophoresis.
2. Run the gel at 120 V for 1.5–2 h in 0.5 $\times$ TBE.
3. Stain the gel with SYBR[®] Gold and visualize on a scanner or UV box.
4. Cut out the DNA bands around 145–150 bp (which contain the cDNA library of sRNAs) with sterile razor blades.
5. Place the gel slice into a gel breaker tube and centrifuge at $20,000 \times g$ for 5 min.
6. Elute the cDNA library in 1 $\times$ NEB restriction enzyme buffer 2 by shaking the gel slurry at room temperature for more than 2 h or overnight at 4 $^{\circ}\text{C}$.
7. Add the gel slurry into a spin-X column and centrifuge the column at the speed of $650 \times g$ for 5 min at room temperature to remove the gel debris.

8. Concentrate the eluted cDNA library with ethanol precipitation overnight at -80°C . For example, when there is $300\ \mu\text{L}$, mix it with $2\ \mu\text{L}$ glycoblue ($5\ \text{mg/mL}$), $30\ \mu\text{L}$ $3\ \text{M}$ NaOAc, and $975\ \mu\text{L}$ absolute ethanol. Leave the mixture at -20°C or -80°C overnight.
9. Centrifuge at $20,000 \times g$ for 20 min at 4°C .
10. Wash the pellet with 80% ethanol.
11. Air dry for 5–10 min and resuspend in $12\ \mu\text{L}$ nuclease-free water.
12. Measure the concentration of the cDNA library by NanoDrop and quantify the library.
13. Send the cDNA library for sequencing.

4 Notes

1. The sRNA library construction using HD adapters can be completed within one day, while the PCR product is gel extracted on the second day. The protocol can be modified as needed for example, extending ligation time.
2. It is important to have high quality total RNA, but the quality of total RNA often varies with the original tissues. Thus, any efforts in improving total RNA purity and reducing other non-RNA contaminants will be helpful for successful library construction.
3. We recommend to test a few PCR amplification cycles to find out the lowest cycle number resulting in clear bands for desired PCR products and least side products if needed. When PCR products from direct 5' and 3' adapter ligation are high, a second gel extraction of the purified library will help.

References

1. Axtell MJ (2013) Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol* 64:137–159
2. Farazi TA, Juranek SA, Tuschl T (2008) The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* 135(7):1201–1214
3. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116(2):281–297
4. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell* 136(4):642–655
5. Pritchard CC, Cheng HH, Tewari M (2012) MicroRNA profiling: approaches and considerations. *Nat Rev Genet* 13(5):358–369
6. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* 44(1):3–12
7. Ho CK, Wang LK, Lima CD, Shuman S (2004) Structure and mechanism of RNA ligase. *Structure* 12(2):327–339
8. Viollet S, Fuchs RT, Munafo DB, Zhuang F, Robb GB (2011) T4 RNA ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotechnol* 11:72
9. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, Ojo T, Luo S, Schroth G, Tuschl T (2011) RNA-ligase-dependent biases in

- miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* 17(9):1697–1712
10. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* 39(21):e141
 11. Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* 3(1):4. doi: [10.1186/1758-907X-3-4](https://doi.org/10.1186/1758-907X-3-4)
 12. Zhang Z, Lee JE, Riemondy K, Anderson EM, Yi R (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol* 14(10):R109
 13. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB (2012) Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res* 40(7):e54
 14. Baran-Gale J, Erdos MR, Sison C, Young A, Fannin EE, Chines PS, Sethupathy P (2015) Addressing bias in small RNA library preparation for sequencing: a new protocol recovers microRNAs that evade capture by current methods. *Front Genet* 6:352 doi: [10.3389/fgene.2015.00352](https://doi.org/10.3389/fgene.2015.00352)
 15. Tian G, Yin X, Luo H, Xu X, Bolund L, Zhang X, Gan SQ, Li N (2010) Sequencing bias: comparison of different protocols of microRNA library construction. *BMC Biotechnol* 10:64
 16. Sun G, Wu X, Wang J, Li H, Li X, Gao H, Rossi J, Yen Y (2011) A bias-reducing strategy in profiling small RNAs using Solexa. *RNA* 17(12):2256–2262
 17. Pease J (2011) Small-RNA sequencing libraries with greatly reduced adaptor-dimer background. *Nat Methods* 8(3) Application Note, iii-iv
 18. Vaidyanathan R, Kuersten S, Doyle K (2013) US Patent 20110104785 A1, 2011
 19. Xu P, Billmeier M, Mohorianu I, Green D, Fraser WD, Dalmay T (2015) An improved protocol for small RNA library construction using High Definition adapters. *Methods Next Generation Seq* 2: 1–10
 20. Xu P, Mohorianu I, Yang L, Zhao H, Gao Z, Dalmay T (2014) Small RNA profile in moso bamboo root and leaf obtained by high definition adapters. *PLoS One* 9(7):e103590
 21. Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* 162(1):156–159
 22. Lovett ST, Kolodner RD (1989) Identification and purification of a single-stranded-DNA-specific exonuclease encoded by the *recJ* gene of *Escherichia coli*. *Proc Natl Acad Sci U S A* 86(8):2627–2631

Surface Acoustic Wave Lysis and Ion-Exchange Membrane Quantification of Exosomal MicroRNA

Katherine E. Richards, David B. Go, and Reginald Hill

Abstract

MicroRNA detection and quantification are commonly explored techniques for diagnostic and prognostic predictions. Typically, microRNAs are extracted and purified from a biological source, converted into complementary DNA (cDNA), and amplified using real time polymerase chain reaction (RT-PCR). The number of RT-PCR cycles required to reach the threshold of detection provides a relative quantification of the target microRNA when this data is normalized to the quantity of a control microRNA. This methodology has several drawbacks, including the need to artificially amplify the target microRNA for detection as well as quantification errors that can occur due to expression level differences of the control microRNAs for normalization in various sample sources. Here, we provide a technique to quantify actual concentrations of target microRNAs directly from any biological source without the requirement of these additional steps. In addition, we describe an alternative approach for obtaining exosomal microRNAs directly from biological samples without the use of harsh detergents and RNA isolation.

Key words MicroRNA, RNA, Detection, Quantification, Exosomes, SAW, Microfluidics, Ion-exchange, Membrane, Biosensor

1 Introduction

One of the many challenges facing clinicians and scientists today is the lack of early detection and prognostic biomarkers for many diseases. MicroRNAs isolated from patient serum samples have the potential to be promising biomarkers because they can be obtained noninvasively and are stable in circulation. However, the process of detecting them can be lengthy and can require larger sample volumes than is sometimes available from routine procedures. In order to detect microRNAs using RT-PCR, the RNA must be isolated from a sample, converted into cDNA, and a specific cDNA is amplified by DNA polymerase, which increases the abundance of the cDNA to reach the minimum threshold of detection. When normalized to a control microRNA, the cycles needed to reach this limit of

detection are used to quantify relative levels of target microRNA in the original sample. Ideally, direct quantification of the target microRNA would be done from a patient sample since the additional RNA isolation and RT-PCR steps leave room for human error and sample loss. We developed a way to quantify small concentrations of microRNA (down to 2 pM) without the need for these steps [1, 2, 10]. A microRNA probe is attached to a positively charged ion-exchange membrane where it can bind to its complimentary microRNA. When the proper current is applied, anions flow through the pores of the membrane, but this flow is disrupted by higher amounts of bound microRNAs, creating a measurable shift in voltage. The ion-exchange membrane can be reused by dehybridizing the probes on the membrane from their complementary miRNA before adding a new sample [2, 3]. This allows us to incrementally add higher known concentrations of a commercially available microRNA to the same membrane to create a calibration curve linking the amount of voltage shift to target microRNA concentration. Thus, the shift in voltage is used to directly detect and quantify the amount of target microRNA in a sample without the need for reverse transcription or amplification. If normalization to a control microRNA is needed, the steps below can be ran side-by-side using two different biosensors/membranes, one with the control probe and one with the target-microRNA probe (one probe is used per membrane). Most microRNAs in circulation are contained in lipid vesicles known as exosomes [4]. Exosomes can be lysed by detergents to release the RNA within, but we found these detergents to be disruptive to both the probe-sensor and the RNA-probe binding. Instead, microRNAs in exosomes can be released by exposing exosome-containing samples to surface acoustic waves, which results in about a 40% exosome lysis rate (which can be optimized by the user by changing variables such as power used) [1]. Surface acoustic waves (SAWs) are Rayleigh waves that can be generated on the surface of a piezoelectric substrate to induce fluid flow in microfluidic channels. When an electric current is applied to a transducer consisting of interdigitated electrodes on the surface of a piezoelectric substrate, electromechanical couple generates a SAW that travels along the surface of the substrate. Although the amplitude of the displacement is small (~nanometers), the energy contained in the wave is high, and when the wave encounters a fluid film or drop on the substrate, it will refract and induce convective motion in the fluid [5, 6]. Here we use a standard electrode-width controlled (EWC) single-phase unidirectional SAW transducer (SPUDT) design to generate SAWs with a wavelength of 132 μm [1, 7–10]. When conditioned cell media flows through a microchannel on the surface of the SAW device, the induced acoustic and electric forces lyse the exosomes, releasing their RNA into the media. The media can then be directly channeled onto the probe-functionalized ion-exchange membrane,

after creation of a calibration curve, to determine the concentration of target microRNA in the cell media sample. Conditions may also be optimized to use patient serum samples as the biological source of microRNA-containing exosomes. This technique required little volume (roughly 100 μ L media) and did not require the RNA isolation, RNA to cDNA conversion, and PCR amplification steps required to quantify microRNA with RT-PCR [1].

2 Materials

2.1 Biosensor Materials

1. Benzophenone-3,3',4,4'-tetracarboxylic acid powder.
2. Sodium hydroxide.
3. pH strips.
4. AMH5E-HD RALEX[®] ion-exchange membrane.
5. Electro-Cure 500 Light.
6. Loctite[®] 3492 UV curing glue.
7. (Poly)vinyl alcohol.
8. TAP Quik-Cast Polyurethane Resin.
9. Acrylic sheet (Poly(methyl methacrylate) or PMMA).
10. DREMEL[®] 4000 drill with 2 mm drill bit.
11. Tygon tubing (inside diameter 2 mm).
12. Gamry[™] potentiostat.
13. MES buffer: 0.1 M MES, pH 5.5.
14. 0.1 $\times$ PBS buffer: 13.7 mM NaCl, 0.27 mM KCl, 1 mM Na₂HPO₄, 0.18 mM KH₂PO₄, pH 7.4.
15. EDC powder (N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride).
16. ssDNA probe complementary to RNA target.
17. Artificial target RNA to create a calibration curve.

2.2 SAW-Incorporated Microfluidics Device Materials

1. Oerlinkon Evaporation System.
2. Lithium niobate piezoelectric wafer (Precision Micro-Optics).
3. L-Edit software.
4. Photoresist spinner.
5. S1813 photoresist.
6. Hot plate.
7. Karl Suss MJB3 Mask Aligner.
8. DRYTEK plasma etcher.
9. Acetone.

10. 300 μm thick polycarbonate sheets.
11. Graphtec Cutting Pro.
12. Glass slides.
13. Agarose powder.
14. ACRIFIX® 1R 0192 UV curing glue.
15. Centrifuge and ultracentrifuge.
16. Agilent 33250A generator.
17. E&I 325LA RF power amplifier.
18. Microcentrifuge tubes.
19. Exosome-free FBS.
20. 200 nm pore size filters.

3 Methods

3.1 RNA Probe Biosensor Fabrication

3.1.1 Membrane Carboxylation

1. Add 3 mg of benzophenone-3,3',4,4'-tetracarboxylic acid powder, to 30 μL of DI water. Add NaOH to reach pH between 6 and 8. Keep this mixture in dark or in aluminum foil (*see Note 1*).
2. Cut the ion-exchange membrane into a piece just large enough to cover a 0.25 mm^2 opening in the biosensor. With a permanent marker, label one side of the membrane as the back side.
3. Soak membrane piece in benzophenone solution for 30 min.
4. Expose the front side of the membrane to UV light from Electro-Cure 500 for 10 min.
5. With forceps, carefully transfer the membrane into DI water to rinse, and then allow to dry.
6. Add benzophenone solution to the front of the membrane piece, covering the surface. Incubate with benzophenone solution in the dark for 15 min.
7. Repeat **steps 4–6** until all of the benzophenone solution has been consumed, and keep membrane piece in DI water until continuing with the membrane biosensor fabrication steps (below).

3.1.2 Membrane Biosensor Fabrication

1. Allow carboxylated membrane to dry then put small drop of polyvinyl alcohol (PVA) in the central post in the bottom piece of the primary mold (*see Note 2*). Put the back surface of the membrane on the polyvinyl alcohol and secure the membrane in place with the top piece of the primary mold.
2. Mix polyurethane resin (casting resin A and B 1:1 ratio) for 1 min, load into syringe, and pour into a hole in the top of the primary mold (*see Note 3*). Allow to solidify for 30 min to 1 h.

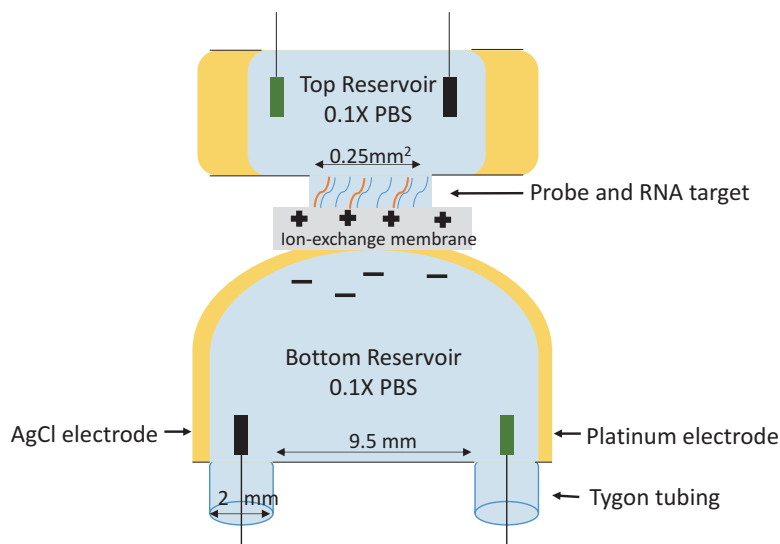


Fig. 1 Biosensor assembly. The ion-exchange membrane is sandwiched between two polyurethane molds with the functionalized probe on top. Both chambers are filled with 0.1× PBS, and the electrodes of the Gamry™ Potentiostat are connected to the inside of the *top chamber* and the inside of the *bottom chamber* through the tygon tubing

3. Obtain a piece of acrylic sheet. Drill two holes (that electrodes will later go into) using 2 mm drill bit. The distance between the two holes should be such that they fit just inside the bottom reservoir of the biosensor (*see* Fig. 1).
4. Cut two pieces of tygon tubing to ~4 mm in length. Glue these to the acrylic sheet using UV curing glue so the inner diameter of tubing is aligned with the holes (*see* **Note 4**). UV treat with Electro-Cure 500 for 10 min to cure.
5. When the polyurethane resin from **step 2** has solidified, detach from mold. This is the biosensor. Attach the acrylic sheet to the biosensor with UV curing glue. Treat with UV light for 10 min with ElectroCure 500. Let cool at room temperature for 10 min (*see* **Note 5**; *see* Fig. 1).
6. Fill the top and bottom reservoir of the biosensor with 0.1× PBS. Let soak overnight.
7. User can take raw current–voltage characteristic measurements at this step using the Gamry™ potentiostat (as a reference prior to probe functionalization and RNA sample exposure; *see* Subheading 3 below).

3.1.3 Gamry™ 500 Potentiostat Current–Voltage Characteristic (CVC) Measurements

1. Wash the top and bottom reservoirs of the biosensor with 0.1× PBS then fill the top and bottom reservoirs with 0.1× PBS while making sure not to accumulate air bubbles (bottom reservoir is filled with 0.1× PBS through tygon tubes).

2. Insert platinum electrode and silver chloride reference electrode of Gamry™ potentiostat into the tygon tubes of the bottom reservoir (*see* Fig. 1).
3. Insert platinum electrode and silver chloride reference electrode into top reservoir, making sure that the two electrodes do not touch each other (*see* Note 6).
4. Start up the Gamry™ Framework software application. Under the “Experiments” menu, select “D – Electrochemical Energy,” then “6 Polarization Curve” from the submenu.
5. When pop-up options occur, make sure “Working Connection” is set to negative, the scan rate is set to “ 5×10^{-7} A/s,” and the sample period to “0.5 s.”
6. The Gamry™ will begin to take the current–voltage characteristic (CVC) curve. Take about three to five measurements until each CVC curve has stabilized and looks approximately the same (*see* Note 7). The user should take measurements before and after probe functionalization. Once the probe is functionalized the user can then add biological sample to the top reservoir prior to CVC measurements.

3.1.4 Probe Functionalization and Target RNA Detection

1. Fill top reservoir of biosensor with pH 12 sodium hydroxide for at least 4 h (overnight permissible).
2. Remove sodium hydroxide and wash top reservoir three times with 0.1 M MES buffer, pH 5.5. Fill top reservoir with MES buffer and allow membrane to soak for 30 min.
3. Bring EDC powder to room temperature. Combine 19 mg EDC powder with 250 μ L of MES buffer, and mix thoroughly. EDC powder should fully dissolve into the solution.
4. Aspirate MES buffer from top reservoir, and add EDC buffer solution. Let the membrane soak in EDC solution for 30 min.
5. Mix probe with 0.1 $\times$ PBS to obtain a probe concentration of at least 10 μ M. Aspirate EDC buffer from top reservoir. Place 40 μ L of probe-containing buffer in top reservoir, on top of membrane, and allow overnight incubation (*see* Note 8).
6. After overnight incubation, current–voltage characteristic curves may be taken using the Gamry™ Potentiostat to determine if probe was successfully functionalized (*see* Note 9; *see* Fig. 2).
7. Once probe is functionalized onto ion transport membrane, wash top reservoir 4 $\times$ with 0.1 $\times$ PBS, then add the target RNA (complementary to functionalized probe) sample directly to the membrane. Wash top reservoir/membrane again with 4 $\times$ PBS to wash away unbound RNA and other biological debris. We recommend starting off with commercially available, known concentrations of your target miRNA of interest before

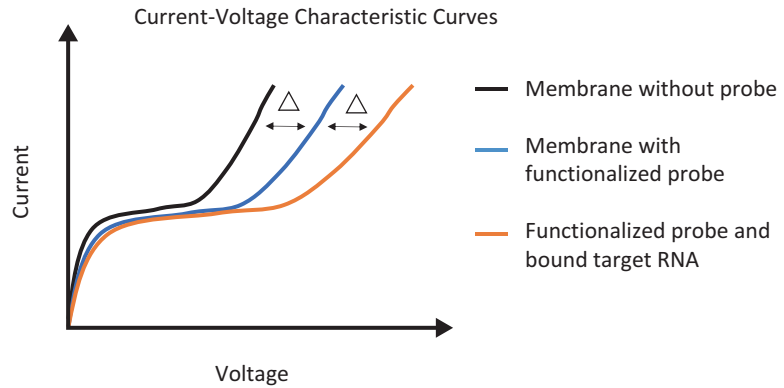


Fig. 2 Current–voltage characteristic curves. The addition of probe and hybridized RNA to the probes creates an increase in the voltage shift, which is measured for incremental known concentrations of target miRNA to establish a calibration curve prior to the addition of the biological sample

adding your biological sample (conditioned cell media sample with exosomes exposed to SAW lysis in our case). This will allow you to obtain voltage data points at known concentrations of target RNA to create a calibration curve using the same membrane.

8. Take CVC measurements as described above. We recommend adding incrementally higher concentrations of RNA target on the same membrane and taking CVC measurements post each concentration to establish successful binding of target RNA to probe (*see Note 10*).
9. In order to quantify target RNA concentration of multiple RNA samples using the same membrane piece, the membrane must be washed with 0.1× PBS, pH 13 (buffered with NaOH) to dehybridize the target RNA from the probes before adding the next RNA sample to the membrane. Wash membrane with 0.1× PBS to reestablish pH. Take CVC measurements to affirm that the voltage shift is back to that of membrane and probe alone. Now another RNA sample may be added to the membrane. Repeat **steps 7–9** as necessary to measure target RNA concentrations of separate biological samples using the same membrane.

3.2 SAW Microfluidics Platform Fabrication

3.2.1 Surface Acoustic Wave (SAW) Device Fabrication

1. Use Oerlikon Evaporator to deposit 20 nm of titanium followed by 200 nm aluminum onto a 128° Y-cut lithium niobate piezoelectric wafer (101.6 mm × 101.6 mm).
2. Dice wafer into six to eight pieces large enough to accompany each individual SAW device (each SAW device is 16 mm × 40 mm).
3. Make a SAW mask using the L-Edit software to design an interdigitated transducer consisting of 20 pairs of interdigitated electrodes. Each electrode leg should be 1/8 of

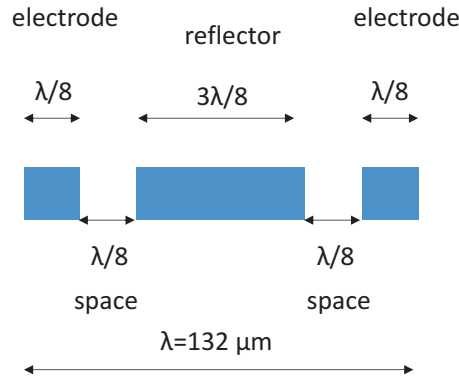


Fig. 3 Mask design. An example of one electrode pair is shown as designed in L-Edit for the etching of the interdigitated electrodes of the SAW device. The device uses the standard SPUDT design

the designed SAW wavelength (λ) and the spacing between electrodes should also be $1/8$ of the designed SAW wavelength. In our case, the SAW wavelength was $\lambda = 132 \mu\text{m}$ (see Fig. 3, see **Note 11**).

4. Spin coat the lithium niobate wafer with S1813 photoresist at 2500 RPM for 30 s.
5. Place wafer on 90 °C hot plate for 1 min; then place wafer on plastic container and allow to cool for 5 min.
6. Expose in Karl Suss MJB3 Mask Aligner using SAW mask for 14 s, then develop for 1 min or until the SAW pattern is clearly visible.
7. Place lithium niobate piece in DRYTEK plasma etcher, and apply the descum process for 1 min.
8. Use the reactive ion etch (RIE) process to etch away the areas without photoresist in order to expose the titanium–aluminum interdigitated electrodes. An etch time of 5 min and 15 s is usually sufficient. This will need to be optimized by checking under a microscope to validate that the pattern appears correctly. Then wipe the substrate with acetone to remove any residual photoresist.

3.2.2 Microfluidic Device Fabrication

1. Obtain a 300 μm thick polycarbonate sheet and attach it to Graphtec Cutting Pro to cut a channel into the sheet with height of 300 μm and width of 2000 μm . With another polycarbonate sheet cut the holes for the inlet and outlet ports of the microfluidic device (see Fig. 4).
2. Clamp two of the three polycarbonate sheets one on top of the other while being sandwiched between glass slides to prevent scratching with binder clips. Place in 180 °C oven for 15 min.

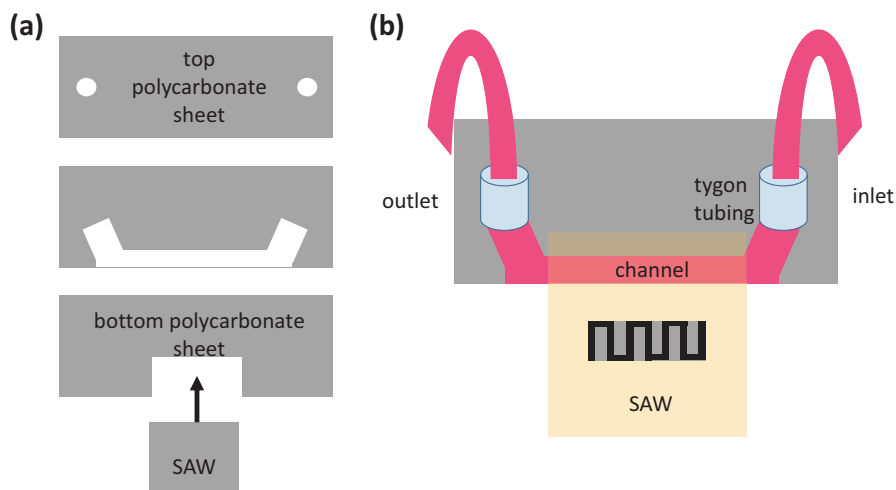


Fig. 4 Integrated microfluidics device design. (a) The three polycarbonate sheets are welded together to create an inlet and outlet hole in the *top* where the biological sample enters and exits, the channel in the *middle* which will come into contact with the SAW device, and the *bottom* polycarbonate sheet which is cut to form to the attached SAW device. (b) Depiction of what the integrated device looks like after addition of the cell media sample and tygon tubing

3. Repeat this process to add the third polycarbonate sheet to the two adhered sheets. The user should now have one polycarbonate sheet (consisting of three combined sheets) which has inlet and outlet holes in the top, which connect to the beginning and end of the interior channel.
4. To attach the SAW device to the microfluidics polycarbonate device beneath the channel, first seal the bottom of the channel (*see Note 12*).
5. Heat up 12% agarose powder in DI water in microwave until it is in liquid form and add it to the channel through the outlet or inlet using a syringe. Allow it to cool and solidify.
6. Once the agarose gel is solidified, remove the sealant on the bottom of the channel and attach the SAW device just beneath the bottom of the channel with UV curing glue (Loctite® 3492). Cure in Electro-Cure 500 for 10 min.
7. Heat the entire device in 80 °C oven for 10 min to liquefy the agarose gel. Aspirate out the liquid agarose from the channel.
8. Tightly seal the channel to the SAW device with ACRIFIX® 1R 0192 UV curing glue to prevent leakage of biological sample from the channel. Cure in Electro-Cure 500 for 20 min.
9. Obtain 4 mm long tygon tube pieces that have the same inner diameter as the inlet and outlet holes and seal them over top of the inlet and outlet holes using the Loctite® 3492 UV curing glue and cure for 5 min.

3.2.3 Exosome Lysis via Surface Acoustic Wave Exposure

10. Finally, tightly seal the tube pieces to the microfluidics device by applying the outside circumference of each tube with ACRIFIX® 1R 0192 UV curing glue and curing for 20 min.
1. To detect a specific miRNA or RNA contained in exosomes collected from cell culture media we suggest culturing cells in exosome-free media to start.
 2. Collect the media from the cells. Wash the cells with 1× PBS and add it to the collected cell media.
 3. Spin down any cells and large cell debris. We typically spin down the media at $500 \times g$ for 5 min, collect the supernatant, spin down the supernatant at $16,500 \times g$ in an ultracentrifuge for 20 min, and collect the supernatant. The supernatant at this point can also be filtered using a 200 nm pore size filter, but this step may result in some exosome sample loss.
 4. This cell media supernatant can now be used and be pumped through the inlet of the microfluidics-SAW device by surface acoustic waves activated by Agilent 33250A generator along with E&I 325LA RF power amplifier (*see Note 13*).
 5. Collect the media exposed to the SAWs from the outlet, allowing the media to drip into a microcentrifuge tube. This media is now ready to be placed onto the probe-functionalized ion-exchange membrane of the biosensor in order to detect the target RNA within (*see Note 14*).

4 Notes

1. pH strips may be used to verify the pH. It is recommended to start by adding 9 μL of concentrated NaOH solution and then add 1–2 μL as needed to reach the desired pH.
2. The bottom and top reservoirs are created by pouring polyurethane into a primary mold. This primary mold was created using CAD designs and 3D printing [2]. After polyurethane reservoirs have solidified the membrane should be sandwiched between the two reservoirs and exposed by an approximately 0.25 mm² opening between the two reservoirs (*see Fig. 1*).
3. We mix resin in a weigh boat while vigorously stirring and avoiding accumulation of air bubbles.
4. Use a syringe needle to poke a hole through the tubing to make sure you did not just glue the tubes shut before you cure the glue with UV light. Also make sure the electrodes you will use will fit through the tubing. If not, adjust the tubing diameter and acrylic sheet hole diameters accordingly.

5. Make sure the acrylic sheet holes and tubing fit just within the diameter of the bottom reservoir (*see* Fig. 1). If too close together, air bubbles often accumulate in the bottom reservoir upon adding 0.1× PBS that will skew data. If outside the diameter of the bottom reservoir, the electrodes will not come into contact with solution and measuring the current–voltage characteristic will be impossible. We also typically glue the outside of the bottom reservoir where it comes into contact with the acrylic sheet and UV cure again to make sure the biosensor is securely attached to the acrylic sheet.
6. To secure the biosensor in place while adding the electrodes to the top reservoir we usually tape the biosensor to a flat platform. To help ensure that the two electrodes do not touch we hold the reference electrode in place with a ring stand.
7. If the system jumps erratically or if the data points appear in red, it could be due to an air bubble in the bottom reservoir. Replacing the 0.1× PBS in the bottom reservoir to eliminate air bubbles may fix the problem (*see* **Note 5**). Otherwise, the user may repeat **steps 4–6** to take additional measurements.
8. For our purposes we used an amine-coupled ssDNA oligonucleotide probe complementary to microRNA 550 which was purchased from Life Technologies.
9. Successful functionalization of the probe onto the membrane can be determined by measuring the current–voltage characteristic (*see* Fig. 2) or by IR spectroscopy.
10. The more target RNA which binds to the probe, the larger the shift in voltage should be, until the membrane has become saturated, with all probes bound to their targets. If no shift in voltage occurs after addition of more RNA, there may be many dilemmas such as RNA degradation or that the starting concentration was too high and membrane became saturated. A few microliters of conditioned cell culture media should be sufficient (detection down to ~2 pM target RNA). Ensure that probe was successfully functionalized (*see* **Note 9**). Try dehybridizing membrane from target RNA (*see* **Note 10**) then adding a more dilute RNA sample. Always keep RNA sample at –80 °C when not in use and on ice when in use.
11. We used standard protocols to produce electrode width controlled (EWC) single phase unidirectional SAW transducer (SPUDT).
12. We found scotch tape to be sufficient for this task.
13. The operating frequency should be 28.3 MHz and the samples we used were pumped through the device at a rate of 250 µL/h with the SAW device operating at 1 W of power. Media within the channel was exposed to the SAW portion for approximately 30 s.

14. Only 100 μL of media exposed to the SAW device was needed to transfer onto the biosensor membrane and detect target miRNA. If wanted, exosome lysis can be validated by examining exosomes via transmission electron microscopy (exosomes should look jagged and broken apart), quantifying the exosomes using a NanoSight, or by quantifying the amount of target RNA prior to and after SAW exposure using the biosensor.

References

1. Taller D, Richards K, Slouka Z et al (2015) On-Chip surface acoustic wave lysis and ion-exchange nanomembrane detection of exosomal RNA for pancreatic cancer study and diagnosis. *Lab Chip* 15:1656–1666
2. Senapati S, Slouka Z, Shah S et al (2014) An ion-exchange nanomembrane sensor for detection of nucleic acids using a surface charge inversion phenomenon. *Biosens Bioelectron* 60:92–100
3. Slouka Z, Senapati S, Chang H (2014) Microfluidic systems with ion-selective membranes. *Annu Rev Anal Chem* 7:317–335
4. Gallo A, Tandon M, Alevizos I, Illei GG (2012) The majority of MicroRNAs detectable in serum and saliva is concentrated in exosomes. *PLoS One* 7(3):e30679
5. Länge K, Rapp B, Rapp M (2008) Surface acoustic wave biosensors: a review. *Anal Bioanal Chem* 391:1509–1519
6. Taller D, Go D, Chang H (2013) Modulated exponential films generated by surface acoustic waves and their role in liquid wicking and aerosolization at a pinned drop. *Physical Review E* 87:053004
7. Hartman CS, Abbott BP (1989) Overview of design challenges for single phase unidirectional SAW filters. *Ultrasonics Symposium IEEE* 1: 79–89
8. Shilton R, Tan M, Yeo L et al (2008) Particle concentration and mixing in microdrops driven by focused acoustic waves. *J Appl Phys* 104:014910
9. Qi A, Yeo L, Friend J et al (2009) The extraction of liquid, protein molecules and yeast cells from paper through surface acoustic wave atomization. *Lab Chip* 10:470–476
10. Taller D (2015) Surface acoustic wave microfluidics: droplets, pinned liquid films, and biodection. Dissertation, University of Notre Dame

Droplet Microfluidic Device Fabrication and Use for Isothermal Amplification and Detection of MicroRNA

Maria Chiara Giuffrida, Roberta D'Agata, and Giuseppe Spoto

Abstract

Droplet microfluidics combined with the isothermal circular strand displacement polymerization (ICSDP) represents a powerful new technique to detect both single-stranded DNA and microRNA sequences. The method here described helps in overcoming some drawbacks of the lately introduced droplet polymerase chain reaction (PCR) amplification when implemented in microfluidic devices. The method also allows the detection of nanoliter droplets of nucleic acids sequences solutions, with a particular attention to microRNA sequences that are detected at the picomolar level. The integration of the ICSDP amplification protocol in droplet microfluidic devices reduces the time of analysis and the amount of sample required. In addition, there is also the possibility to design parallel analyses to be integrated in portable devices.

Key words Circular strand displacement polymerization, Isothermal amplification, Droplet microfluidics, MicroRNA, Nucleic acids amplification

1 Introduction

Droplet-based microfluidics offers many important advantages in biomolecular detection compared to continuous microfluidics and has emerged as a disruptive technology for nucleic acids detection [1]. Advantages include reduced analysis time and sample volume, simplified automation of analytical procedures and integration of different functions in a single device. All the abovementioned advantages are particularly desirable for DNA and microRNA detection [2].

The most widely used approach for nucleic acid amplification is based on the polymerase chain reaction (PCR). Integration of PCR amplification in droplet microfluidic devices has resulted in a dramatic expansion of possibilities in nucleic acid detection and quantification [3]. However, this method suffers from a number of limitations [4] being the need for a thermal cycling in vitro to reach a maximum of twofold amplification of the nucleic acid molecules after each cycle one of the most important when the full exploitation of advantages associated with the integration of nucleic

acid amplification protocols in microfluidic-based devices is going to be pursued [5, 6]. A successful strategy to overcome this limitation is offered by isothermal-amplification methods [7, 8], among which the isothermal circular strand displacement polymerization (ICSDP) stood out as a promising method for the simple nucleic acid amplification [9]. ICSDP exploits a molecular beacon (MB) fluorescent probe with a stem-loop structure. The loop carries a sequence complementary to the sequence to be targeted, while the stem ends comprise a fluorophore (6-FAM) and a quencher (Black Hole Quencher 1, BHQ). In the presence of the target MB undergoes a conformational change leading to the separation of the quencher from the fluorophore and the emission of fluorescence radiation. The annealing of a primer complementary to the stem region at 3'-end follows thus triggering the action of the DNA polymerase I (exo- Klenow fragment) leading to the primer extension and the displacement of the hybridized target sequences. Once displaced the target is made available for another hybridization/displacement cycle, thus providing a linear amplification of the target sequence (Fig. 1).

We used a Roundup-Ready soybean related DNA sequence [10] to evaluate possibilities offered by ICSDP when implemented in droplet microfluidic devices. Then the detection protocol was adapted for the detection of microRNA (miR) sequences [11].

miRs are small noncoding RNAs with important implications in clinical diagnostics [12]. miRs are involved in a large number of different biological processes, including cell fate determination, proliferation, and death. Their expression profiles have been linked to the diagnosis and prognosis of a variety of human cancers

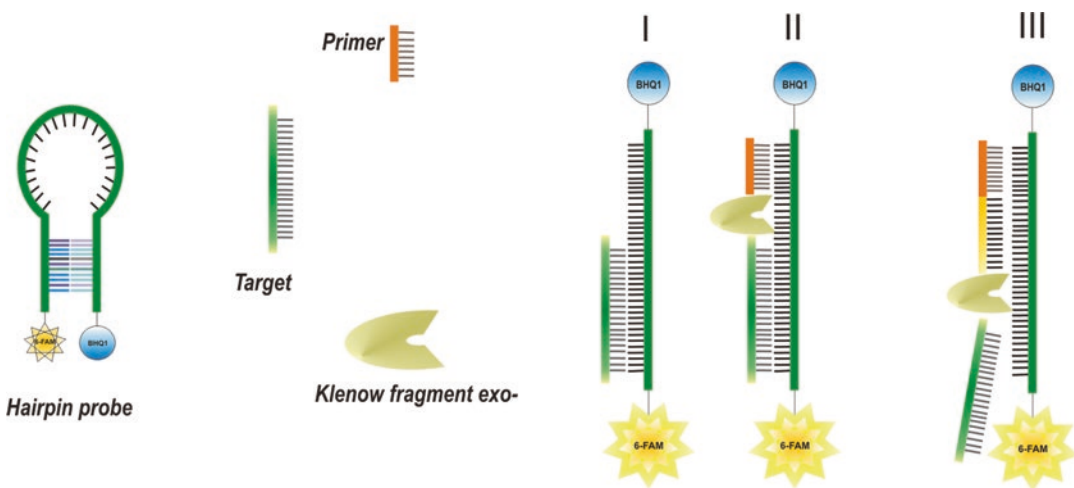


Fig. 1 Schematic description of the isothermal circular-strand displacement polymerization strategy (Reproduced from ref. 11 with permission from Springer)

Table 1
MiR-210, primers, molecular beacons sequences, and acronyms used

Name	Sequence	Acronym
Target miR-210	5'CUGUGCGUGUGACAGCGGCUGA3'	miR-210 FM
Molecular beacon	5'(6FAM) TCTTGACTTCCTTATATAGAGGAA GGGTCTTGGAAGTCAAGA (BHQ1)3'	MB
Primer	5'TCTTGTC3'	7-mer
Control sequence	5'UCAUAGAGUACUAAUAUCAUCU3'	miR-210 CTR

[13–16]. Moreover, extracellular miRs circulating in biological fluids play a key role as biomarkers [17].

miR detection is challenging because of the combination of their small size (19–23 nucleotides), low abundance, and sequence similarity among different family members [18]. Recently, different approaches for the ultrasensitive detection of nucleic acids have been described [19–22]. Nevertheless, the reliable miR detection is still a challenging.

The procedure explained in this work consists in a molecular beacon-assisted ICSDP in polydimethylsiloxane (PDMS) droplet microfluidic devices for the detection of hsa-miR-210-3p (miR-210 FM) sequence. The method is able to discriminate between the target sequence (miR-210 FM) and a control unrelated microRNA scrambled sequence (miR-210 CTR). Each droplet, generated by the combination of the aqueous phase and the immiscible carrier at the T-junction of the microfluidic device, acts as a sort of bioreactor and the fluorescence generated by the conformational change of the molecular beacon is detected at the end of the device. The MB-generated fluorescence intensity is calculated considering the average brightness expressed in pixel of a selected region of interest (ROI) inside each droplet. The average fluorescence of droplets carrying the full match sequence was subtracted from that intrinsically produced by the amplification mixture (blank).

Oligonucleotide sequences and acronyms used are indicated in Table 1.

2 Materials

1. Ultrapure nuclease-free water (resistivity at 25 °C 18.2 MΩ cm) and analytical grade reagents.
2. 10:1 mixture of 3 M Fluorinert Electronic Liquid FC-3283 and 1H,1H,2H,2H-perfluoro-1-octanol 97% (PFO).
3. 80 μm thick polyvinyl chloride adhesive film.
4. Cutting plotter.

5. Petri dish.
6. PDMS curing agent and prepolymer (Sylgard 184, Dow Corning, USA).
7. PEEK tubes (UpChurch Scientific, ID 0.508 mm; OD 0.774 mm).
8. Acrylic glue.
9. Adhesive tape.
10. Femto Diener Electronics plasma cleaner.
11. Microscope cover glass.
12. Aluminum foil.
13. Hot plate.
14. Syringe pump.
15. Pharmed BPT tubings (ID = 0.25 mm).
16. Buffer: 50 mM Tris-HCl, pH 8.0, 6% dimethylsulfoxide (DMSO), 1 mM dithiothreitol (DTT), 5 mM MgCl₂.
17. ICSDP amplification mixture: 0.31 mM molecular beacon for miR-210 (MB-miR) (Thermo Fisher Scientific), 0.31 mM 7-mer primer, 300 μ M dNTPs, 40 U Recombinant Ribonuclease Inhibitor 40 U, 10 U Klenow fragment.
18. Leica DM IL Fluo inverted microscope equipped with a Leica DFC 450C digital camera and a Lumen 200 (Prior Scientific Inc., metal-halide lamp) source.
19. Hamilton syringes model 1750 RN (volume 500 μ L) and model 1725 RN (volume 250 μ L).
20. Automated 5 μ L syringe (eVOL XR automated syringe).
21. Temperature control system consisting of a silicon heater (12 V; 1.25 W), a platinum sensing resistor (Pt 1000), a toroidal transformer (115–230 Vac Primary, 30 VA and 50 VA) (RS Components), and a microprocessor-based digital electronic controller (THP 48, Technologic) to control the temperature (± 0.1 °C) during the experiment.

3 Methods

3.1 *Microfluidic Devices*

PDMS microfluidics devices are fabricated by replica molding.

1. Create a master for replica molding by xerography using a 80 μ m thick polyvinyl chloride adhesive film and a cutting plotter [23]. After cutting the adhesive film in order to reproduce the desired shape for the microfluidic circuit stick it on the internal surface of a Petri dish.
2. Replicas are formed from a 1:10 mixture of PDMS curing agent and prepolymer.

3. Degas the PDMS mixture under vacuum for 30 min, and poured it onto the master to create a layer of about 3 mm (*see Note 1*), then allow the mixture to polymerize for 48 h at room temperature.
4. Peel off the polymerized PDMS mold and repeatedly wash it with ethanol and ultra-pure water. Dry it before use.
5. Drill holes (0.5 mm) into the PDMS microfluidic device to insert inlets and outlets represented by PEEK tubes (UpChurch Scientific, ID 0.508 mm; OD 0.774 mm). Seal tubes with acrylic glue.
6. Remove visible particulate present on the surface of the PDMS mold by using an adhesive tape (*see Note 2*).
7. Air-plasma etch the cleaned mold for 30 s using a Femto Diener Electronics plasma cleaner (40 kHz) and put in contact the treated surface with the plasma etched surface of a microscope cover glass to obtain an irreversible adhesion.
8. Place the new device wrapped into an aluminum foil (to allow a better distribution of the heating) onto a hotplate at 40–50 °C for 30 min.
9. Use microfluidic device at least 24 h after the fabrication to allow the PDMS hydrophobic surface to be recovered after the plasma treatment. The hydrophobic surface will facilitate the flow of water-in-oil droplets into the device.
10. Connect the device to syringe pumps by using Pharmed BPT tubings. An example of microfluidic devices used for the experiments is shown in Fig. 2 (*see Note 3*).
11. Carry out experiments at 37 °C (*see Note 4*).

3.2 Experiment Preparation

1. Place properly the microfluidic device on the stage of the microscope.
2. Fill two Hamilton syringes (250 μL) with the buffer solution, one with the ICSDP mixture (250 μL) and one (500 μL) with the carrier. Position syringes on the syringe pump and connect them in order to flow the ICSDP mixture through the inlet 1, the buffer solution through inlets 2, and 3 and the carrier through inlet 6.
3. Clamp inlets 4 and 5 as shown in Fig. 2c.
4. Set the automatic 5 μL syringe with the appropriate dispensing method, choosing the low dispensing speed (speed 1).
5. In order to prevent the mixing of solutions to be analyzed inset the carrier, the sample solution and the buffer solution into the automatic syringe as follow (Fig. 3):
 - (a) 200 nL of carrier;
 - (b) 200 nL of the sample (miR-210 FM) or the control (miR-210 CTR) solution (highest investigated concentration).

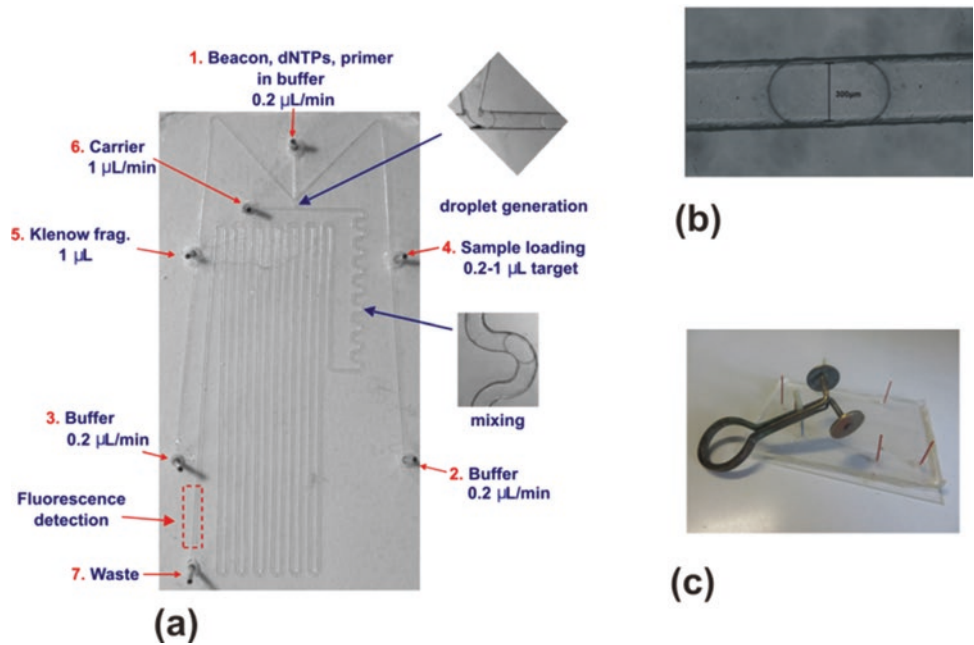


Fig. 2 (a) Microfluidic device used for the droplet-based nucleic acid detection. (b) Representative optical image of a droplet generated in the microfluidic device. (c) Clamped inlet used for the sample loading. (Adapted from ref. 11 with permission from Springer)

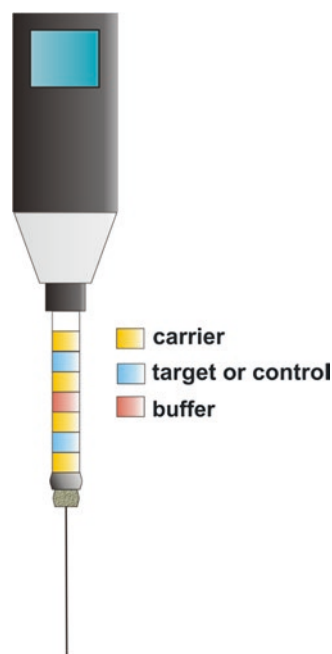


Fig. 3 Schematic representation of sample loading into the automatic syringe used for the experiment

- (c) 200 nL of carrier;
- (d) 200 nL of buffer solution (fluorescence-signal reference);
- (e) 200 nL of carrier.
- (f) Repeat **steps** from (b) to (e) for each sample solution to be analyzed.

The above described method allow to separate miscible solutions by using the immiscible carrier solution.

6. Start syringe pumps: flow rate of buffer and ICSDP mixture 0.2 $\mu\text{L}/\text{min}$; flow rate of carrier 1 $\mu\text{L}/\text{min}$.
7. When droplets with the same size are formed with a stable frequency the samples can be introduced into the device.
8. Stop syringe pumps.
9. Unclamp inlet 4 and connect it to the automatic 5 μL syringe. Then start the dispensing method. When the method is completed and the sample and control solutions are loaded into the microfluidic device, clamp again inlet 4.
10. Unclamp inlet 5 and introduce 1 μL of Klenow fragment solution by using a manually driven 1 μL Hamilton syringe model 7101 KH, then clamp again inlet 5.
11. Start syringe pumps.
12. Set the fluorescence microscope in order to detect the fluorescence generated by the molecular beacon fluorophore and position the microfluidic device under the microscope objective (10 $\times$) in order to image the final part of the microfluidic channel (Fig. 2a).
13. Start recording fluorescence images from droplets taking into accounts that after their generation droplets take about 28 min to reach the fluorescence detection area of the device under the described experimental conditions. The carrier solution used to separate miscible solutions is useful to identify droplets containing the different samples and control solution because the carrier stops the continuous generation of droplet for about 1 min.

The described method has been used to detect microRNA (miR-210 FM) hundreds of pmol/L in concentration [13].

4 Notes

1. If necessary gently remove remaining bubbles by using a pipette tip.
2. This is a critical step in the device fabrication since surface deposited dust can seriously affect the plasma bonding.
3. Insert the PEEK tubes of inlet 5 and 4 into short segments of Pharmed tube and clamp them as shown in Fig. 2c.

4. Choose a silicon heater with the appropriate dimension in order to heat the microfluidic device by leaving the fluorescence detection area accessible for the fluorescence microscope detection.

Acknowledgments

MIUR (PRIN 20093N774P) and Ministry of Health, Italy (n. 098/GR-2009-1596647), are acknowledged for partial financial support.

References

1. Schneider T, Kreutz J, Chiu DT (2013) The potential impact of droplet microfluidics in biology. *Anal Chem* 85:3476–3482
2. Spoto G, Corradini R (2012) Detection of non-amplified genomic DNA. Springer, Verlag
3. Hindson CM, Chevillet JR, Briggs HA, Gallichotte EN, Ruf IK, Hindson BJ, Vessella RL, Tewari M (2013) Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods* 10:1003–1005
4. Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* 7:e43093
5. Craw P, Balachandran W (2012) Isothermal nucleic acid amplification technologies for point-of-care diagnostics: a critical review. *Lab Chip* 12:2469–2486
6. Asiello PJ, Baeumner AJ (2011) Miniaturized isothermal nucleic acid amplification, a review. *Lab Chip* 11:1420–1430
7. Zanolini LM, Spoto G (2013) Isothermal amplification methods for the detection of nucleic acids in microfluidic devices. *Biosensors* 3:18–43
8. Kim J, Easley CJ (2011) Isothermal DNA amplification in bioanalysis: strategies and applications. *Bioanalysis* 3:227–239
9. Guo Q, Yang X, Wang K, Tan W, Li W, Tang H, Li H (2009) Sensitive fluorescence detection of nucleic acids based on isothermal circular strand-displacement polymerization reaction. *Nucleic Acids Res* 37(3):e20
10. D'Agata R, Corradini R, Ferretti C, Zanolini L, Gatti M, Marchelli R, Spoto G (2010) Ultrasensitive detection of non-amplified genomic DNA by nanoparticle-enhanced surface plasmon resonance imaging. *Biosens Bioelectron* 25:2095–2100
11. Giuffrida MC, Zanolini LM, D'Agata R, Finotti A, Gambari R, Spoto G (2015) Isothermal circular-strand-displacement polymerization of DNA and microRNA in digital microfluidic devices. *Anal Bioanal Chem* 407(6):1533–1543
12. Nana-Sinkam SP, Croce CM (2013) Clinical applications for microRNAs in cancer. *Clin Pharmacol Ther* 93:98–104
13. Dong H, Lei J, Ding L, Wen Y, Ju H, Zhang X (2013) MicroRNA: function, detection, and bioanalysis. *Chem Rev* 113:6207–6233
14. Mendell JT, Olson EN (2012) MicroRNAs in stress signaling and human disease. *Cell* 148:1172–1187
15. Zampetaki A, Mayr M (2012) MicroRNAs in vascular and metabolic disease. *Circ Res* 110:508–522
16. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
17. Tzimagiorgis G, Michailidou EZ, Kritis A, Markopoulos AK, Koudou S (2011) Recovering circulating extracellular or cell-free RNA from bodily fluids. *Cancer Epidemiol* 35:580–589
18. Yan L, Yan Y, Pei L, Wei W, Zhao J (2014) A G-quadruplex DNA-based, label-free and ultrasensitive strategy for microRNA detection. *Sci Rep* 4:7400
19. D'Agata R, Breveglieri G, Zanolini LM, Borgatti M, Spoto G, Gambari R (2011) Direct detection of point mutations in non-amplified human genomic DNA. *Anal Chem* 83:8711–8717
20. D'Agata R, Spoto G (2013) Surface plasmon resonance imaging for nucleic acid detection. *Anal Bioanal Chem* 405:573–584
21. Zanolini LM, D'Agata R, Spoto G (2012) Functionalized gold nanoparticles for the ultrasensitive DNA detection. *Anal Bioanal Chem* 402:1759–1771
22. Spoto G, Minunni M (2012) Surface plasmon resonance imaging: what next? *J Phys Chem Lett* 3:2682–2691
23. Grasso G, D'Agata R, Zanolini L, Spoto G (2009) Microfluidic networks for surface plasmon resonance imaging real-time kinetics experiments. *Microchem J* 93:82–86

Interrogation of Functional miRNA–Target Interactions by CRISPR/Cas9 Genome Engineering

Yale S. Michaels*, Qianxin Wu*, and Tudor A. Fulga

Abstract

Post-transcriptional silencing by microRNAs (miRNAs) is a critical constituent of eukaryotic gene regulation. miRNAs are short (~22nt) noncoding RNAs capable of specifically targeting the miRNA-induced-silencing-complex (miRISC) to transcripts bearing a complementary miRNA response element (MRE). Although recent methodological advances have greatly improved our understanding of miRNA biogenesis and the mechanisms by which miRNAs repress their cognate targets, exploring the physiological relevance of direct miRNA–target interactions in vivo has remained an outstanding challenge. Here we describe the experimental protocol underlying a novel approach, which allows direct interrogation of specific miRNA–MRE interactions by CRISPR/Cas9-mediated genome engineering. In this instance, the CRISPR/Cas9 system is first used to catalyze homology-directed replacement of candidate MREs with molecular barcodes at endogenous loci. Subsequently, the effect of MRE mutation on transcript abundance (i.e., MRE activity) can be rapidly evaluated by routine quantitative PCR. This strategy enables functional investigation of a putative miRNA–target pair in a pool of transiently transfected cells, obviating the need for generation of clonal cell lines or transgenic animals. This protocol can be implemented in any cell line in less than 2 weeks, and can readily be scaled up for multiplex studies. To facilitate the conceptual workflow underlying this strategy, we also describe a genome-wide resource for automated design and computational evaluation of CRISPR/Cas9 guide RNAs targeting all predicted MREs in various species (miR-CRISPR).

Key words microRNAs, miRNA, miRNA response elements (MRE), Genome engineering, CRISPR/Cas9

1 Introduction

microRNAs (miRNAs) post-transcriptionally down-regulate target gene expression in a sequence-specific manner [1]. Effective target repression depends on complementarity between miRNAs and short miRNA response elements (MREs) in the target transcript [2]. Several studies have implicated nucleotides 2–8 of the mature

*These authors contributed equally.

miRNA as critical determinants for target selection and repression [2, 3]. Consequently, a number of algorithms have been developed to computationally identify, in a genome-wide fashion, putative MREs based on 6nt “seed” regions and other sequence determinants including evolutionary conservation, AU content, position within the 3′-UTR, etc. [4–6]. This endeavor revealed that, by and large, each individual miRNA is predicted to bind numerous MREs, suggesting that collectively miRNAs may target nearly the entire coding transcriptome in mammals [7]. As a result, discriminating unique functional MREs is a major challenge for studies seeking to understand miRNA-mediated gene regulatory networks *in vivo*. miRNA gain-of-function (GOF) and loss-of-function (LOF) studies have the power to identify differentially expressed genes upon alterations in miRNA homeostasis [8–10]. However, definitively distinguishing between primary and secondary miRNA targets on the basis of differential transcriptional or proteome profiling alone has proven a difficult task.

The prevailing approach developed to address this issue is the use of synthetic MRE sensors. In this instance, the 3′UTR of a gene bearing an MRE of interest is fused to a fluorescent or bioluminescent reporter transcript (XFP or luciferase) [11]. A reduction in reporter expression upon inclusion of a functional MRE, and the loss of this effect following mutation of the putative seed sequence, is employed to infer a functional miRNA–target interaction. Although such reporters are easy to use, they display several limitations, including most notably the failure to capture the endogenous miRNA–target stoichiometry. Therefore, they can only serve as a proxy for physiological miRNA-mediated repression *in vivo*. Alternatively, capture-based techniques (CLIP variants, CLASH, etc.) have been used to detect spatial-temporal binding events between miRNAs and endogenous targets [12–14]. However, since it became apparent that miRNA–target binding does not necessarily mirror repressive action, these methods fall short of definitively uncovering functional interactions [15]. Therefore, drawing a causative link between specific miRNA–MRE interactions and endogenous gene repression has remained elusive. We have recently pioneered a cell-based genome engineering strategy for direct functional validation of miRNA–MRE interactions under physiological conditions and with minimal impact on cellular homeostasis [16].

The recent implementation and optimization of the clustered, regularly interspaced, short palindromic repeats (CRISPR)-associated Cas9 programmable nuclease, has rendered genome editing a routine laboratory procedure [17]. The native Cas9 nuclease is capable of generating precise DNA double strand breaks in the genome and can be targeted to any 20nt locus adjacent to a protospacer adjacent motif (PAM) of the sequence NGG or NAG. A chimeric synthetic guide RNA (sgRNA) delivered along

with the Cas9 protein specifies the target genomic locus. In addition to generation of insertions or deletions (*indels*) at the targeted locus, specific nucleotide changes or insertions can be introduced at/near the Cas9 cleavage site when a homology template is supplied alongside the sgRNA and active Cas9 nuclease.

Here, we describe a four-step protocol to rapidly and accurately determine the functionality of a MRE in an immortalized human cell line (Fig. 1) [16]. First, the dedicated web tool miR-CRISPR is used to select a PAM within, or in close proximity of an MRE of interest and design a suitable sgRNA. Second, a pair of barcoded single stranded DNA oligonucleotides (ssODN) is designed, of which one mutates the MRE under investigation and replaces it with a barcode (ssODN^{mut-MRE}), while the other maintains the MRE and adds a second barcode immediately adjacent to the MRE seed sequence (ssODN^{WT-MRE}). Third, the sgRNA is synthesized and cloned into an appropriate Pol-III-driven sgRNA/Pol-II-driven Cas9 dual expression vector. Fourth, the mixed barcoded ssODNs are simultaneously co-transfected with the sgRNA/Cas9 expression vector into a mammalian cell line. Finally, quantitative PCR (qPCR) is used to determine the MRE^{mut}/MRE^{WT} barcode ratio in genomic DNA (gDNA) and RNA (cDNA) extracted simultaneously from the same pool of cells (Fig. 1b). The ratio of the two barcodes in the gDNA is used to estimate the relative homology directed repair (HDR) integration efficiency of the two ssODNs. The activity of each MRE is established as a function of MRE^{mut} barcode enrichment in the cDNA sample, resulting in an altered MRE^{mut}/MRE^{WT} ratio when compared to the gDNA.

2 Materials

2.1 sgRNA and ssODN Template Design Resources

1. sgRNA design: miR-CRISPR: <http://mir-crispr.molbiol.ox.ac.uk/fulga/miR-CRISPR.cgi>.
2. Prediction of ectopic MREs generation or inadvertent removal of other MREs by ssODN: TargetScan source code Perl script http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_70), or PITA (http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html).

2.2 sgRNA Cloning and ssODN Synthesis Reagents

1. Dual sgRNA/Cas9 expression vector backbone (human): *pX330*, *pX458*, *pX459*.
2. Restriction enzyme: BbsI Fast-Digest + 10xFastDigest Buffer.
3. T4 DNA ligase and buffer.
4. Antarctic Phosphatase and buffer.
5. High efficiency competent *E. coli* 5- α .
6. QIAprep Spin Miniprep Kit.
7. QIAquick PCR Purification kit.

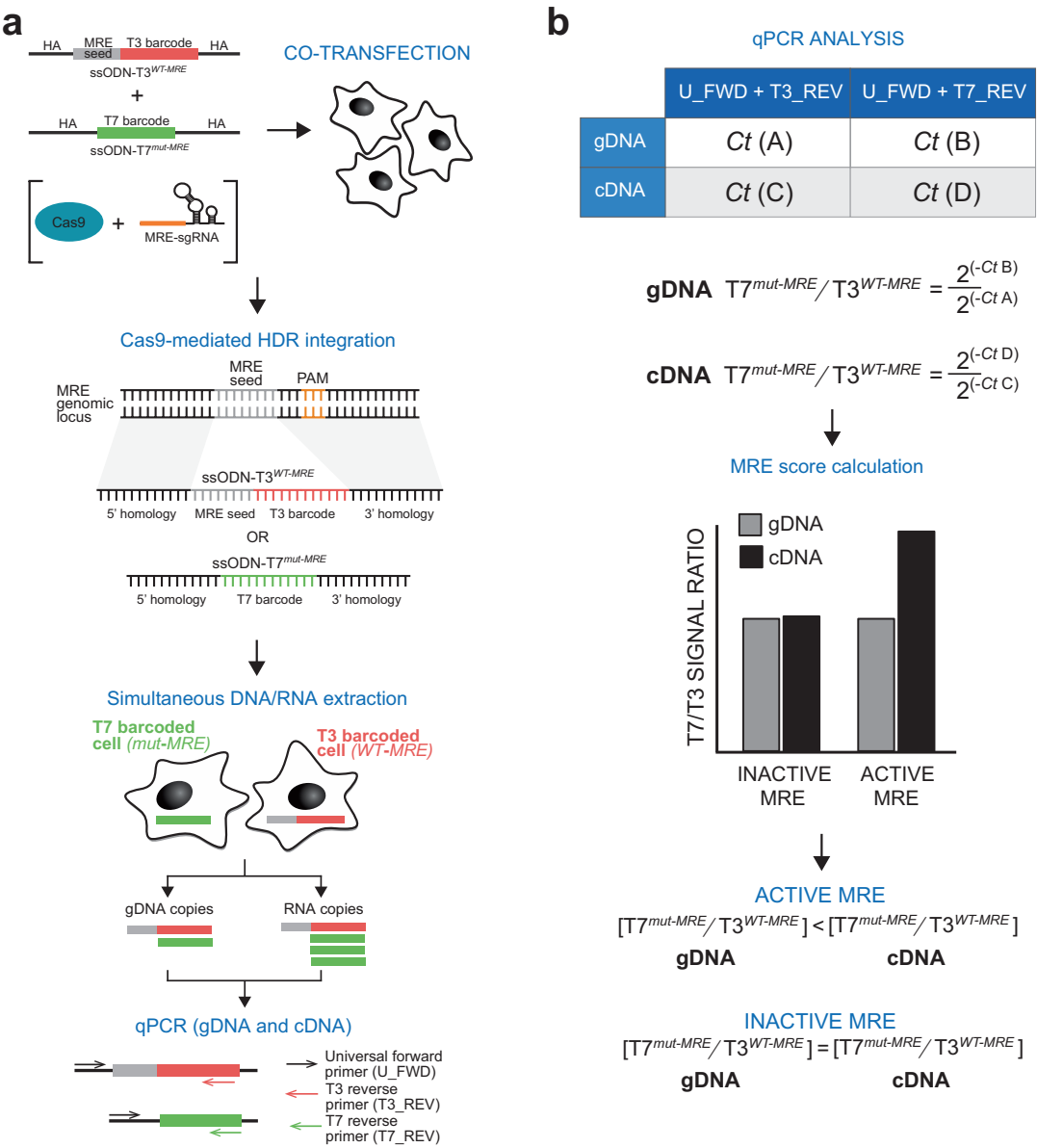


Fig. 1 Functional interrogation of MRE activity by paired barcode genome engineering—experimental workflow. **(a)** Mammalian cells are transiently co-transfected with a dual expression plasmid (Pol-II driven Cas9 and Pol-III driven chimeric sgRNA, targeting an MRE genomic locus of interest), and a pair of ssODN HDR templates. ssODN-T3^{WT-MRE} appends a T3 barcode (red) downstream of the MRE seed sequence of interest (grey) and ssODN-T7^{mut-MRE} will replace the MRE with a T7 barcode (green). Following Cas-9 mediated HDR at the MRE genomic locus, cells are harvested, DNA and RNA are simultaneously extracted from the same pool of cells, and used as templates for qPCR using a gene-specific forward primer, and either a T3 or T7 reverse primers. **(b)** The ratio of T7^{mut-MRE} to T3^{WT-MRE} in gDNA is used to estimate the relative integration efficiency of the two oligos. The activity of each MRE is established as a function of T7^{mut-MRE} to T3^{WT-MRE} ratio in cDNA relative to the corresponding ratio in the gDNA sample

8. Oligos for cloning sgRNA spacer sequence.
9. *px330_Seq* primer: 5'-ACTATCATATGCTTACCGTAAC-3'.
10. Ultramers for ssODN HDR template generation.
11. NanoDrop 2000c.

2.3 Cell Culture and Transfection Reagents

1. Dulbecco's Modified Eagle Medium.
2. Opti-MEM® I Reduced Serum Media.
3. Fetal bovine serum.
4. Penicillin–streptomycin.
5. Phosphate-buffered saline.
6. Culture flasks.
7. 12-well plates.
8. Polyethylenimine, branched (PEI).

2.4 RNA and DNA Extraction Reagents

1. DNA extraction buffer: 10 mM Tris–Cl pH 8.2, 1 mM EDTA, 25 mM NaCl, and 200 µg/ml Proteinase K (add the Proteinase K immediately before use).
2. DNA and RNA parallel extraction kit.
3. ExoSAP-IT.

2.5 High Resolution Melt Analysis (HRMA) Components

1. HRMA primer design (<http://primer3.ut.ee>).
2. HRMA primers.
3. Hotshot Diamond PCR mastermix.
4. LC-Green dye.
5. CFX384 Touch™ Real-Time PCR Detection System with Precision Melt Analysis Software.

2.6 RT-qPCR

1. QuantiTect Reverse Transcription Kit.
2. Gene specific forward primer: same as the HRMA FWD primer.
3. Gene specific reverse primer: complementary to T7 and T3 barcodes.
4. SsoAdvanced™ Universal SYBR® Green Supermix.
5. 384-well PCR Plates Hard Shell.
6. CFX384 Touch™ Real-Time PCR Detection System.

3 Methods

3.1 sgRNA Design: miR-CRISPR

The miR-CRISPR web tool (<http://mir-crispr.molbiol.ox.ac.uk/fulga/miR-CRISPR.cgi>) is an integrated interface for selecting appropriate CRISPR/Cas9 target sequences in the proximity of MRE(s) predicted by the miRanda algorithm (<http://www>.

microrna.org), or in any user-defined custom sequence (Fig. 2). General guidelines on using the miR-CRISPR platform are provided below.

1. Select the organism of interest (Fig. 2a). Currently, the database allows genome-wide analysis of all Human, Mouse, Rat, *Drosophila* and *C. elegans* predicted MREs.
2. Choose a search mode (Fig. 2b). Available modes are *target gene name*, *miRNA name*, or *custom sequence* (user-defined). For high-throughput applications, a batch mode option is available for all modes. When selected, this option will output sgRNA sequences for all predicted MREs for a particular miRNA or all MREs within a particular gene in a table format.
3. Choose sgRNA design parameters (Fig. 2c). If using the *pX330* backbone vector in mammalian cells as described in this protocol, input “Mammalian U6” and “4 bp (*BsmBI*, *BbsI*, *BsaI*)” in the *Promoter/RE* field. Specify the remaining parameters based on preference and MRE architecture. As noted, NGG PAMs generally yield more efficient cleavage. Select “*Get CRISPR Targets?*” to access the output page.
4. On the output page (Fig. 2d), the algorithm will return:
 - (a) Information on the selected target gene and miRNA of interest.
 - (b) links to relevant repositories and web-based resources (NCBI, miRBase, miRanda).
 - (c) Putative MRE(s) sequence.
 - (d) Characteristics of output sgRNAs: on-target sequence, PAM identity, PAM distance from the MRE, ready to use oligos for appending the desired spacer sequence to a chimeric sgRNA backbone in a destination vector of choice, including the appropriate overhangs required for cloning (highlighted in blue).

Fig. 2 Design of CRISPR/Cas9 sgRNAs for MRE targeting—the miR-CRISPR interface. **(a)** miR-CRISPR front page. The algorithm uses miRanda-based MRE predictions for human, mouse, rat, *Drosophila*, and *C. elegans*. **(b)** miR-CRISPR search modes: all MREs within a gene of interest (Target gene query), all MREs targeted by one miRNA (miRNA query) or all MREs within a custom sequence (User defined sequence query). Both a candidate gene/miRNA output and a batch mode output are available. **(c)** All modes are supported by a number of flexible search parameters, which integrate features relevant to the design of CRISPR/Cas9 genome engineering experiments. These features include sgRNA expression system and cloning strategy, PAM identity, and PAM position with respect to the MRE seed sequence. If the “*miRNA*” search mode is selected, a drop-down list will display all genes predicted to carry cognate MREs. If the “*Gene target*” or “*User-defined MRE*” search mode are selected, the list will display all miRNAs predicted to target the selected gene/sequence. **(d)** The output page returns a list of candidate CRISPR target sites sorted by distance from the MRE of interest, as well as the oligos required to clone the spacer sequence into a chosen chimeric sgRNA expression vector. Details about potential off-target sites can be accessed on demand for each sgRNA

a

miR-CRISPR
Genome engineer
microRNA response elements

Home About Help Contact + CBRG

Welcome to miR-CRISPR

The miR-CRISPR tool is an integrated algorithm designed to facilitate CRISPR / Cas9 genome engineering of putative microRNA response elements (MREs). We appreciate your feedback or suggestions, as we aim to continuously improve and expand the scope of this resource.

[Read the publication +](#)

Organism: **Human** Mouse Rat Drosophila C. elegans

Search type: **Gene Targets** miRNAs User-defined MRE

Search:

19281 human genes available

Get miRNAs **Batch Mode**

Batch mode will display only the nearest CRISPR target for all miRNAs predicted to interact with the selected gene.

b

Target gene query

Organism: **Human** Mouse Rat Drosophila C. elegans

Search type: **Gene Targets** miRNAs User-defined MRE

Search:

19281 human genes available

Get miRNAs **Batch Mode**

miRNA query

Organism: **Human** Mouse Rat Drosophila C. elegans

Search type: **Gene Targets** miRNAs User-defined MRE

Search:

249 human miRNAs available

Get Genes **Batch Mode**

User defined sequence query

Organism: **Human**

Sequence:

MRE seed position:

MRE seed length:

View example data **Get CRISPR Targets** **Cancel**

c

Search parameters

Please specify your CRISPR design parameters...

Promoter/RE choice: **Mammalian U6** 4bp (BsmBI, BbsI, BsaI)

PAM sequence: ☐ NGG-only ☒ NGG or NAG
NGG PAMs are more efficient but less frequent

MRE Overlap: ☒ Allow CRISPR PAM site to overlap predicted miRNA Response Element (MRE)

Choose Gene: **A2BP1**
2499 genes predicted to interact with Human hsa-miR-21

d

Output page

PDCD1 hsa-miR-374a

Gene: **PDCD1**

miRNA: **hsa-miR-374a**

Filters: **PAM sequence: any**
PAM/MRE overlap: yes

Links: **Gene:** [NCBI](#)
miRNA: [miRBase](#)
Gene + miRNA: [miRanda](#)

Promoter RE Overlap: **miR, 4bp (BsmBI, BbsI, BsaI) CAG: (BbsI, AAC (rev)**

MREs: **1. CUUCCUGUGGUUUAUUAU**

MRE #1 miRanda-predicted MRE: **CUUCCUGUGGUUUAUUAU**

Genomic co-ordinates: **chr2:242792058-242792069** [UCSC](#)

Transcript: **NM_005018**

Transcript strand: **-ve**

CRISPR sites to target this MRE:

CRISPR target sequence	PAM	Distance from MRE seed (bp)	Forward oligo (5'-3')	Reverse oligo (5'-3')	Off-target search
TGATATATATATATAT	TAG	Overlaps MRE	GACCGTAATATATATATATATAT	AAACTATATATATATATATAC	search
TGGTCTATATATATAT	CAG	6	GACCGATATATATATATATATAT	AAACTGGTCTATATATATATAT	search
CAAGGATCCCTCCCTG	TGG	6	GACCGCAAGGATCCCTCCCTG	AAACCAAGGATCCCTCCCTG	search

The CRISPR target sequences are ranked based on the distance from the MRE and the identity of the PAM sequence. For each sgRNA, a genome-wide off-target search option is available which will return detailed information on all possible off-targets containing up to four mismatches from the on-target site (chromosomal location, mismatch number and position, genetic context, links to relevant repositories). Although an off-target evaluation option is provided, for the purpose of this protocol, off-target events are unlikely to interfere with the assay output for several reasons: (1) the assay is designed to detect changes in gene expression associated with altering a unique candidate MRE. Therefore, any off-target editing event could only elicit detectable readout interference if it overlaps within the same gene locus, which is a very unlikely occurrence. (2) Since the assay is carried out at the level of pooled cell populations, the rate of off-target mutagenesis would have to be very high to produce any detectable effects.

3.2 ssODN Design for HDR-Mediated MRE Replacement

Although it is possible to alter or ablate MREs in the genome by CRISPR/Cas9-mediated non-homologous-end-joining (NHEJ), to generate a more predictable outcome we recommend replacing candidate MREs with specific barcodes by HDR [16]. This method also offers the advantage that the consequence of MRE deletion can be assessed by standard real-time quantitative PCR. This strategy can be readily implemented by co-transfecting two barcoded single stranded oligonucleotide (ssODN) donor templates alongside the Cas9/sgRNA expression plasmid. The barcodes described in this protocol (T3 and T7) were chosen because these sequences provide high specificity PCR priming sites. However, the combination of barcodes must be carefully considered on a case-by-case basis, in order to avoid the possibility of generating new MREs or inadvertently removing others. In addition, because the field is rapidly evolving and improved HDR template generation principles are constantly emerging [18, 19], the design described below is only meant to serve as a guideline (*see Note 1*).

1. Design two ssODNs: ssODN^{mut-MRE} and ssODN^{WT-MRE} of the format 5'-N₆₀-tag-N₆₀-3' (Fig. 3). To generate tag^{mut}, replace the MRE with the T7 barcode (5'-TAATACGACTCACT ATAGGG-3'). To generate tag^{WT}, insert the T3 barcode (5'-AATTAACCCTCACTAAAGGGA-3') downstream of the MRE seed sequence. N₆₀ stands for homology arms ~60nt upstream and ~60nt downstream of the MRE seed sequence of interest. The final ssODNs should have the following structure:

$$\begin{aligned} \text{ssODN}^{\text{mut-MRE}} &: 5' \text{-N}_{60}\text{-[MRE seed}^{\text{deletion}}\text{]T7-N}_{60}\text{-3'}. \\ \text{ssODN}^{\text{WT-MRE}} &: 5' \text{-N}_{60}\text{-MRE-T3-N}_{60}\text{-3'}. \end{aligned}$$
2. Order the ssODNs as 4 nmol Ultramers from IDT. In our experience, standard desalting appears to be sufficient for reproducible results (PAGE purification of oligos is not necessary) (*see Note 2*).

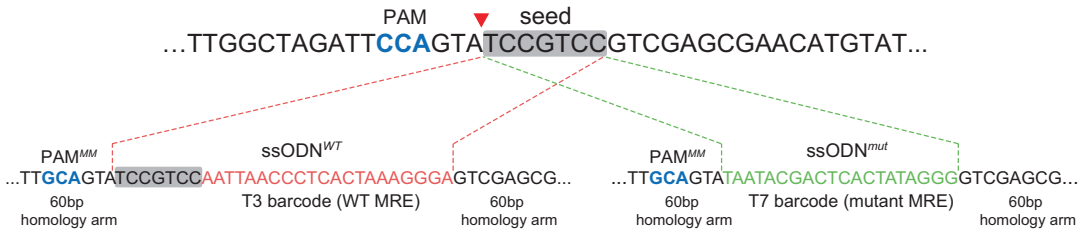


Fig. 3 Design of paired single-stranded oligonucleotide homology donors (ssODNs) for MRE replacement and barcoding. One ssODN inserts a T3 barcode at the endogenous MRE locus (*red*) while the second ablates the MRE and replace it with a distinct T7 barcode (*green*). Both ssODNs include ~60nt homology arms upstream and downstream of the MRE seed sequence. To prevent further cleavage following HDR replacement, the PAM is mutated from CCA (genomic sequence) to GCA in the ssODN (*blue*)

3.3 sgRNA Cloning (*pX330* Destination Vector)

This protocol is adapted from genome-engineering.org and Neena K. Pyzocha, et al. [20].

For many cell-culture based applications, rapid and efficient MRE deletion and replacement can be achieved using a plasmid-based CRISPR system. *pX330* is a highly optimized backbone vector for genome editing in mammalian cells [20]. This plasmid encodes the human codon optimized *S. pyogenes* Cas9 gene driven by a *CBh* promoter as well as a chimeric sgRNA scaffold under the control of a human *Pol-III U6* promoter. This vector only requires the insertion of a user-defined spacer sequence targeting an MRE(s) of interest, using short dsDNA oligonucleotides and Type-IIS restriction enzyme cloning. For difficult to transfect cell types, the *pX330* backbone can be replaced with *pX459*, which allows for enrichment of transfected cells by puromycin selection (*see Note 3*). The oligonucleotide sequences, as generated by the miR-CRISPR algorithm, can be purchased from any favorite supplier (we recommend IDT oligos for their reproducibility), and in our experience standard desalting is sufficient and no additional purification steps are necessary.

1. *pX330* backbone vector preparation for oligo cloning: Mix 2 µg *pX330*, 0.5 µl BbsI Fast-Digest (10 k U/ml), 2 µl 10xFastDigest Buffer and make it up with water to 20 µl. Centrifuge briefly and incubate at 37 °C for 1 h.
2. Dephosphorylate digested *pX330* vector: Add 0.5 µl Antarctic Phosphatase + 2.2 µl Antarctic Phosphatase buffer to the reaction above. Incubate for 10–15 min at 37 °C.
3. Purify the digested and dephosphorylated *pX330* plasmid using the Qiagen QIAquick PCR Purification Kit. Elute the linearized product in 30 µl of nuclease-free H₂O or Qiagen elution buffer (EB) and determine the concentration of recovered plasmid DNA by NanoDrop.
4. Resuspend lyophilized forward and reverse sgRNA spacer oligos in nuclease-free H₂O to a final concentration of 100 µM.

Pipette/vortex thoroughly to mix. This step can be carried out in parallel with **steps 1–3**.

5. Oligo annealing and phosphorylation.

Forward oligo (100 μ M)	1 μ l
Reverse oligo (100 μ M)	1 μ l
T4 PNK	0.5 μ l
10xT4 ligase buffer	1 μ l
H ₂ O	6.5 μ l
Total	10 μ l

Incubate reaction at 37 °C for 30 min/95 °C for 5 min/ramp down to 25 °C at 5 °C/min.

6. Ligation into *pX330* vector.

Oligo duplex <i>from 5</i> (1/100)	1 μ l
pX330 <i>from 3</i>	50 ng
10xT4 ligase buffer	1 μ l
T4 Ligase	0.5 μ l
H ₂ O	X μ l
Total	10 μ l

Incubate at room temperature for 30–90 min (*see Note 4*).

7. Transform chemically competent bacteria: Thaw a 50 μ l aliquot of NEB 5-alpha Competent *E. coli* cells for 10 min on ice, gently pipette 2 μ l of ligation reaction from **step 6** and incubate for 30 min on ice. Heat-shock for 30 s at 42 °C and then immediately place on ice for 2 min. Add 950 μ l SOC media and shake at 225 rpm for 1 h at 37 °C. During this incubation step, pre-warm an LB-agar-ampicillin plate at 37 °C. Pellet bacteria by slow centrifugation at 2500 $\times g$ for 3 min. Remove 800 μ l SOC and resuspend the pellet in the remaining 200 μ l media. Plate approximately 50–100 μ l on a pre-warmed agar plate and incubate overnight at 37 °C.
8. Pick 1–3 colonies and grow overnight in 5 ml LB+Amp at 37 °C and 225 rpm.
9. Isolate plasmid DNA using the Qiagen QIAprep Spin Miniprep Kit (elute in 50 μ l H₂O or EB) and confirm presence of correct sgRNA spacer by Sanger sequencing using the *pX330_Seq* primer.

3.4 MRE Engineering in HEK-293T Cells

This protocol applies to transfection of CRISPR/Cas9 DNA reagents in HEK-293T cells. Other cell lines may require additional optimization of transfection conditions and reaction parameters. Cultured cells should be manipulated in a biosafety cabinet using sterile reagents in accordance with standard protocols.

1. HEK-293T cells are cultured at 37 °C and 5% CO₂ in DMEM supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (complete media). Passage cells every 2–3 days (at approximately 70% confluency).
2. Seed 1×10^5 cells per well (in 1 ml media) in a 12 well plate 24 h prior to transfection.
3. HEK-293T cells can be transfected using a range of commercial reagents. We recommend PEI as a cost effective alternative [21], which in our hands performs equivalently to Lipofectamine 2000 or other commercial reagents.

We recommend setting up three independent transfections for each sgRNA/target MRE locus. Set up control wells using a fluorescence reporter plasmid such as pmaxGFP to establish transfection efficiency, as well as an un-transfected control. For each transfection set up the following reaction:

pX330-sgRNA plasmid DNA	0.5 µg
ssODN ^{WT-MRE} oligo (100 µM)	0.25 µg
ssODN ^{mut-MRE} oligo (100 µM)	0.25 µg
OPTIMEM	100 µl
Mix well	
Add PEI	3 µl
Vortex 10–15 s	
Incubate 15 min RT	

During incubation, aspirate the media from the 12 well plate (**step 2**), wash once with PBS and replace with 1 ml fresh DMEM supplemented with 2% FBS. Add transfection mix to cells dropwise and incubate at 37 °C/5% CO₂ for 48 h. To maximize cell viability replace transfection media after 24 h with complete media.

4. Harvest cells 48 h post-transfection. Remove media, wash once with PBS and incubate in 50 µl trypsin (0.01%) for 5 min at 37 °C. Add 1 ml complete media to neutralize trypsin.
5. Centrifuge cells at $500 \times g$ for 3 min, aspirate the supernatant and resuspend in 500 µl cold PBS to wash. Repeat this wash step three times, on the final wash, remove all PBS and either proceed immediately to DNA/RNA extraction (Subheading 3.4.2) or snap-freeze cells in liquid nitrogen and store at –80 °C.

3.4.1 HDR-Mediated MRE Replacement

1. Lyse cells by adding 350 µl buffer RLT (Qiagen) directly to each fresh or frozen pellets.
2. Simultaneously extract DNA and RNA from each well using the Qiagen AllPrep DNA/RNA Mini Kit following the manufacturer's instructions (*see Note 5*).

3. Since residual ssODNs can interfere with the qPCR analysis, it is necessary to treat the gDNA sample with ExoSAP-IT. Alternatively, the same result can be obtained using the following custom-made protocol.

gDNA	7.5 µl
Shrimp-AP	1 µl
Exonuclease-X	0.1 µl
NEB Cutsmart buffer	1.5 µl
H2O	4.9 µl
Total	15 µl
Incubate 60 min at 37 °C	
Inactivate 20 min at 80 °C	

4. Generate cDNA from 1 µg total RNA using the Qiagen QuantiTect Reverse Transcription Kit according to the manufacturer's instructions (*see* **Note 6**).
5. Design specific forward primers approximately 100–150nt upstream of each MRE seed sequence of interest, with a T_m of 60–65 °C. Synthesize reverse primers complementary to the designed MRE^{mut} and MRE^{wt} barcodes (in this case, T7 and T3 respectively) (*see* **Note 7**).
6. For each MRE run the following combinations of template and primers.

Sample	Fwd Primer	Rev Primer	Template	Annotation
A	MRE specific	T3 rev	gDNA	WT MRE gDNA
B	MRE specific	T7 rev	gDNA	Mutant MRE gDNA
C	MRE specific	T3 rev	cDNA	WT MRE cDNA
D	MRE specific	T7 rev	cDNA	Mutant MRE cDNA

7. For each sample above set up the following reaction. We recommend using the SsoAdvanced Universal SYBR Green Supermix (BioRad).

Template DNA (gDNA or cDNA)	1 µl
Fwd. primer (10 µM)	1 µl
Rev. primer (10 µM)	1 µl
SsoAdvanced Universal SYBR Green Supermix	10 µl
Nuclease free H ₂ O	7 µl
Total	20 µl

8. Run the following protocol on a real-time PCR instrument (BioRad CFX384 Touch™ Real-Time PCR Detection System).

	Temperature	Time	Cycles
Denaturation	98 °C	30 s	1
Annealing	Primer specific	5–15 s	35–40
Extension + Plate read	60 °C	15–30 s	
Melt-curve analysis	65–95 °C	0.5 °C increment 2–5 s/step	1

9. Calculate the MRE score as shown in Fig. 1b and Fig. 4 and detailed below. Obtain the raw cycle of threshold (Ct) for reactions A, B, C and D. Calculate technical triplicates average Ct for each transfected well. For each transfected well, determine the MRE^{mut}/MRE^{WT} gDNA (gDNA T7/T3) ratio $[2^{(-B)}/2^{(-A)}]$

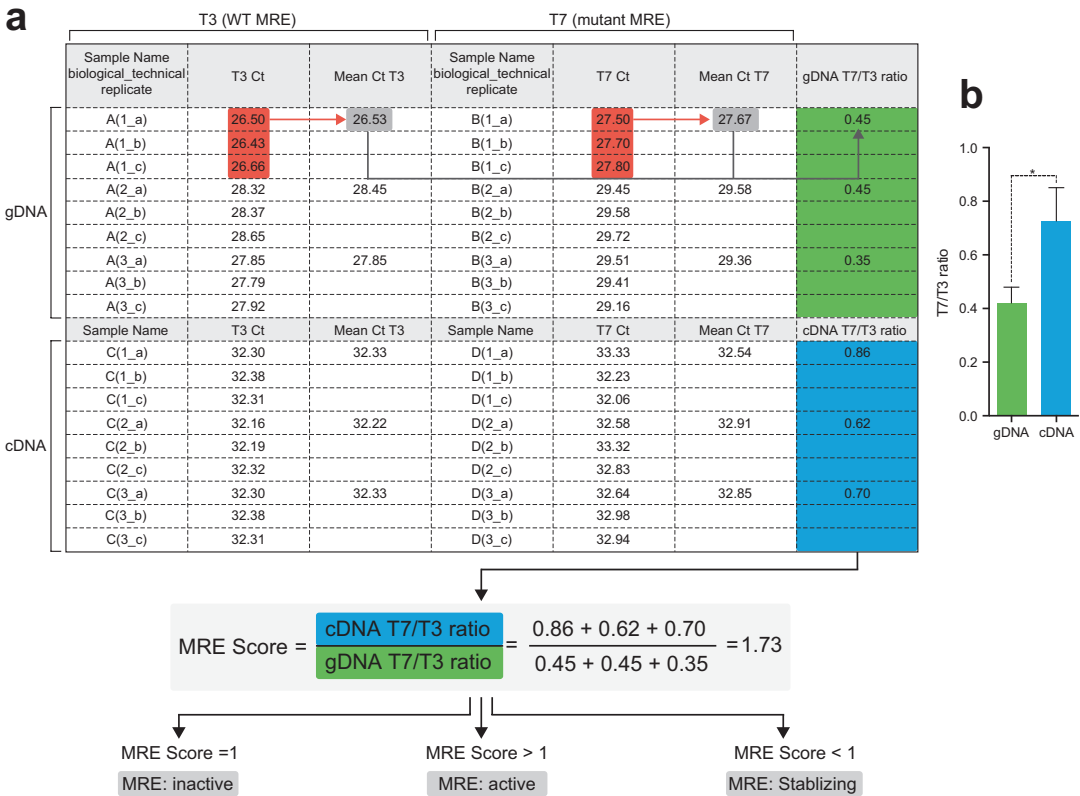


Fig. 4 Quantitative PCR analysis and MRE score calculation. **(a)** Raw cycle of threshold (Ct) values are recorded for technical and biological triplicates in gDNA and cDNA using reverse primers complementary to either the T3 barcode (WT MRE) or T7 barcode (mutant MRE). Technical triplicates are averaged and a T7/T3 ratio is calculated for gDNA and cDNA within each biological sample. An MRE score is calculated by dividing the average cDNA T7/T3 ratio by the average gDNA T7/T3 ratio. **(b)** Experimental MRE score calculated for a candidate miR-92a MRE in C9orf7 in HEK-293T cells (data source Basset et al. [16])

and the MRE^{mut}/MRE^{WT} cDNA (cDNA T7/T3) ratio $[2^{(-D)}/2^{(-C)}]$. Average ratios across triplicate transfections (biological replicates). Data can be displayed as gDNA MRE^{mut}/MRE^{WT} ratio [averaged $2^{(-B)}/2^{(-A)}$ from 3 biological replicates] and cDNA MRE^{mut}/MRE^{WT} ratio [averaged $2^{(-D)}/2^{(-C)}$ from 3 biological replicates]. Alternatively, an MRE score can be calculated by dividing cDNA MRE^{mut}/MRE^{WT} /gDNA MRE^{mut}/MRE^{WT} ratio $[(2^{(-D)}/2^{(-C)})]/[(2^{(-B)}/2^{(-A)})]$. An MRE score >1 suggests functional miRNA–target regulation. A t -test can be carried out by comparing the gDNA MRE^{mut}/MRE^{WT} ratios between biological replicates to the cDNA MRE^{mut}/MRE^{WT} ratio, to determine whether the MRE^{mut} is significantly enriched in the cDNA samples.

The optimal Cycle of Threshold (Ct) number is around 18–30. Insufficient HDR integration (high Ct number [>35] in the qPCR) may be a consequence of low transfection efficiency, or unsuccessful sgRNA-mediated DNA cleavage at the target locus. We suggest validating sgRNA cleavage by high-resolution melt analysis (HRMA) and employing a puromycin selection strategy to enrich for HDR events (*see* **Notes 7 and 8**).

4 Notes

1. To validate the dependence of the obtained MRE score on the unique deletion of the MRE seed sequence of interest, a parallel experiment can be carried out using swapped barcodes. In this instance, the tag^{mut} will replace the MRE with the T3 barcode, while the tag^{WT} will insert the T7 barcode downstream of the MRE seed sequence. A similar T7/T3 and T3/T7 ratio will provide high confidence in the analysis. If a difference between these ratios is observed, it suggests the generation of ectopic *cis*-regulatory elements (i.e., another MRE) by one of the barcodes, positively or negatively influencing RNA transcript stability. In this case, we recommend generation of alternative custom barcode sequences. These should be pre-screened computationally for the presence of ectopic MREs in a simulated post-integration genomic locus, using the TargetScan or PITA algorithms.
2. To increase the stability of ssDNA HDR templates and prevent them from degradation by cellular exonucleases, oligos can be modified with 2-phosphorothioate bonds between the first two and last two DNA bases. 2-phosphorothioate can be ordered as a standard oligonucleotide modification from IDT (<https://www.idtdna.com/site>). Furthermore, in order to prevent Cas9-mediated cutting of the integrated barcodes, it is imperative to disrupt the CRISPR target site in the ssODN HDR

template. Depending on the design and the targeted locus, this can be achieved either by interrupting or deleting part of the CRISPR target site due to the insertion of the T3/T7 barcodes, or when this is not the case, by mutating the PAM sequence.

3. When attempting to assess MRE functionality in a difficult to transfect cell-line it may be helpful to enrich for transfected cells using a selectable marker. In Subheading 3.3, *pX330* can be replaced by *pX458* or *pX459* which co-express GFP and puromycin resistance, respectively. In Subheading 3.4, enrich for edited cells by sorting on GFP (*pX458*) or using a brief 48-h puromycin selection (*pX459*). Experimental conditions should be optimized for each cell line.
4. To assess successful cloning of the sgRNA spacer sequence into the destination vector, set-up a control ligation reaction in which the insert (annealed sgRNA oligo duplex) is replaced with an equivalent volume of H₂O. If the cloning reaction was successful, a very small number of colonies if any will be observed on the control LB-Amp plates compared to the sgRNA ligation plates. If a large number of colonies are observed on the control plates, the cloning reaction likely failed, typically due to incomplete vector digestion or dephosphorylation.
5. To avoid RNA degradation and obtain maximum yield, add Qiagen RLT lysis buffer directly onto the frozen cell pellets (keep on dry ice until adding the lysis buffer). Thawing cell pellets on ice is sufficient to degrade the RNA in as little as 10 min, due to cell lysis upon freeze-thawing. Elute RNA from Qiagen RNeasy column (included in the Qiagen AllPrep DNA/RNA Mini Kit) with 30–50 µl nuclease free H₂O. Pass eluted fraction through the column twice to recover all RNA. For maximum DNA recovery from Qiagen AllPrep DNA/RNA Mini columns, elute with 90 °C nuclease free H₂O or Qiagen elution buffer (EB) and pass the eluate through the Qiagen DNA column twice.
6. In most cases, the DNase step integrated in the QuantiTect Reverse Transcription Kit (Qiagen) is sufficient to remove most contaminating genomic DNA from RNA samples. Using this kit, the difference between RT samples and NO-RT controls generally exceeds 20 cycles. However, if DNA contamination is observed or a different RT kit is used, a TURBO DNase (Thermo Fisher Scientific) step can be included before the RT reaction. In our hands, this reagent allows extremely effective removal of genomic DNA contaminations.
7. Successful qPCR reactions rely on amplification of unique PCR products, devoid of any primer dimers. Therefore, it is crucial to obtain a reliable PCR primer pair. The PCR reaction

should not generate any unspecific products or primer dimers. We recommend using the *Primer 3* website (<http://primer3.ut.cc>) to design three pairs of PCR primers. Use very stringent conditions: optimal melting temperature [60–70 °C]; minimal primer pair melting temperature difference [<1 °C]; optimal size of PCR product [100–300 bp]; no unspecific targeting throughout the genome as defined by BLASTn. Test all primer pair combinations to find optimal conditions (i.e., single band at expected size). Experimentally validate the optimal annealing temperature for every primer pair/template combination. The following workflow can be used as a general guideline for this purpose. Set-up a large PCR reaction volume, and distribute into individual tubes/PCR stripe (10 µl/tube). Run a temperature gradient according to the thermal cycler manufacturer's instructions. For example, a gradient from 53 to 63 °C will generate eight annealing temperatures (50 °C, 50.9 °C, 52.5 °C, 55 °C, 57.9 °C, 60.4 °C, 62 °C, and 63 °C). Determine optimal PCR conditions by agarose gel electrophoresis and/or melt curve analysis.

8. Erroneous NHEJ causes the accumulation of short insertions and deletions (indels) at the targeted genomic locus. The frequency of such indels can be used as an indirect proxy for Cas9-mediated cleavage activity. This metric is potentially an important determinant of HDR efficiency and can be qualitatively assessed by High-Resolution-Melt-Analysis (HRMA) in transiently transfected cells [22]. HRMA is a sensitive assay capable of detecting single nucleotide mutations/deletions (under optimal conditions). Consequently, any technical variation can cause deviations in melt curves, rendering potential false positive or false negative results. Therefore, we recommend always carrying out this assay in biological and technical triplicates (three individual transfections, DNA extractions, PCR and HRMA reactions). The following protocol is optimized for HRMA indel detection in HEK-293T cells, using the C1000 Touch™ Thermal Cycler (Bio-Rad).
 - (a) Culture HEK-293T and transfect with 1 µg *pX330-sgRNA* (from Subheading 3.2) as described in Subheading 3.4.2.
 - (b) Resuspend cell pellets in 100 µl DNA extraction buffer with 200 µg/ml proteinase K. Incubate for 2 h at 55 °C. Centrifuge at $13,000 \times g$ for 3 min. The supernatant containing genomic DNA can be directly used as template for PCR.
 - (c) Design a pair of PCR primers that flank the sgRNA target site. The amplicon size should range between 150 and 200nt. In general, smaller amplicons offer increased sensitivity of indel detection by HRMA. Select PCR primers as described in **Note 6**.

- (d) Setup the following reaction in a PCR plate (384 well). It is useful to perform this reaction in duplicate or triplicate for each transfected sample.

Genomic DNA	1 μ l
Fwd. primer (10 μ M)	0.5 μ l
Rev. primer (10 μ M)	0.5 μ l
2xHotshot Diamond PCR mix	5 μ l
10xLC Green	1 μ l
H ₂ O	2 μ l
Total	19 μ l

- (e) Run the following PCR program in the C1000 Touch™ Thermal Cycler (Bio-Rad).

Temperature	Time	Cycles
95 °C	5 min	1
95 °C	20 s	45
60 °C	20 s	
72 °C	30 s	
95 °C	30 s	1
25 °C	30 s	1
10 °C	hold	

- (f) Run the following program to generate heteroduplexes using the Precision Melt Analysis Software interface.

Temperature	Time
95 °C	30 s
25 °C	30 s
95 °C	30 s
65–95 °C (plate read)	0.1 °C increment 2 s/step

- (g) Analyze the resulting melt curves using the BioRad Precision Melt Analysis software. Briefly, set gates above and below the melting temperature of the amplicon to generate a normalized melt curve and normalized difference curve. To correct for well-to-well temperature variations, temperature-shifted curves can be generated. Control samples (un-transfected) should produce reproducible melt profiles. Indel generation, and coincidentally sgRNA efficiency, can be estimate by the presence of melt-curves that are qualitatively distinct (shifted) from the un-transfected control curves.

Acknowledgments

We thank Andrew Basset for critical comments on the manuscript and the Fulga lab for feedback and advice. Y.S.M is supported by the Clarendon Scholarship; WIMM Prize Fellowship; Christopher Welch Scholarship. Q.W is supported by the Medical Research Council (WIMM Strategic Award, MRC #G0902418 to T.A.F.). T.A.F. is supported by the Medical Research Council (WIMM Strategic Award, MRC #G0902418); Biotechnology and Biological Sciences Research Council (Project Grants #BB/L010275/1 and #BB/N006550/1); Wellcome Trust ISSF Award (#105605/Z/14/Z).

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297
2. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
3. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20
4. Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB (2014) Common features of microRNA target prediction tools. *Front Genet* 5:23
5. Li Y, Zhang Z (2015) Computational biology in microRNA. *Wiley Interdiscip Rev RNA* 6:435–452
6. Oulas A, Karathanasis N, Louloupis A, Pavlopoulos GA, Poirazi P, Kalantidis K, Iliopoulos I (2015) Prediction of miRNA targets. *Methods Mol Biol* 1269:207–229
7. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19:92–105
8. Lim LP, Lau NC, Garrett-Engle P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433:769–773
9. Selbach M, Schwanhauser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature* 455:58–63
10. Xu G, Fewell C, Taylor C, Deng N, Hedges D, Wang X, Zhang K, Lacey M, Zhang H, Yin Q, Cameron J, Lin Z, Zhu D, Flemington EK (2010) Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. *RNA* 16:1610–1622
11. Mansfield JH, Harfe BD, Nissen R, Obenaus J, Srineel J, Chaudhuri A, Farzan-Kashani R, Zuker M, Pasquinelli AE, Ruvkun G, Sharp PA, Tabin CJ, McManus MT (2004) MicroRNA-responsive ‘sensor’ transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat Genet* 36:1079–1083
12. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479–486
13. Hafner M, Landthaler M, Burger L, Khorshid M, Haussler J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141:129–141
14. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153:654–665
15. Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* doi: [10.7554/eLife.05005](https://doi.org/10.7554/eLife.05005)
16. Bassett AR, Azzam G, Wheatley L, Tibbit C, Rajakumar T, McGowan S, Stanger N, Ewels PA, Taylor S, Ponting CP, Liu JL, Sauka-Spengler T, Fulga TA (2014) Understanding functional miRNA-target interactions in vivo by site-specific genome engineering. *Nat Commun* 5:4640
17. Sternberg SH, Doudna JA (2015) Expanding the biologist’s toolkit with CRISPR-Cas9. *Mol Cell* 58:568–574
18. Maruyama T, Dougan SK, Truttmann MC, Bilate AM, Ingram JR, Ploegh HL (2015)

- Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat Biotechnol* 33:538–542
19. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, Kuhn R (2015) Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol* 33:543–548
 20. Pyzocha NK, Ran FA, Hsu PD, Zhang F (2014) RNA-guided genome editing of mammalian cells. *Methods Mol Biol* 1114:269–277
 21. Aricescu AR, Lu W, Jones EY (2006) A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr D Biol Crystallogr* 62: 1243–1250
 22. Fraley SI, Hardick J, Jo Masek B, Athamanolap P, Rothman RE, Gaydos CA, Carroll KC, Wakefield T, Wang TH, Yang S (2013) Universal digital high-resolution melt: a novel approach to broad-based profiling of heterogeneous biological samples. *Nucleic Acids Res* 41:e175

Cell-Free Urinary MicroRNAs Expression in Small-Scale Experiments

Ludek Zavesky, Eva Jandakova, Radovan Turyna, Daniela Duskova, Lucie Langmeierova, Vit Weinberger, Lubos Minar, Ales Horinek, and Milada Kohoutova

Abstract

Cell-free microRNAs (miRNAs) have become one of the novel promising diagnostic and prognostic biomarkers for various diseases recently. Blood serum and plasma along with urine are the most common sources of clinically well, almost noninvasively available samples containing various types of miRNAs. Here, we present a protocol for a small-scale study investigating expression of several candidate miRNAs. Small-scale experiments may be worth investigating in cases where no information is available on miRNAs expression in particular diseases, for validation of previously published miRNAs with promising diagnostic potential, particularly in situations where follow-up study is aimed at validating miRNAs coming from array or NGS experiments, or where funding for these large-scale experiments is not available.

Using urine miRNAs expression as the novel diagnostic tools is challenging and currently this approach is still in its infancy. Therefore, various methods may result in different conclusions depending on clinical sample sets and differences among methods used for the miRNAs isolation and quantitation. In this protocol, we present the method evaluated in the study focused on cell-free urinary miRNAs in ovarian and endometrial cancers. We recommend using stabilization tubes for the urine collection, as this step may be necessary to stop activity of RNases. Further, routine real-time PCR methods are described. We demonstrate that assessment of urinary miRNAs expression may reveal as a feasible method to explore the potential for finding novel diagnostic and prognostic markers.

Key words Urine, MicroRNA, Diagnostics, Real-time PCR, Ovarian cancer, Endometrial cancer

1 Introduction

Diagnostic methods remain insufficient for many diseases. This may be particularly due to lacking markers or techniques, or their compromised sensitivity and specificity. In the last two decades, microRNAs (miRNAs) have appeared as the large group of small,

The original version of this chapter was revised. An erratum to this chapter can be found at DOI [10.1007/978-1-4939-6866-4_21](https://doi.org/10.1007/978-1-4939-6866-4_21)

non-protein-coding RNAs with potent regulatory roles in key cellular processes. They function as posttranscriptional regulators of gene expression targeting many mRNAs and affecting eventually the particular protein expression. However, miRNAs have been shown to occur not only within the cells but also in various body fluids and their extracellular fractions [1], suggesting their potential as novel diagnostic and prognostic markers.

Urine as a non-invasively available source of body fluid miRNAs has attracted the attention of diagnostics-related research in various diagnoses. Recently, several pilot studies have explored this promising potential for example in urological diagnoses [2], particularly in bladder cancer [3], and further for example in breast cancer [4], or in prostate cancer [5]. In ovarian cancer, two reports have been published, ovarian and endometrial cancers being examined in our study [6], and serous ovarian cancer in the study of Zhou et al. [7].

Urine is a complex body fluid. Within the urinary miRNAs, there may be those miRNAs coming from blood filtration in kidneys, but also miRNAs originating in cells of urogenital system. In particular diagnoses associated with urological diseases (e.g., in bladder cancer), the tumor cells may contribute to the miRNAs appearing in stored urine. In other cancer diagnoses, it is suggested that the differential expression of urinary miRNAs mirrors the altered expression in blood, with many sources of different origin [8]. Within the urinary miRNome, two main different fractions of miRNAs may be found, i.e., cellular and extracellular, the latter fraction occurring particularly in exosomes-associated miRNAs and protein complex/lipoprotein-associated miRNAs. It should be noted that finding the exact origin of particular urinary miRNAs may be beyond reach of the current research possibilities.

We present here the protocol usable for a small-scale study focused on several candidate urine cell-free miRNAs evaluated in gynaecological cancer study [6]. However, other different protocols may be tested and applied, particularly those performed in a large-scale miRNAs expression manner. Further research attempts to evaluate urinary miRNAs expressions as the novel diagnostic tools are warranted for diseases with lacking relevant diagnostic and screening markers.

2 Materials

1. Benchtop microcentrifuge.
2. Swing-out rotor centrifuge (e.g., Eppendorf 5804).
3. Centrifuge Eppendorf 5430.
4. Fume hood.
5. Decontamination box.
6. Urine Collection and Preservation Tubes (15 ml/50 ml).

7. Urine Total RNA Purification Maxi Kit (Slurry Format).
8. TaqMan microRNA Reverse transcription kit.
9. MicroAmp Optical 96-Well Reaction Plate. MicroAmp Optical Adhesive Film.

3 Methods

3.1 Urine Collection

1. Collect and process urine samples in the same manner within the whole study.
2. Use stabilization tubes or other stabilization chemicals for the urine collection to stop RNases and bacterial activity in urine. Otherwise, urine processing should be done as rapid as possible (*see Note 1*).
3. Collect urine samples within the same day-time range, ideally as the fresh void second morning urine. Avoid using the first morning urine [9].
4. Pre-surgery/diagnostic urine samples have to be collected before the surgery. Consider all medication applied to patients in the course of sampling may affect the results.
5. Ensure the same gender/menopausal status for the appropriate healthy controls in relation to patients.

3.2 Urine Processing

1. Urine stabilized in Norgen Biotek Collection and Preservation tubes may be stored at room temperature for long time. However, we recommend processing the samples within 30 days after delivery to laboratory. In case of other collection techniques, particularly without stabilization reagents, an immediate processing (centrifugations at 4 °C and RNA isolation/storage at –80 °C) should be done (*see Note 2*).
2. Eliminate the potential precipitates and cell/bacterial debris contamination in urine by centrifugation steps. Centrifuge the tubes at $300 \times g$ for 10 min, and then centrifuge the supernatant at $2,000 \times g$ for 10 min at RT/4 °C. Decant cell-free urine into new tubes (e.g., 2 ml aliquots). In case of stronger centrifugations, do not use higher speed than $2,500 \times g$ (some miRNAs fractions may be lost).
3. Process the cell-free urine according to RNA isolation protocol, or store the aliquoted urine at –80 °C until further processing. Do not remove potential precipitates in the latter case after thawing as they may contain miRNAs bound with protein!

3.3 Isolation of Total RNA

1. Use frozen urine aliquots (thaw them at RT), stabilized urine samples, or freshly void urine samples for RNA isolation (*see Note 3*). Use Urine Total RNA Purification Maxi Kit (Slurry Format) for RNA isolation as follows. Alternatively, use another isolation kit for urinary total RNA isolations ensuring

involvement of microRNA fraction according to manufacturer instructions. Note that for cell-free urine fractions, centrifugation steps have to be included in the urine processing.

2. The minimum recommended urine input is 5 ml and the maximum recommended urine input is 10 ml, according to manufacturer. However, 2 ml urine inputs may work as well. The following volumes should be applied for processing 5 ml urine samples and could be changed accordingly to volumes processed. If available, isolate total RNA from the same volume of urine.
3. Add 10 μ l of β -mercaptoethanol to each 1 ml of Lysis Buffer A required.
4. Mix 5 ml of urine sample with 0.35 ml of Slurry C3 and 4.65 ml of Lysis Buffer A in a 50 ml conical tube and vortex for 15 s.
5. Add 5 ml of 96–100% ethanol to the lysate and vortex for 10 s.
6. Centrifuge for 5 min at $2,000 \times g$ and discard the supernatant.
7. Add 500 μ l of Wash Solution A, mix well by pipetting or vortexing and then transfer the entire contents into a Mini Filter Spin column. Centrifuge for 1 min at $14,000 \times g$. Discard the flowthrough.
8. There may be applied an optional On-Column DNA Removal using Norgen's RNase-Free DNase I Kit in case of high DNA contamination; however, this step may have detrimental effect on the resulting RNA concentration and we do not recommend it.
9. Repeat the washing procedure according to **step 7** two times more.
10. Apply 500 μ l of 96–100% ethanol, centrifuge for 1 min at $14,000 \times g$, and discard the flowthrough.
11. Spin the column, empty, for 3 min at $14,000 \times g$ and incubate the column with the lid open, at 60 °C for 3 min.
12. Transfer the spin column to a fresh 1.7 ml Elution tube. Apply 50–100 μ l (e.g., 60 μ l) of Elution Solution A to the column and centrifuge for 2 min at $200 \times g$, and then for 3 min at $14,000 \times g$.
13. Mix the RNA by pipetting, prepare aliquots and store them at -80 °C ready for downstream applications (*see Note 4*).

3.4 Quality and Quantity Control of Total RNA

Currently, there is no generally applicable test for microRNA quality and quantity. In case of urine samples, low RNA concentration should be expected. Check the quality and quantity of RNA using sensitive applications, such as Agilent Pico Chip or Small RNA Chip. Note, however, that there may be not any association between the

RNA concentration (even in case of Agilent Small RNA kit and “miRNA fraction,” consisting of mRNA degradation products in addition to miRNAs) and the performance in real-time PCR. Alternatively, check the quality of microRNA fraction by running the real-time PCR using particular miRNA(s) (e.g., miR-1228 might be highly abundant in urine) before starting the experiment.

3.5 Reverse Transcription

1. Perform reverse transcription of total RNA to cDNA using TaqMan microRNA Reverse transcription kit and miRNA-specific stem-loop primers. Instead of full-scale reactions, scaled-down reactions up to one-third or one-half of the recommended original input volumes of all reagents and RNA may be used. Additionally, increased amount of RNA may be applied replacing the ddH₂O volume, in order to improve resulting Ct values in low RNA-concentration samples. Use the same input RNA volume for samples in the experiment.
2. Prepare the reaction mix on ice. In scaled-down reactions (7.5 µl), each reaction should comprise 2.08 µl nuclease-free water (ddH₂O), 0.75 µl 10× Reverse Transcription Buffer, 0.095 µl RNase-Inhibitor (20 U/µl), 0.075 µl 100 mM dNTPs with dTTP, and 0.5 µl Multiscribe Reverse Transcriptase (50 U/µl), prepared in large volume as a master mix. For each miRNA-specific master mix add 1.5 µl of RT primer (5×) and finally add 2.5 µl of specific RNA sample for each reaction. Mix gently by inversion (do not vortex) and centrifuge briefly.
3. In thermal cycler, perform the reverse transcription reaction in 0.2 ml PCR tubes with following conditions: 16 °C for 30 min, 42 °C for 30 min, 85 °C for 5 min, hold at 4 °C. The RT product may be stored at −20 °C at least for 1 week prior to the real-time PCR.

3.6 Real-Time PCR and Expression Analyses

1. In addition to pathological and control samples included within the plate, no-template controls (only master mix with specific primer without cDNA), no-RT controls (total RNA instead of cDNA) and inter-plate calibrators for each miRNA assay (sample with good expression) may be applied in each plate to observe potential contamination, or differences between the plates.
2. Perform real-time PCR reactions in duplicates or triplicates, using scaled-down (10 µl) reactions. Prepare a reaction pre-mix consisting of TaqMan Universal PCR Master Mix (2×) (5 µl), nuclease-free water (3.8 µl), and miRNA-specific small RNA assay (20×) (0.5 µl) (volumes are per one reaction/sample), including excess volumes (15%) for pipetting losses.
3. Aliquot reaction pre-mix in tubes for duplicate/triplicate reactions (i.e., 21.4 µl), add cDNA product (1.61 µl) from the reverse transcription for each sample by gentle pipetting. Centrifuge briefly.

4. Aliquot 10 μ l of pre-mix with cDNA in duplicates/triplicates for each sample into the MicroAmp Optical 96-Well Reaction Plate and seal it with MicroAmp Optical Adhesive Film. Centrifuge plate at 2,600 rpm using appropriate centrifuge and rotor (e.g., Eppendorf 5430) for 1.5 min.
5. Run real-time PCR reactions using 7900HT Fast Real-Time PCR System (Life Technologies). In SDS 2.4 software, select $\Delta\Delta$ Ct (RQ) as the “assay”, and 96 Wells Clear Plate as “Container.” Then set the particular miRNAs as targets and endogenous controls (when available), and following real-time PCR conditions: Hold 50 °C for 2 min, hold 95 °C for 10 min, PCR (40 cycles): 95 °C for 15 s, 60 °C for 60 s, saving the experiment. Load the plate into instrument and Start Run.
6. After completing the real-time PCR analyses, save the study and open the dataset in the RQ Manager 1.2.1. Prepare “New Study from SDS file” by loading the original sds file. Save the study. In Analysis setting, set the Threshold manually to be the same within the whole experiment (e.g., 0.2). In the Plate Centric Table orientation, perform the analysis, save the study and export the data as “clip data,” to have the raw amplification data for each cycle and sample (further used in Real-time PCR Miner). Alternatively, save the data as results data to have the Ct values as determined by RQ manager to be processed in other applications.
7. Open the exported data file in Excel, select raw Rn data (*not* delta Rn), copy and “paste special” as transposed data to another sheet. In the first column (designated as Cycle), there should be 1–40 cycle numbers. Within the first row, there are sample names (miR-200b_A1 etc.).
8. Select, copy and export these data to Real-time PCR Miner (online application at <http://ewindup.info/miner/index.htm>, [10]) and fill in appropriate data fields requested. Save the results of analysis primarily as txt. file, and then open the file in Excel.

There are several kinds of data coming from the PCR Miner (these are, along with Ct, for example Average Efficiency of Genes, Stdev of Genes Efficiency Mean CV% of Genes Efficiencies, Mean CV% of Genes Cts) etc. Check them carefully. To use efficiency-corrected data, calculate the relative expressions as follows.

3.7 Calculating Relative miRNA Expression Using Real-Time PCR Miner Data

1. Use average efficiency (E) for each gene analyzed instead of using individual sample efficiency data.
2. Calculate $(1+E)^{CT}$.
3. Calculate relative expression R0 as relative R0: $1/(1+E)^{CT}$.
4. Normalize the relative expression R0 for the sample and particular miRNA by dividing the R0 value by R0 value of

the endogenous control or the geometric mean of the R0 expression of all miRNA targets determined in the experiment. In the latter case, note that geometric mean normalization is suitable particularly for a wider selection of miRNAs coming from different chromosomes and regions; at least five miRNAs (ideally at least ten miRNAs) should be used. As an alternative normalization approach, the percentage proportions of particular miRNAs within all the tested miRNAs R0 (total) expression (*not* normalized) in the experiment may be used. Such normalized expression reflects the proportion relations between the miRNAs [11]. As further alternative, when no endogenous control is available in small-scale experiments, ratio proportions of each miRNA to one another may be applied and tested.

5. Perform statistical analyses using normalized and log-transformed data for pathological and control samples to evaluate the differences between the two groups.
6. Calculate fold-differences between pathological and control samples by dividing the normalized R0 value of the pathological samples by the mean normalized R0 value of the control samples (*see* **Notes 5** and **6**).

4 Notes

1. Stabilization of urine samples within collection tubes is highly recommended to stop RNases and bacterial activity in urine.
2. Precipitation may occur in stored urine samples. Avoid removing the precipitates occurring in cell-free samples obtained after centrifugations steps.
3. Working with RNA requires using RNase-free laboratory equipment and consumables. Avoid all potential sources of sample contamination.
4. Avoid frequent thaw–freeze cycles of stored RNA. Store RNA at -80°C .
5. Carefully selected miRNAs with sufficient abundance/expression may be necessary for relevant real-time PCR analyses. Efficiency-corrected expression data may be advisable to be used in analyses. Validation/verification of results by alternative algorithms and normalization procedures should improve relevance of conclusions.
6. Urine as the complex body fluid consists of various miRNA fractions of different origin. Cell-free urine includes exosomal miRNAs, which may be isolated separately by various methods.

Acknowledgments

The financial support from the Charles University Prague (projects PRVOUK-P27/LF1/1, Progres Q28/LF1 and PRVOUK-P25/LF1/2) and from the Ministry of Health of the Czech Republic (project FNBr 65269705) is appreciated. We would like to thank Avast Foundation, Feron, a.s., ČEPS, a.s. and CEZ Foundation for their kind support of our research.

References

1. Weber JA, Baxter DH, Zhang SL et al (2010) The microRNA spectrum in 12 body fluids. *Clin Chem* 56:1733–1741
2. Mlcochova H, Hezova R, Meli AC, Slaby O (2015) Urinary microRNAs as a new class of noninvasive biomarkers in oncology, nephrology, and cardiology. *Methods Mol Biol* 1218:439–463
3. Chen L, Cui Z, Liu Y et al (2015) MicroRNAs as biomarkers for the diagnostics of bladder cancer: a meta-analysis. *Clin Lab* 61:1101–1108
4. Erbes T, Hirschfeld M, Rucker G et al (2015) Feasibility of urinary microRNA detection in breast cancer patients and its potential as an innovative non-invasive biomarker. *BMC Cancer* 15:193
5. Korzeniewski N, Tosev G, Pahernik S et al (2015) Identification of cell-free microRNAs in the urine of patients with prostate cancer. *Urol Oncol* 33:16.e17–16.e22
6. Zavesky L, Jandakova E, Turyna R et al (2015) Evaluation of cell-free urine microRNAs expression for the use in diagnosis of ovarian and endometrial cancers. A pilot study. *Pathol Oncol Res* 21:1027–1035
7. Zhou J, Gong G, Tan H et al (2015) Urinary microRNA-30a-5p is a potential biomarker for ovarian serous adenocarcinoma. *Oncol Rep* 33:2915–2923
8. Zavesky L, Jandakova E, Turyna R et al (2015) New perspectives in diagnosis of gynaecological cancers: Emerging role of circulating microRNAs as novel biomarkers. *Neoplasma* 62:509–520
9. Witwer KW, Buza EI, Bemis LT et al (2013) Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *J Extracell Vesicles* 2:20360
10. Zhao S, Fernald RD (2005) Comprehensive algorithm for quantitative real-time polymerase chain reaction. *J Comput Biol* 12:1047–1064
11. Zavesky L, Jandakova E, Turyna R et al (2016) Supernatant versus exosomal urinary microRNAs. Two fractions with different outcomes in gynaecological cancers. *Neoplasma* 63:121–132

Peptide-Based Isolation of Argonaute Protein Complexes Using Ago-APP

Judith Hauptmann and Gunter Meister

Abstract

Argonaute (Ago) proteins bind small RNAs such as microRNAs (miRNAs) or short interfering RNAs (siRNAs), which guide them to distinct mRNAs for post-transcriptional gene silencing. Mammalian miRNA-guided gene silencing pathways mainly lead to translational repression and mRNA destabilization. To facilitate these processes, Ago proteins bind members of the GW protein family, which form central interaction platforms for the recruitment of downstream effector proteins. GW proteins use tryptophane residues (W) to bind to the surface of Ago proteins. This high affinity interaction is retained when a short, GST-fused GW peptide is used in biochemical pull-down experiments—an approach referred to as “Ago Affinity Purification by Peptides” (Ago-APP). Since the binding interface is conserved among different paralogues and different species, Ago-APP represents a universal tool to purify Ago proteins and associated small RNAs using samples from species with conserved miRNA pathways.

Key words Argonaute proteins, Small RNAs, miRNAs, Ago-APP

1 Introduction

Ago proteins are central components of small RNA-guided gene silencing pathways. They bind to small RNAs and regulate target mRNAs via different mechanisms depending on the degree of complementarity between the small RNA and the target mRNA [1, 2]. Ago proteins are characterized by an N-terminal domain, which contributes to miRNA loading, a PAZ domain that binds the 3' end of the miRNA, a MID domain that anchors the 5' end of the miRNA and a PIWI domain, which is similar to RNase H and thus can function as endonuclease [3, 4]. Some Ago proteins are endonucleases and cleave their targets in case of full complementarity [5, 6]. This slicer function depends on the presence of four catalytic residues in the PIWI domain [7] and is additionally influenced by structural elements in the PIWI and N-terminal domains [7–11].

Most mammalian miRNAs bind targets only with partial complementarity and thus such targets are not directly cleaved. Instead, repression involves mechanisms of exonucleolytic mRNA destabilization and translational repression [12, 13]. To achieve silencing, Ago proteins directly interact with GW proteins (also known as GW182 proteins), which are termed TNRC6A, -B, and -C (trinucleotide repeat-containing protein 6 A, B, and C) in mammals [14–17]. These proteins form interaction platforms for additional protein factors such as the poly(A) binding protein and deadenylase complexes leading to translational inhibition, mRNA deadenylation, decapping, and subsequent decay [12]. GW proteins use their C-terminal part to interact with the CCR4-NOT complex, the major deadenylase complex in mammals [18, 19]. It has been demonstrated recently that the major contact between CCR4-NOT is facilitated by the subunit CNOT9. Furthermore, CNOT9 forms a complex with DDX6, which is required for translational repression and mRNA decay [20, 21]. In addition, the deadenylase complex PAN2-PAN3 can be recruited as well [22].

GW proteins are largely unstructured and contain multiple glycine/tryptophan (GW) repeats. Most of these motifs are present in the N-terminal half where two tryptophans are essential for Ago binding [23–25]. The crystal structures of human Ago2 and Ago1 revealed that these residues insert into two specific binding pockets at the protein surface [7, 24]. Whereas the sequence context of Ago-interacting tryptophans is hardly defined, their spacing seems to be critical [23]. Importantly, the human Ago proteins show high sequence similarity and all four paralogues interact with GW proteins [15, 25, 26]. Moreover, members of the GW protein family can be found in all organisms with functional miRNA silencing pathways suggesting that also the tryptophan-mediated interaction is conserved.

Here, we describe a novel purification strategy for Ago proteins. The approach is based on the Tryptophan-mediated interaction between Ago and GW proteins. In particular, a short, GST-tagged peptide derived from the human GW protein TNRC6B is purified recombinantly, immobilized to an adequate affinity matrix, and used to enrich for Ago proteins [27]. The conserved interaction of Ago and GW proteins facilitates antibody-independent Ago purifications from all species with functional miRNA silencing machineries. Thus, the approach enables cross-paralogue and species-independent analyses of endogenous Ago proteins. Moreover, Ago-APP precipitates Agos that are loaded with miRNAs and thereby facilitates the detailed analysis of small RNAs involved in gene silencing.

2 Materials

2.1 Expression and Purification of the APP Peptide

1. APP peptide expression constructs: TNRC6B 599-683 in pGEX-6P-1 or pEC-K-3C, tagged with N-terminal FLAG-GST or 6xHis-GST.
2. LB medium: 10 g/l NaCl, 5 g/l yeast extract, 10 g/l tryptone (pH 7.2–7.5).
3. Expression strain *E. coli* BL21.
4. Isopropyl β -D-1-thiogalactopyranoside (IPTG).
5. Phosphate-buffered saline (PBS): 137 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, 1.8 mM KH₂PO₄ (pH 7.5).
6. Lysozyme.
7. Dithiothreitol.
8. GST wash buffer: PBS with protease inhibitors.
9. Glutathione elution buffer: PBS containing 20 mM Tris-HCl pH 8.0 and 10 mM Glutathione.
10. Sonicator.
11. PALL filters.
12. ÄKTA purification system.
13. Nanodrop or spectrophotometer.

2.2 Cell Culture and Lysis

1. Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum and 1 $\times$ Penicillin/Streptomycin.
2. PBS.
3. NET buffer: 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 10% glycerol, 1 mM NaF; supplemented with 0.5 mM DTT and 1 mM AEBSF before used.
4. FastPrep[®]-24.
5. Lysing Matrix D.

2.3 Ago-APP

1. Glutathione Sepharose 4 Fast Flow.
2. Anti-FLAG M2 affinity gel.
3. PBS.
4. APP peptide: 6xHis-GST-tagged TNRC6B 599-683 or FLAG-GST-tagged TNRC6B 599-683.
5. 5 $\times$ Laemmli sample buffer: 300 mM Tris-HCl (pH 6.8), 10% SDS, 62.5% glycerol, 0.05% bromophenol blue, 10% β -mercaptoethanol.
6. Glutathione elution buffer: 20 mM Glutathione, 50 mM Tris-HCl pH 8, 120 mM NaCl.

7. PreScission Cleavage Buffer: 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM DTT.
8. PreScission Protease.

2.4 miRNA Analysis

1. TRIzol.
2. Isopropanol.
3. Glycogen RNA grade.
4. 70% Ethanol.
5. RNA sample buffer: bromophenol blue and xylene cyanol in formamide.
6. 12% Urea gels.
7. TBE: 89 mM Tris, 89 mM boric acid, 2 mM EDTA.
8. Whatman paper.
9. Amersham Hybond-N membrane.
10. Semidry Blotting chamber.
11. EDC crosslinking solution: 0.16 M EDC (1-ethyl-3-(3-dimethyl-aminopropyl)-carbodiimide) in 0.13 M 1-methylimidazole (pH 8).
12. PNK-labeled probe: use DNA oligos with perfect complementarity to the miRNA of interest. Label them with T4 polynucleotide kinase according to the manufacturer's protocol. Remove free γ - ^{32}P -ATP by Sephadex G25 columns.
13. 20 $\times$ SSC: 3 M NaCl, 0.3 M tri-sodium citrate dihydrate (pH 7).
14. Hybridization solution: 1 $\times$ SSC, 20 mM Na_2HPO_4 (pH 7.2), 7% SDS, 0.02% BSA, 0.02% Polyvinylpyrrolidone K30, 0.02% Ficoll 400.
15. Washing solution 1: 5 $\times$ SSC, 1% SDS.
16. Washing solution 2: 1 $\times$ SSC, 1% SDS.
17. Phosphoimaging system.

3 Methods

3.1 Peptide Expression and Purification (See Notes 1 and 2)

1. Prepare an overnight culture of the TNRC6 peptide constructs in BL21.
2. Inoculate 1–4 L of LB medium (containing the respective antibiotics) with the overnight culture (1:100).
3. Grow the culture at 37 °C until an optical density (OD) of 0.4 is reached. Continue growing at 18 °C.
4. At an OD of 0.6–0.8, induce protein expression by the addition of 1 mM IPTG and grow overnight at 18 °C.

5. Harvest the culture by 15 min centrifugation at $4000 \times g$. Discard the supernatant. Per 1 L of culture, resuspend the bacteria pellet in 12–25 ml cold PBS containing protease inhibitors, 1 mM dithiothreitol (DTT) and 1 mg/ml lysozyme.
6. Sonicate the resuspended cells, clear the lysate by centrifugation for 30 min at $40,000 \times g$, and filter the lysate to remove particles.
7. Apply the GST-tagged peptide to a glutathione sepharose column. Wash with 10 column volumes (cv) of GST wash buffer.
8. Elute the peptide with 4 cv Glutathione elution buffer.
9. Pool the peptide-containing fractions and remove Glutathione by application to a desalting column, gel filtration, or dialysis (always in PBS).
10. Determine the peptide concentration.
11. Prepare aliquots and store these at -80°C . Repeated thawing and freezing of the peptide is possible to a certain extent, but should be restricted to the minimum.

3.2 Lysate Preparation

3.2.1 Lysate Preparation from Cell Culture Material

1. Remove the cell culture media and add PBS to the cell culture plate.
2. Use a rubber policeman to scrape the cells from the plate. Transfer them to a 15 ml or 50 ml tube.
3. Spin the cells down for 10 min at $200 \times g$ (4°C).
4. If cell culture medium is still visibly present in the supernatant, repeat the washing step with PBS and again centrifuge for 10 min at $200 \times g$.
5. Remove the supernatant and resuspend the cell pellet in an appropriate amount of lysis buffer. As a rule of thumb, use 0.5–1 ml per cell culture plate (15 cm in diameter).
6. Incubate the extract on ice for 10 min.
7. Centrifuge for 10 min at $15,000 \times g$.

3.2.2 Lysate Preparation from Primary Tissues

1. Add 800 μl of lysis buffer to one tube of Lysing Matrix D. Place one piece of tissue (ca. 0.5 cm in diameter) in the tube.
2. Use the FastPrep24 device for tissue disruption. Therefore, set it to 45 s at 6.5 M/s. Spin the tube for 1 min at $13,000 \times g$ (4°C).
3. Use a 200 μl tip to transfer the extract from between the ceramic beads to a 2 ml reaction tube. If large amounts of material are needed, repeat these steps with a second piece of tissue. You may recover more material by washing Lysing Matrix D one more time with lysis buffer. Therefore, add the buffer to the beads, vortex thoroughly, and centrifuge 1 min at $13,000 \times g$.

4. Centrifuge the lysates for 20 min at $15,000 \times g$. Recover the supernatants. You may determine their concentration and keep input samples for Western Blot analysis.

3.3 Peptide Coupling

1. Depending on the used peptide, use Glutathione or anti-FLAG M2 beads. Wash the beads with PBS twice. Therefore, add 1 ml of PBS to the beads, invert the tube several times, and centrifuge for 1 min at $1000 \times g$ (4°C). Carefully remove the supernatant and repeat this step.
2. Per 50 μl beads (50% slurry, i.e., 25 μl packed bead volume) add at least 100 μg of GST- or FLAG-GST-tagged recombinant peptide in a total volume of 1 ml PBS. A total volume of 1 ml can be used for coupling up to 150 μl beads (Fig. 1).
3. Incubate at 4°C on the rotating wheel for at least 1 h, possibly as long as overnight.
4. Wash the coupled beads with PBS three times to remove unbound peptide. For each washing step, invert the tube several times and centrifuge for 1 min at $1000 \times g$ (4°C).

3.4 Ago-APP (See Notes 3–6)

1. Add the lysate to the peptide-coupled beads and incubate on the rotating wheel for 3 h at 4°C .
2. Wash the beads four times with NET buffer and one time with PreScission cleavage buffer (for elution by PreScission protease) or PBS (for any other elution or RNA extraction). Each time, invert the reaction tube well and centrifuge for 1 min at $1000 \times g$ (4°C).

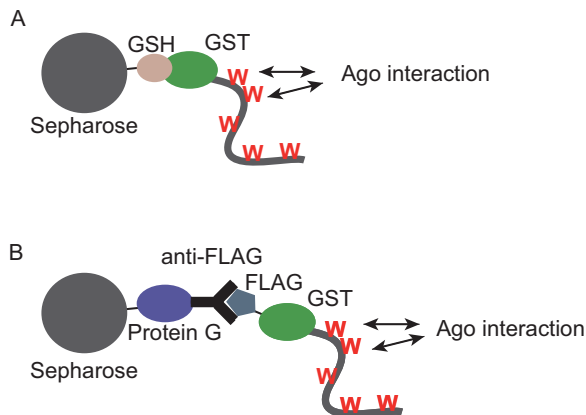


Fig. 1 Schematic representation of Ago-APP using the GST-tagged (a) or FLAG-GST-tagged (b) APP peptide

3. For elution, add at least the packed bead volume of the respective elution buffer. Use one of the following elution strategies:
 - (a) Add 2× SDS sample buffer and incubate at 95 °C for 5 min.
 - (b) Add Glutathione elution buffer and incubate at room temperature for 30 min.
 - (c) Add PreScission cleavage buffer and PreScission and incubate at 4 °C overnight (Fig. 2).

3.5 Protein Analysis

For protein analysis, refer to SDS-PAGE, Western Blot, and Coomassie staining protocols.

3.6 miRNA Analysis by Northern Blot (See Note 7)

1. Perform Ago-APP as described in Subheading 3.4. Use the last washing step to remove a part of the beads (20–25%) for Western Blot analysis.
2. Add 1 ml of Trizol directly to the beads and proceed as stated in the manufacturer's protocol. Precipitate the RNA with Glycogen overnight at –20 °C.
3. Spin the samples for 30 min at 15,000 × *g*. Remove the supernatant and wash the RNA one time with 70% ethanol. Spin for 10 min at 15,000 × *g*. Remove the supernatant and leave the tube open to let the pellet dry for ca. 5 min.
4. Solve the RNA pellet in 10–20 µl water and add 2× RNA sample buffer. Alternatively, the pellet can be directly solved in RNA sample buffer. Incubate for 10 min at 65 °C while shaking at 1200 rpm.

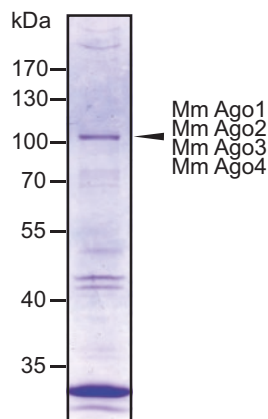


Fig. 2 Purification of mouse Ago1-4 from primary tissue. GST-tagged APP peptide was used with glutathione sepharose beads and mouse spleen extract. The samples were analyzed by SDS-PAGE, Coomassie staining, and mass spectrometric measurements

5. Prepare a 12% Urea gel, let it polymerize for at least 1 h and prerun the gel for 30 min at 400 V.
6. Directly before loading, wash the wells to remove urea. Load the RNA to the bottom of the wells and run for 1–1.5 h at 400 V. miRNAs will run between the two dye fronts.
7. For blotting, soak three Whatman papers with water. Place them on a semidry blotting chamber. Soak a membrane of exactly the same size and place it on the top. Cut the gel to the same size, wash it with water, and place it on the membrane. Soak three more Whatman papers with water and place them on the top of the stack. Carefully remove air.
8. Blot for 30 min at 20 V.
9. Prepare the EDC crosslinking solution just before use. Cut a Whatman paper slightly bigger than the membrane. Soak it with EDC crosslinking solution. Place the membrane on the top and wrap everything in plastic film. Crosslink the blot for 1 h at 50 °C.
10. Rinse the blot with water twice and let it dry.
11. Place the blot in a hybridization bottle and add hybridization solution. Add the labeled probe to the hybridization solution and incubate overnight at 50 °C.
12. Wash the blot two times with washing solution 1 and one time with washing solution 2. For each washing step, incubate at 50 °C for 10 min.
13. Wrap the blot in plastic film and expose it to a phosphorimager screen (Fig. 3).

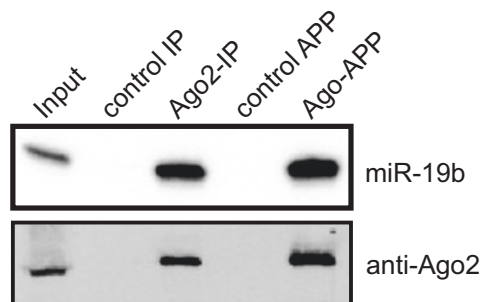


Fig. 3 MiRNA isolation by Ago-APP and anti-Ago2 immunoprecipitation. Anti-Ago2 immunoprecipitation and Ago-APP from HEK 293 cell extracts were analyzed for co-isolated miRNAs. Proteins and RNA samples from both purifications were analyzed by Western blotting (*lower panel*) and Northern blotting (*upper panel*). An unrelated control antibody or the mutated peptide (control APP) served as controls

4 Notes

1. Here, peptide purification is described using a liquid chromatography system. Batch purification can be used as well, although it might result in lower yields.
2. The use of GST- or FLAG-GST-tagged peptide depends on the experimental system that is used. GSH beads can be sensitive to oxidizing reagents such as the phosphatase inhibitor vanadate. Moreover, the use of anti-FLAG beads is well established for applications such as PAR-CLIP.
3. The hydrophobic interaction of TNRC6B and Ago is resistant to high salt concentrations, but only low detergent levels are tolerated. The method has been established with NET buffer, but the use of other buffers can be optimized with respect to the mentioned restrictions.
4. Ago-APP depends on the Tryptophan-mediated binding of Ago and GW proteins. As a control peptide, a version with all Tryptophans mutated to Alanins can be used (TNRC6B W623A W634A W653A W666A W680A). Alternatively, GST alone can be used.
5. Ago-APP has been established using a peptide derived from human TNRC6B. However, this region is well conserved in the human paralogues TNRC6A (699-786) and TNRC6C (429-512) and therefore peptides from these proteins might work as well.
6. Ago-APP can be used to isolate Ago protein complexes from many species. However, if species are evolutionary distant, precipitation efficiency might decrease. This could be circumvented by using adapted GW peptides, e.g., those derived from Ago-interacting regions in *Arabidopsis thaliana* Polymerase V [28] or *Schizosaccharomyces pombe* Tas3 [29].
7. Here, the co-purification of miRNAs has been described. Importantly, target mRNAs can only be co-purified when an additional crosslinking step is used.

Acknowledgments

Our research is supported by grants from the Deutsche Forschungsgemeinschaft (SFB 960, FOR2127), the European Research Council (ERC grant 242792 “sRNAs”, ITN RNATrain), the Bavarian Genome Research Network (BayGene), the German Cancer Aid and the Bavarian Systems-Biology Network (BioSysNet).

References

- Meister G, Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature* 431:343–349
- Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
- Jinek M, Doudna JA (2009) A three-dimensional view of the molecular machinery of RNA interference. *Nature* 457:405–412
- Ipsaro JJ, Joshua-Tor L (2015) From guide to target: molecular insights into eukaryotic RNA-interference machinery. *Nat Struct Mol Biol* 22:20–28
- Meister G et al (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* 15:185–197
- Liu J et al (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305:1437–1441
- Nakanishi K et al (2013) Eukaryote-specific insertion elements control human ARGONAUTE slicer activity. *Cell Rep* 3:1893–1900
- Hauptmann J et al (2013) Turning catalytically inactive human Argonaute proteins into active slicer enzymes. *Nat Struct Mol Biol* 20:814–817
- Hauptmann J et al (2014) Generation of catalytic human Ago4 identifies structural elements important for RNA cleavage. *RNA* 20:1532–1538
- Faehnle CR et al (2013) The making of a slicer: activation of human Argonaute-1. *Cell Rep* 3:1901–1909
- Schurmann N et al (2013) Molecular dissection of human Argonaute proteins by DNA shuffling. *Nat Struct Mol Biol* 20:818–826
- Jonas S, Izaurralde E (2015) Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet* 16:421–433
- Pfaff J, Meister G (2013) Argonaute and GW182 proteins: an effective alliance in gene silencing. *Biochem Soc Trans* 41:855–860
- Jakymiw A et al (2005) Disruption of GW bodies impairs mammalian RNA interference. *Nat Cell Biol* 7:1267–1274
- Meister G et al (2005) Identification of novel argonaute-associated proteins. *Curr Biol* 15:2149–2155
- Liu J et al (2005) A role for the P-body component GW182 in microRNA function. *Nat Cell Biol* 7:1161–1166
- Rehwinkel J et al (2005) A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. *RNA* 11:1640–1647
- Behm-Ansmant I et al (2006) mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev* 20:1885–1898
- Chekulaeva M et al (2011) miRNA repression involves GW182-mediated recruitment of CCR4:NOT through conserved W-containing motifs. *Nat Struct Mol Biol* 18:1218–1226
- Chen Y et al (2014) A DDX6-CNOT1 complex and W-binding pockets in CNOT9 reveal direct links between miRNA target recognition and silencing. *Mol Cell* 54:737–750
- Mathys H et al (2014) Structural and biochemical insights to the role of the CCR4:NOT complex and DDX6 ATPase in microRNA repression. *Mol Cell* 54:751–765
- Braun JE, Huntzinger E, Izaurralde E (2013) The role of GW182 proteins in miRNA-mediated gene silencing. *Adv Exp Med Biol* 768:147–163
- Pfaff J et al (2013) Structural features of Argonaute-GW182 protein interactions. *Proc Natl Acad Sci U S A* 110:E3770–E3779
- Schirle NT, MacRae IJ (2012) The crystal structure of human Argonaute2. *Science* 336:1037–1040
- Baillat D, Shiekhattar R (2009) Functional dissection of the human TNRC6 (GW182-related) family of proteins. *Mol Cell Biol* 29:4144–4155
- Lazaretti D, Tournier I, Izaurralde E (2009) The C-terminal domains of human TNRC6A, TNRC6B, and TNRC6C silence bound transcripts independently of Argonaute proteins. *RNA* 15:1059–1066
- Hauptmann J et al (2015) Biochemical isolation of Argonaute protein complexes by Ago-APP. *Proc Natl Acad Sci U S A* 112:11841–11845
- El-Shami M et al (2007) Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes Dev* 21:2539–2544
- Till S et al (2007) A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nat Struct Mol Biol* 14:897–903

Predicting Functional MicroRNA-mRNA Interactions

Zixing Wang and Yin Liu

Abstract

MicroRNAs (miRNAs) are small RNA molecules that play key regulatory roles in general biological processes and disease pathogenesis. These small RNA molecules interact with their target mRNAs to induce mRNA degradation and/or inhibit the translation of mRNAs into proteins. Therefore, identifying miRNA targets is an essential step to fully understand the regulatory effects of miRNAs. Here, we describe a regularized regression approach that integrates the sequence information with the miRNA and mRNA expression profiles for detecting miRNA targets. This method takes into account the full spectrum of gene sequence features of miRNA targets, including the thermodynamic stability, the accessibility energy, and the context features of the target sites,. Given these sequence features for each putative miRNA-mRNA interaction and their expression values, this model is able to quantify the down-regulation effect of each miRNA on its targets while simultaneously estimating the contribution of each sequence feature for predicting functional miRNA-mRNA interactions.

Key words MiRNA target identification, miRNA and mRNA expression profiles, Regularized regression, Sequence features, Context score, Thermodynamic stability

1 Introduction

MicroRNAs (miRNAs) are small noncoding RNA molecules found in both the invertebrate and vertebrate genomes [1, 2]. These small RNA molecules exert their regulatory effects on target gene mRNAs by inducing mRNA degradation and/or inhibiting protein translation [3, 4]. They represent one of the most abundant classes of gene-regulatory molecules in mammals [5] and there have been more than two thousand distinct miRNAs identified in humans [6]. It is also demonstrated that more than 60% of human mRNAs are subject to miRNA-mediated post-transcriptional regulation [7]. Among the miRNA-mRNA regulatory relationships, multiple miRNAs are often required to act cooperatively to target the same gene. On the other hand, a given miRNA is able to pair with up to hundreds of genes [8]. This multiplicity effect poses a great challenge to comprehensively depict a functional miRNA-mRNA interaction network. Over the past decade, many miRNA

algorithms and tools have been independently developed to study the miRNA-mRNA interactions, including TargetScan [2, 7, 9], PicTar [10], MicroT_CDS [11], miRanda [12], RNAhybrid [13], and TargetS [14, 15]. The majority of these algorithms are based on several well-known rules derived from important sequence features of target recognition, including sequence complementarity (seed matching), thermodynamic stability, target site context, target site accessibility, and the degree of site conservation [2, 7, 9, 16]. In recent years, expression profiling has been proposed as a complementary information resource to study dynamic miRNA-mRNA regulatory relationship. As a result, many approaches tend to integrate the expression information into sequence-based target predictions, including the GenMiR++ [17] and the Lasso-based methods [18, 19]. However, when integrating the sequence-based predictions on miRNA-mRNA interactions with the expression data, these approaches perceive these putative interactions as equally meaningful and neglect the full spectrum of available sequence features of miRNA targets.

Here, we describe a novel regularized regression approach for detecting functional miRNA-mRNA interactions. Our approach takes into account the miRNA and mRNA expression profiles from matched samples while incorporating full spectrum of sequence features of putative targets, including the thermodynamic stability, the target site accessibility, and the target site context. Simultaneously, our method estimates the contribution of each sequence feature in predicting miRNA targets. By combining these sequence features and the expression profiles of miRNAs and mRNAs together, we are able to quantify the down-regulation effect of each miRNA on its putative targets and rank the candidate miRNA-mRNA interactions [20].

2 Data Sources

1. MiRNA sequences: The mature miRNAs sequences can be downloaded from miRBase [6]. There are more than 28,000 reported miRNAs entries, including 2557 entries for humans in the latest version (Release 21, 2014).
2. Gene sequences: The sequences of the Promoters, 5'UTRs, CDSs, and 3'UTRs for each human gene can be downloaded from the UCSC Genomes database using the UCSC Table Browser. When there are multiple sequences available for a single gene (e.g., multiple UCSC IDs correspond to a single gene name), the longest sequence is chosen for subsequent analysis.
3. Expression data: The matched miRNA and mRNA expression data for the same set of samples are required for the method.

In this chapter, the Multi Class Cancer (MCC) dataset is used as an example to demonstrate the method. The dataset is composed of 88 paired normal and tumor samples, where 21 samples are from normal tissues and the rest are obtained from tumor tissues, spanning 11 different tumor types [18]. The compiled expression data consist of 114 human miRNAs and 16,063 mRNAs.

4. Gold standard dataset: MiRTarBase 4.5 is used as the gold standard for method evaluation. It hosts 14,659 experimentally verified miRNA-mRNA interactions between 358 miRNAs and 8400 target genes in humans [21]. These interactions have been validated by low-throughput assays including reporter assay and Western blot, or high-throughput transcriptomic experiments, such as microarray and pSILAC experiments.

3 Methods

3.1 The General Framework for Predicting miRNA-mRNA Interaction (See Note 1)

The framework of the approach is illustrated in Fig. 1. Given the matched miRNA and mRNA expression profiles, a set of constraints will be applied to the dataset for initial feature selection purpose. The constraint used in this approach is the list of putative miRNA-mRNA pairs obtained from sequence-based algorithms such as the TargetScan [7]. The constrained gene expression dataset and the full spectrum of sequence features for each putative miRNA-mRNA pair are then used as the input for the adaptive lasso model to predict functional miRNA-mRNA interactions.

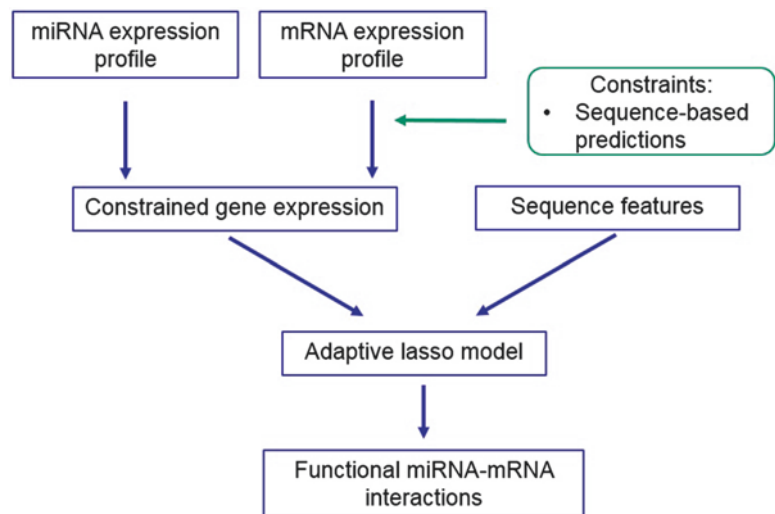


Fig. 1 A framework of inferring functional miRNA-mRNA interactions

3.2 Preprocessing of Sequence Features

Three sequence features are required for the method. (1) Thermodynamic stability, measured by the hybridization energy between miRNA and its candidate target site. For each miRNA-mRNA pair, ΔG_{duplex} is calculated by RNAhybrid [13], using the miRNA sequence and 58 nucleotides flanking the seed match sites in the mRNA sequences. If the mRNA sequence contains multiple seed match sites in different gene regions, the ΔG_{duplex} is computed as the weighted sum of all seed match sites, using different sets of weights deduced from our algorithm [14]; (2) Target site accessibility, which is energetically measured by $\Delta\Delta G$, the free energy required to unpair the nucleotides on the target site to make the target accessible. It is calculated as the difference between ΔG_{duplex} and $\Delta G_{\text{disruption}}$, where the latter is the cost of altering the local structure of the mRNAs and is calculated by RNAfold [22]; (3) Target site context. Several context features are related to miRNA targeting and have been found to boost the target site efficacy [9, 23]. Among these features, the local AU content near the target site, the target position within 3' UTR, and the residue pairing at 3' end of miRNA are used as the determinants to generate the target site context score by the TargetScan algorithm [9]. To ease the implementation, the combined context score of target sites can be extracted directly from the latest version of TargetScanHuman (Release7.0).

All of these three features yield negative values when functional miRNA targeting occurs. The smaller the value obtained from the formation of miRNA-target duplex, the more likely it suggests a functional target miRNA binding. To fit the adaptive lasso framework, each of the features needs to be transformed into a positive value as below,

$$f_t^{jk}(\text{tran}) = \text{sigmoid} \left(b_t * f_t^{jk} \right) = \frac{1}{1 + \exp(-b_t * f_t^{jk})} \quad (1),$$

where f_t^{jk} represents the t th feature for the mRNA j and miRNA k pair. In the data preprocessing step, b_t is a scale factor for the t th feature with a positive value. Therefore, the transformed features will keep their original rankings and fall into the same range $(0, 1/2)$. An appropriate value of b would be critical for variable selection in the adaptive lasso procedure. In the practical implementation of this study, the parameter is tuned for each feature set individually to achieve an optimal performance of the model.

3.3 Mathematical Model

Given the expression dataset corresponding to totally G mRNAs and K miRNAs in N samples, we let vectors $x_j = [x_{j1}, x_{j2}, \dots, x_{jN}]$ and $z_k = [z_{k1}, z_{k2}, \dots, z_{kN}]$ represent the expression levels of mRNA j and miRNA k , in samples 1 to N , respectively. Then, a linear relationship between the expressions of each mRNA j and the K miRNAs is assumed and it is represented by the following linear model,

$$x_j = \sum_{k=1}^K \beta_{jk} \cdot z_k + \varepsilon_j \quad (1)$$

where ε_j is a standard Gaussian noise, and β_{jk} represents the degree of regulatory effect for each miRNA-mRNA interaction. Considering the down-regulation effect of miRNA, β_{jk} is preferentially a negative value. Within this context, the goal of the study is to find a subset of miRNAs that regulate each mRNA with $\beta < 0$. The number of miRNAs that putatively regulate an mRNA is often larger than the number of samples, which possibly results in a high-dimensional model selection problem. To ensure proper estimation of parameter β , a L1 penalized least square estimator, known as the Lasso technique, has been widely adapted as a model selection method [19, 24, 25]. With the nature of sparse solution of Lasso, most of these regression coefficients β shrink to 0 and the nonzero ones will be fully recovered. Therefore, an L1 norm constraint on the coefficients is enforced in the model as the following:

$$\min \left\{ \frac{1}{2} x_j - \sum_{k=1}^K \beta_{jk} \cdot z_{k2} + \lambda_j \sum_{k=1}^K |\delta_{jk} \cdot \beta_{jk}| \right\} \quad (2)$$

subject to $\beta_{jk} \leq 0$, for $k=1, 2, \dots, K$, and K is the total number of miRNAs in the dataset. Here λ_j determines the degree of regularization of nonzero β_{jk} in the Lasso framework. This regularization parameter controls the sparsity of the selected model and the optimal value of λ_j can be selected based on cross-validation. There are two ways to select λ_j : the global and the local parameter tuning methods. It has been noted that the global tuning method performed better in selecting miRNA-mRNA interactions from the whole set of putative interactions than the local tuning method [18], so the global tuning method is chosen in our implementation.

To incorporate the prior knowledge of sequence features, the adaptive version of lasso method is applied. Here, the scale parameter δ from data is automatically learned from the data. We define δ_{jk} as a weighted sum of sequence features on the interaction between the j th mRNA and the k th miRNA, i.e.,

$$\delta_{jk} = \sum_t \omega_t f_t^{jk} \quad (3)$$

where f_t^{jk} is the t th feature for the interaction between mRNA j and miRNA k . Since we are interested in the relative contributions of different features, we further add the constraints that $\sum \omega_t = 1$ with constraint on the weights of each feature $\omega_t \geq 0$. The model can be illustrated as a network in Fig. 2. This optimization problem is solved by an iterative algorithm that (1) obtains the scale parameter δ by fixing the weight ω ; (2) minimizes the equation over β ; and (3) minimizes the equation over ω by fixing β . The steps are repeated iteratively until convergence. Both subproblems (2) and (3) are convex and they can be solved efficiently via an interior-point method and a gradient descent method, respectively. The model is implemented in MATLAB and the source code can be downloaded from <http://nba.uth.tmc.edu/homepage/liu/miRNALasso/>.

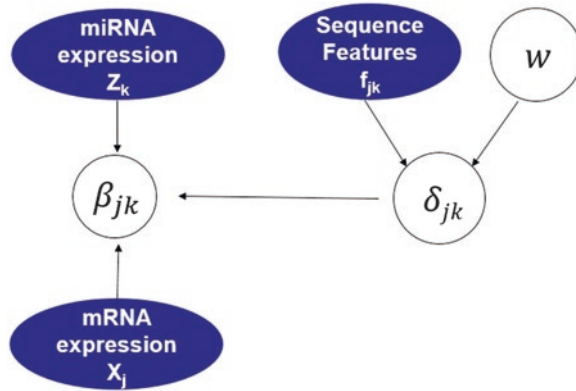


Fig. 2 Graphical model representation of adaptive Lasso. Shaded nodes denote known variables and unshaded ones are unknown parameters. w , weight vector for the three sequence features. δ , scale parameter in the adaptive lasso framework. β , correlation coefficients that are constrained to be negative values

3.4 Model Evaluations

3.4.1 Evaluation of Sensitivity and Specificity
(See **Notes 2 and 3**)

Experimentally validated miRNA targets datasets can be downloaded from the miRTarBase 4.5 as the gold standard [21]. To evaluate the model performance, the prediction results will be compared with the list of experimentally validated miRNA-mRNA pairs. The sensitivity and specificity of the method can then be systematically calculated according to the overlap between the predicted targets and the set of experimentally validated targets in miRTarBase. The performance of different methods can also be compared based on their corresponding receiver operating characteristic (ROC) curves and the area under the curve (AUC) values, using the ROCR package in R [26]. Here, six well-known methods are included in the comparison with our adaptive lasso method. Among them, the GenMiR++ [17], TaLasso [18], and Lasso [19] methods combine the sequence-based predictions and expression profiles, but neglect the full spectrum of sequence features. For the other three methods, the Context Score from TargetScan, the hybridization energy (ΔG), and the accessibility energy ($\Delta\Delta G$) are based on sequence information alone and do not consider expression data. In this chapter, we use the Multi Class Cancer (MCC) dataset as an example to demonstrate the superior performance of our adaptive lasso approach compared to other methods (Fig. 3).

3.4.2 Gene Ontology Enrichment Analysis

The Gene Ontology Biological Process (GO-BP) annotations can be extracted from the Gene Ontology Annotation database. The GO terms with less than five genes or with no biological Data available need to be removed, giving totally 2321 GO-BP terms. Then we examine the Gene Ontology categories enrichment for the predicted targets of each miRNA in each prediction method. The GO enrichment p-values are calculated by the hyper-geometric test, subject to the FDR control by Benjamini and Hochberg (BH)

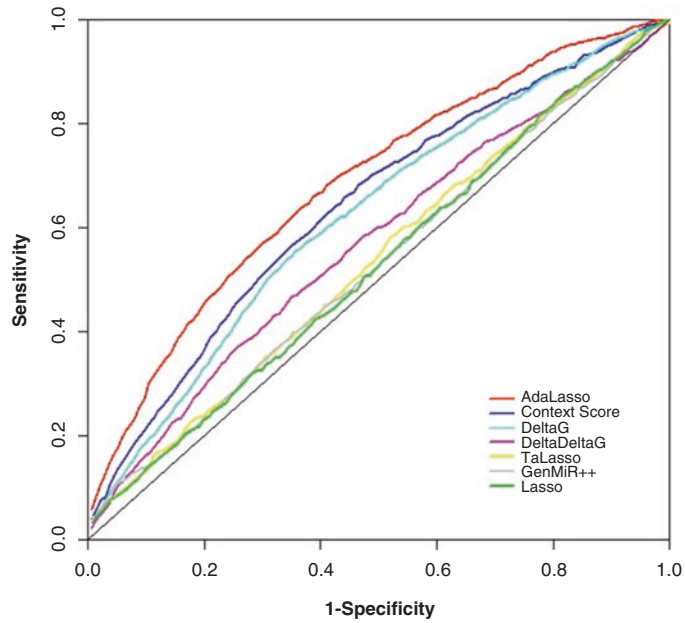


Fig. 3 ROC plots based on the experimentally validated miRNA-mRNA pairs from the miRTarBase. AdaLasso refers to the adaptive lasso method described in this chapter. Context Score, DeltaG, and DeltaDeltaG are predictions based on the context score, ΔG , and $\Delta\Delta G$, respectively. TaLasso, GenMiR++, and Lasso refer to methods that integrate the expression information into sequence-based target predictions [17–19]

multiple testing procedure. The GO terms with $FDR < 0.1$ are considered significantly enriched by predicted miRNA target sets. For the MCC dataset, GenMiR++ method yielded totally 400 significantly enriched GO functional terms with $FDR < 0.1$, whereas TaLasso and our adaptive lasso approach resulted in 440 and 468 significantly enriched GO terms, respectively.

3.4.3 Examining the Clinical Relevance of Predicted Targets

As an additional evaluation metric, we also examine whether the predicted miRNA targets are biologically meaningful by analyzing the correlation between their gene expression profiles and the samples. Specifically, for the MCC dataset, we examined the association between the expression levels of each target gene with the tumor samples. The signal-to-noise (S2N) statistics was obtained for each gene, as in a previous study [27]. Here, the S2N of each gene was calculated as $(\mu_{\text{tumor}} - \mu_{\text{normal}})/(\sigma_{\text{tumor}} + \sigma_{\text{normal}})$, where μ and σ represent the mean and standard deviation of expression levels of each gene, respectively. To calculate whether the gene is a statistically significant tumor marker, we run 1000 permutations of the sample labels, and recalculate the S2N values for each gene in each permutation. The target gene is considered to be a statistically significant tumor marker if the observed S2N value exceeds the permuted ones at least 99% of the time ($p\text{-value} \leq 0.01$).

For MCC dataset, out of the top 1000 miRNA-mRNA pairs predicted by the adaptive lasso method, 22.8% of target genes were considered significant tumor markers ($p\text{-value} = 7.38\text{E-}35$, chi-squared test). In contrast, the Context Score method yielded the smallest percentage (14.4%) of significant tumor markers compared to all the expression-based methods for MCC dataset. Using the full set of genes in the MCC dataset as the background, only 7% of them can be considered tumor markers based on the S2N metric calculation.

4 Notes

1. The method described in this chapter is based on three assumptions: First, the number of predicted interactions is small. This assumption is ensured by the sparse solution of the linear regression problem. Second, we assume miRNAs down-regulate their corresponding mRNA targets. This assumption could lead to a potential problem that the targets affected by translation repression only would not be identified by the method described here. For such targets, the relevant miRNAs might not inhibit the mRNA expression profiles, but would reduce the protein levels [8, 28]. It would be more comprehensive to examine the changes at both mRNA and protein levels if datasets are available. Third, the full spectrum of sequence features will make a significant contribution on improving target prediction.
2. It has been well known that plant miRNAs often have targets with perfect or near-perfect sequence complementarity. In contrast, animal miRNA target sites only exhibit partial complementarity [29]. The distinct pairing requirements and target breadth of animal and plant miRNAs suggest simpler target identification in plants than that in animals. Therefore, the method described here should not be used to predict miRNA-mRNA interactions in plants, but can be broadly applied for miRNA target predictions in other animal species.
3. Besides the three sequence features examined in this model, the degree of target site conservation can also be used as an additional sequence feature to represent the conservation of the target sites. However, the coverage of this feature in existing databases is significantly lower than the three features we used in this study. Therefore, to achieve high coverage of prediction results, we chose not to incorporate the target site conservation information in the method. The future work will focus on leveraging the additional prior knowledge of miRNA targets to improve the prediction accuracy, since this adaptive form of regularized regression approach is capable of efficiently incorporating data from different sources into the model.

Acknowledgment

This work is supported in part by NIH grant R01 LM010022 and the seed grant from the University of Texas Health Science Center at Houston.

References

1. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
2. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20
3. Fabian MR, Sonenberg N, Filipowicz W (2010) Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* 79:351–379
4. Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* 12:99–110
5. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A et al (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129:1401–1414
6. Kozomara A, Griffiths-Jones S (2011) miR-Base: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39:D152–D157
7. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19:92–105
8. Selbach M, Schwanhauss B, Thierfelder N, Fang Z, Khanin R, Rajewsky N (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature* 455:58–63
9. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 27:91–105
10. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ et al (2005) Combinatorial microRNA target predictions. *Nat Genet* 37:495–500
11. Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G et al (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics* 10:295
12. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res* 36:D149–D153
13. Kruger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 34:W451–W454
14. Xu W, San Lucas A, Wang Z, Liu Y (2014) Identifying microRNA targets in different gene regions. *BMC Bioinformatics* 15:11
15. Xu W, Wang Z, Liu Y (2014) The characterization of microRNA-mediated gene regulation as impacted by both target site location and seed match type. *PLoS One* 9:e108260
16. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39:1278–1284
17. Huang JC, Morris QD, Frey BJ (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J Comput Biol* 14:550–563
18. Muniategui A, Nogales-Cadenas R, Vazquez M, Aranguren XL, Agirre X, Luttun A et al (2012) Quantification of miRNA-mRNA interactions. *PLoS One* 7:e30766
19. Lu Y, Zhou Y, Qu W, Deng M, Zhang C (2011) A lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics* 27:2406–2413
20. Wang Z, Xu W, Liu Y (2015) Integrating full spectrum of sequence features into predicting functional microRNA-mRNA interactions. *Bioinformatics* 31:3529–3536
21. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH et al (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 42:D78–D85
22. Gruber AR, Bernhart SH, Lorenz R (2015) The ViennaRNA web services. *Methods Mol Biol* 1269:307–326
23. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nat Struct Mol Biol* 18:1139–1146
24. Cao S, Qin H, Deng HW, Wang YP (2014) A Unified sparse representation for sequence

- variant identification for complex traits. *Genet Epidemiol* 38:671–679
25. Wang Z, Xu W, San Lucas FA, Liu Y (2013) Incorporating prior knowledge into gene network study. *Bioinformatics* 29:2633–2640
26. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21:3940–3941
27. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M et al (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 98:15149–15154
28. Zhuang X, Li Z, Lin H, Gu L, Lin Q, Lu Z et al (2015) Integrated miRNA and mRNA expression profiling to identify mRNA targets of dysregulated miRNAs in non-obstructive azoospermia. *Sci Rep* 5:7922
29. Axtell MJ, Westholm JO, Lai EC (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol* 12:221

Chapter 11

Computational and Experimental Identification of Tissue-Specific MicroRNA Targets

Raheleh Amirkhah, Hojjat Naderi Meshkin, Ali Farazmand,
John E.J. Rasko, and Ulf Schmitz

Abstract

In this chapter we discuss computational methods for the prediction of microRNA (miRNA) targets. More specifically, we consider machine learning-based approaches and explain why these methods have been relatively unsuccessful in reducing the number of false positive predictions. Further we suggest approaches designed to improve their performance by considering tissue-specific target regulation. We argue that the miRNA targetome differs depending on the tissue type and introduce a novel algorithm that predicts miRNA targets specifically for colorectal cancer. We discuss features of miRNAs and target sites that affect target recognition, and how next-generation sequencing data can support the identification of novel miRNAs, differentially expressed miRNAs and their tissue-specific mRNA targets. In addition, we introduce some experimental approaches for the validation of miRNA targets as well as web-based resources sharing predicted and validated miRNA target interactions.

Key words MicroRNA, Computational target prediction, Machine learning, Next-generation sequencing, Cross-linking and immunoprecipitation

1 Introduction

Computational predictions suggest that miRNAs can target hundreds of transcripts, many of which are regulated by multiple miRNAs. Most algorithms assume a global pattern governing miRNA–mRNA interactions. Such approaches ignore potential tissue-specific functions of miRNAs, and thereby generate results that may include many false positive predictions.

Fortunately, cross-linking and immunoprecipitation (CLIP) provides the opportunity to study direct RNA–protein interactions in a tissue-specific manner [1]. The combination of CLIP with high-throughput sequencing (HITS) allows the large-scale identification of mRNA interactions with components of the RNA induced silencing complex (RISC). Since its introduction HITS-CLIP has increased the number of experimentally verified

miRNA–target pairs while facilitating the development of tissue-specific miRNA target predictors. In this chapter we review a newly developed tissue-specific machine learning (ML)-based algorithm. We describe features of the algorithm that are crucial for determining tissue-specific miRNA target interactions. In addition, we provide a brief overview of the advances in high-throughput target identification techniques along with established and new web resources.

1.1 *MicroRNAs*

MiRNAs, a highly studied class of short noncoding RNAs (ncRNAs), are able to negatively regulate target gene expression by guiding the RISC to miRNA response elements (MREs), and thereby mediating translation repression or target mRNA degradation. The multiplicity of interactions between miRNAs and their targets contributes significantly to the complexity of gene regulation. By affecting the expression of numerous genes in parallel, miRNAs are able to fine-tune various cell-biological processes, including differentiation, proliferation, and apoptosis [2–4]. A recently introduced concept suggests that miRNAs could be mediators of a regulatory language, through which RNA transcripts such as mRNAs, pseudogene transcripts, and ncRNAs crosstalk with each other, thus forming a novel gene regulatory layer composed of competing endogenous RNAs (ceRNAs) [5, 6]. RNA transcripts involved in this regulatory network affect each other's expression levels by competing for the same pool of miRNAs via MREs [6].

The deregulation of miRNA expression is known to induce pathological events underlying diverse diseases, including different cancers [7, 8]. Mounting evidence also shows that miRNA function is tissue-specific and thus adds complexity to the problem of finding true miRNA targets [9, 10]. To address this problem a web resource (miTALOS) was developed to incorporate the potential tissue-specific effects of miRNAs and target mRNAs to improve the analysis of miRNA regulation in biological pathways [11, 12]. Another study conducted by Ritchie et al. highlighted the importance of conservation of miRNA and target expression patterns as an indicator of functional miRNA target regulation. They presented an approach that integrates expression data from human and mouse to determine conserved patterns of expression between orthologous miRNAs and mRNA genes [13]. Clark et al. showed that most tissue-specific miRNA–mRNA interactions carry a noncanonical seed region [9]. However, Agarwal et al. demonstrated that these noncanonical binding sites do not mediate gene repression, and that the vast majority of functional sites are canonical [14].

1.2 *Functional Characterization of Predicted Targets*

One approach to determine the function of miRNAs is through transcriptomics profiling of miRNA expression. Differentially expressed miRNAs (DEMs) in microarray, RT-PCR, or next-generation sequencing (NGS) experiments may be linked to

genetic profiles underlying distinct samples or experimental conditions. Known or predicted targets of DEMs showing an inversely correlated expression pattern compared to their miRNA regulators are good candidates for phenotypic effectors responding to post-transcriptional regulation by miRNAs in certain experimental conditions. This idea has been adopted in many studies investigating the impact of miRNA regulation on pathogenesis and disease progression [15–18].

Typically, a functional enrichment analysis is performed on sets of predicted and inversely correlated target genes to find co-occurrences in biological processes or pathways. CORNA for example, a package for the statistical programming language R (www.r-project.org), allows researchers to analyse gene lists obtained from high throughput experiments for significant miRNA–target associations. CORNA links miRNA targets to functional annotations and permits visualization of quantitative expression data [19]. This is achieved by selecting one of three statistical methods (HyperGeometric test, Fisher’s exact test, or the c2-test) for testing a gene/transcript list for significant miRNA associations and for testing miRNA targets for enriched associations with Gene Ontology (GO) terms or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The web-based software miR2GO analyzes the functional impact of mutations in miRNA seed regions by integrating miRNA target prediction with GO enrichment analysis and semantic comparison of functional annotations [18, 20]. In the web resource mimiRNA miRNA and mRNA expression data from 1483 samples across 188 different tissues and cell types was integrated. mimiRNA helps researchers to identify expression profiles and to discover functional relations between miRNAs and mRNAs [17].

Biologically relevant molecular networks and signaling pathways affected by DEM targets can be determined with pathway analysis tools and resources, such as iPathwayGuide (www.advaita-bio.com), KEGG (www.kegg.jp), PathVisio (www.pathvisio.org), Panther (www.pantherdb.org), GeneMANIA (www.genemania.org), or DIANA-miRPath [21]. DIANA-miRPath performs enrichment analyses of predicted miRNA target genes in biological pathways and examines the combinatorial effect that co-expressed miRNAs have on a given pathway. Since version 3 users can perform a hierarchical clustering analysis of miRNAs and pathways based on the levels of their interactions and can create heat maps of miRNA–pathway associations. Furthermore, DIANA-miRPath identifies pathological single nucleotide polymorphisms (SNPs) at miRNA binding sites [21].

Once important target genes are identified in silico, experimental approaches for the validation of the functionally inverse relationship between miRNAs and their targets are employed and used to prove either loss-of-function or gain-of-function upon miRNA induction and knockdown.

2 Principles of miRNA Target Prediction

Most miRNA target prediction algorithms use similar general principles for the identification of putative target sites in mRNA 3' untranslated region (UTR) sequences. These principles include (1) sequence complementarity between a miRNA and a putative target site, (2) target site conservation in related species, (3) target site accessibility in a secondary structure model, (4) the thermodynamic stability of a putative miRNA–mRNA duplex, and (5) defining the expression profiles of target genes [22]. Multiple binding sites for one miRNA in an mRNA 3' UTR sequence usually add to the score that some algorithms assign to a predicted miRNA–target pair [23, 24]. In the following section the differences between existing approaches are further explained.

2.1 History of miRNA Target Prediction

Over the last decade, an array of computational approaches has been proposed to predict miRNA target sites and to guide investigations towards the understanding of miRNA function. These approaches can be broadly classified into rule-based (ab initio) and data-driven (ML-based) approaches.

2.1.1 Rule-Based Approaches

Rule-based methods make use of experimentally derived insights into mechanisms underlying miRNA target recognition to facilitate the design of computational models. For example, target sequence complementarity to the miRNA 5' ends is a widely accepted prerequisite for effective target recognition in most miRNA target interactions and has been consistently used since early miRNA target prediction algorithms [25, 26].

DIANA-microT is a rule-based method which predicts miRNA targets by searching for sequence complementarity with a focus on seed match composition [27]. TargetScanS expands this rule and scores target sites based on the evolutionary conservation of their seed region in five species (human, mouse, rat, dog and chicken), assuming that conserved target sites are more likely to be functional [26, 28]. Ritchie et al. additionally considered the conserved inverse correlation of miRNAs and putative targets in their miRNA target prediction approach [13].

Rule-based approaches are often complemented by another criterion namely a low Gibbs free energy value of the putative miRNA–mRNA duplex. The free energy is an indicator for the thermodynamic stability of a RNA duplex [29]. This measure is for example considered in the algorithms RNAhybrid and PITA [29, 30]. RNAhybrid calculates the minimum free energy of potential hybridization sites and requires perfect sequence complementarity between nucleotides 2 and 7. PITA evaluates the accessibility of the binding site by calculating the difference between the minimum free energy of the duplex and the energy required to open the target site (making it accessible to the miRNA).

2.1.2 ML-Based Approaches

As the number of experimentally verified miRNA–mRNA interactions has increased, statistical or ML-based miRNA target prediction approaches have gained popularity [31]. ML-based approaches make direct use of the experimental data arising from validated instances of miRNA target interactions [32]. Consequently, the efficacy of ML-based approaches is dependent on the availability and quality of experimental data, including negative samples.

Yousef et al. presented a Naïve Bayes (NB) classifier, named NBmiRTar, which was trained with sequence-based features of miRNA–mRNA pairs derived from validated miRNA–target interactions (positive data) and artificially generated negative samples [33]. A NB classifier is a probabilistic model derived from empirical data. It calculates the probability that a given instance belongs to a certain class and assumes there is strong independence between the features used in the model. Despite its simplicity, the NB classifier is often used due to its accuracy and performance superiority compared to more sophisticated classification methods [34].

In contrast to NB, support vector machines (SVM) are classifiers that combine all features of each class into a single vector in an n -dimensional space and try to find a *hyperplane* that produces the best separation between the two classes [34]. SVMicrO is an example of a two-stage SVM-based algorithm for miRNA target prediction that was developed by Liu and colleagues [35]. Their predictions, based on training data and results from high confidence experiments including immunoprecipitation pull down and proteomics experiments like stable isotope labeling with amino acids in cell culture (SILAC), showed an improved sensitivity, specificity, and precision in comparison to several popular target prediction algorithms, including TargetScan, PicTar, miRanda, and PITA.

mirMark is a ML-based method that utilizes experimentally verified miRNA targets from the miRecords [36] and mirTarBase [37] databases as training sets and considers over 700 features at both the site and UTR level [38]. Features incorporated hybridization energy, seed match type, pattern of the miRNA–target duplex pairing for the first 20 nucleotides of the miRNA, target site accessibility, composition, conservation, and location, and 3' UTR length.

Despite the continuing improvement in performance of recent methods, the small overlap between their results (5–70%) suggests that current methods still suffer from high false positive rates [39]. For that reason, databases such as miRò [40] or miRWalk [41] provide information on interactions commonly predicted by several tools with the intention to add confidence to these miRNA–target interactions [42]. Ritchie et al., however, demonstrated that this strategy does not necessarily identify true miRNA–target interactions [24]. Moreover, approaches based on combinations of algorithms may lead to reduced sensitivity wherein valid interactions that are not part of the intersection of several algorithms may

be omitted. A major disadvantage in most of the existing algorithms is that they are not designed to determine tissue- or disease-specific interactions, but rather assume universal patterns underlying miRNA–mRNA interactions. They are thus unreliable for the purpose of identifying tissue-specific miRNA target interactions. To the best of our knowledge CRCmiRTar [43] is the first miRNA target prediction algorithm that was developed based on cancer specific training data. In Subheading 3.1 we describe this algorithm in more detail.

**2.2 Features
of miRNAs that Affect
Target Recognition**

The prediction of miRNA targets in vertebrates is a particularly difficult challenge because of the complexity of miRNA target recognition. Thus, an important step in miRNA target prediction is the selection of features that optimise the discrimination between positive and negative data (Fig. 1). It is well established that the inclusion or exclusion of certain groups of features can significantly influence algorithm performance [43, 44]. Thus, a careful selection of relevant features for classification is very important. Many studies incorporate seed region complementarity and the evolutionary conservation of putative target sites as critical features that enhance the specificity of predictions for functional target sites [45]. Wen et al., using data from CLIP-seq and miRNA transfection experiments, showed that conservation and seed matching are most beneficial for the prediction of bona fide miRNA–target interactions [1]. With respect to seed region complementarity, most miRNA target prediction methods seek canonical seed matches, although it is now known that functional miRNA–target interactions can also harbour noncanonical seed

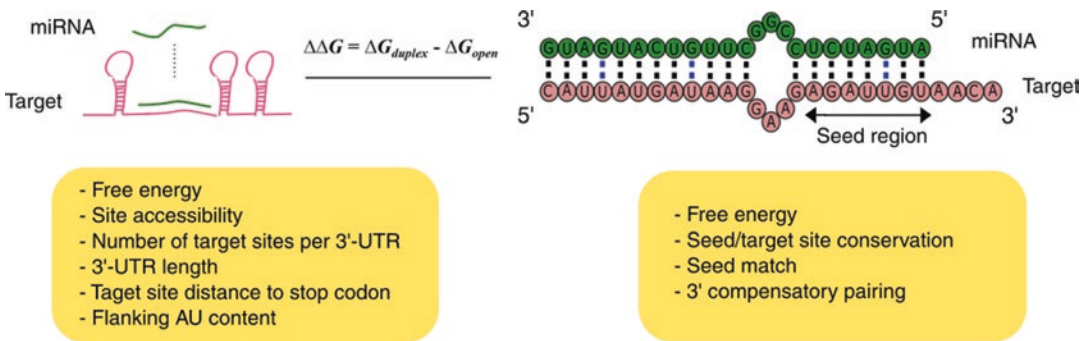


Fig. 1 Schematic overview of a miRNA–mRNA interaction and the features used in current miRNA–target prediction algorithms. Canonical sites refer to the regions in the mRNA that contain the exact sequence of partner bases for the bases in the miRNA seed. Noncanonical refers to the pairing between the miRNA seed and mRNA that is not completely complementary. 3' compensatory pairing refers to supplementary pairing outside of the seed region, especially to nucleotides 12–17 of the miRNA. Flanking AU content refers to the concentration of A and U nucleotides within 30 nucleotides upstream and downstream from the seed of the miRNA. ΔG_{duplex} stands for the minimum free energy of the miRNA–target hybrid and ΔG_{open} stands for the energy required for the target site to be accessible to the RISC complex

matches, e.g., involving G:U wobble base pairs or bulges [9]. MIRZA-G, a recently published algorithm, incorporates canonical and noncanonical seed matches in a unified way [46]; however it does not take into account the spatial configuration of the mRNA target based on its secondary structures [47]. The mRNA secondary structure plays an important role in miRNA target recognition, as it determines the target site accessibility to the docking miRNA–RISC complex [30, 47]. Another observation is that target sites are often located in AU-rich regions. This feature is considered in the TargetScanS algorithm, which rewards an AU-rich nucleotide composition around putative target sites with a higher prediction score [48]. There is also evidence that miRNAs preferentially target genes with longer 3' UTRs and preferably bind target sites near any of the two ends of the 3' UTR [49].

2.3 Tissue-Specific miRNA Target Prediction

Although ML-based methods have improved the sensitivity of current miRNA–target prediction algorithms, the rate of false positive predictions remains problematical. Given the highly tissue-specific expression signatures of miRNAs and target transcripts, tissue-specific miRNA functions must be considered, e.g., in the analysis of miRNA regulation in specific pathological conditions [50]. A recent study confirmed that miRNA target interactions, predicted by current computational approaches, are not functional in all tissues or diseases [51]. In fact some binding sites of previously validated targets were not accessible for miRNAs in some tissues because they are masked by RNA binding proteins or occluded by the mRNA secondary structure [52]. The role of RNA binding proteins in the co-translational and posttranscriptional regulation of gene expression is well established. However, an analysis of RNA-binding protein motifs has not been included in any miRNA target prediction method [39].

Alternative cleavage and polyadenylation (APA) of mRNA 3' UTRs are other phenomena that may explain tissue-specific miRNA function. Nam and colleagues reported APA as a major cause of differential miRNA targeting [53]. Their results showed that many transcripts contain alternative 3' UTR isoforms, which affect miRNA targeting. They proved that 3' UTR lengths can change across different vertebrate cell types and developmental stages. In fact, 3' UTR shortening by APA can result in the deletion of miRNA binding sites in a predicted target and consequently give rise to false positive predictions. This certainly complicates reliable predictions for interactions in a given tissue. However, it was shown that the prediction of miRNA regulatory efficacy can be improved by considering the 3' UTR landscape of putative targets [42]. Another factor that can affect miRNA binding to a given transcript in different tissues is single-nucleotide polymorphisms (SNPs). 3' UTR heterogeneity caused by SNPs or APA affects miRNA-mediated posttranscriptional regulation and must be taken into

account for the prediction of miRNA target sites within different types of tissues. NGS technologies, particularly the increasing availability of RNA-seq data capturing whole transcriptomes in one experiment, provide new opportunities for miRNA target predictions at an isoform level. Thus, nondominant isoform targets as well as their isoform-specific target sites can be detected, which are otherwise invisible to other approaches [54]. The identification of tissue and disease-specific miRNA target genes would ultimately contribute to a greater understanding of their biological role and may also facilitate the identification of potential drug targets.

2.4 Naïve Bayes Classifier Predicts Colorectal Cancer Specific miRNA Targets

As described earlier in Subheading 2.1 most computational miRNA target prediction methods assume general patterns underlying miRNA–mRNA interactions and therefore produce a large portion of false positive predictions. Recently, we reported a ML-based algorithm, named CRCmiRTar, to examine colorectal cancer (CRC)-specific miRNA target relationships. CRCmiRTar, an NB classifier, was trained with data from experimentally validated miRNA target interactions in CRC and other cancers. We applied two strict filters in the data selection step to ensure the reliability of our dataset and an increased specificity of our classifier: (1) the interactions had to be experimentally validated in CRC; and (2) the exact binding site must have been identified by luciferase reporter and mutagenesis assays. We also constructed two negative datasets, one set from validated miRNA–mRNA interactions reported in other cancers (non-CRC interactions) and another set that integrates tissue-specific negative examples, previously presented in [55].

Since many tissue-specific miRNA–mRNA interactions depend on noncanonical seed regions [9], our model does not filter out miRNA–target site pairs with non-perfect seed matches in order to predict tissue-specific and 3′-compensatory target sites.

For the identification of CRC-specific miRNA target interactions, 14 relevant features including a set of position-based, sequence, structural, and thermodynamic features have been selected among the 72 features using two feature selection methods: (1) correlation-based feature selection (CFS) and (2) ReliefF. We used these features and our training data to compare the performance of several ML approaches (NB, SVM, Random forest (RF), Artificial Neural Network (ANN)). Based on the results we used a NB classifier to identify interactions between miRNAs and target mRNAs in CRC. *See* Fig. 2 for an illustration of the CRCmiRTar workflow.

Evaluation of the algorithm based on tenfold cross-validation showed a significant improvement in performance with respect to the area under the receiver operating characteristic curve (AUC) and sensitivity compared to previous non-tissue-specific ML-based algorithms (Fig. 3).

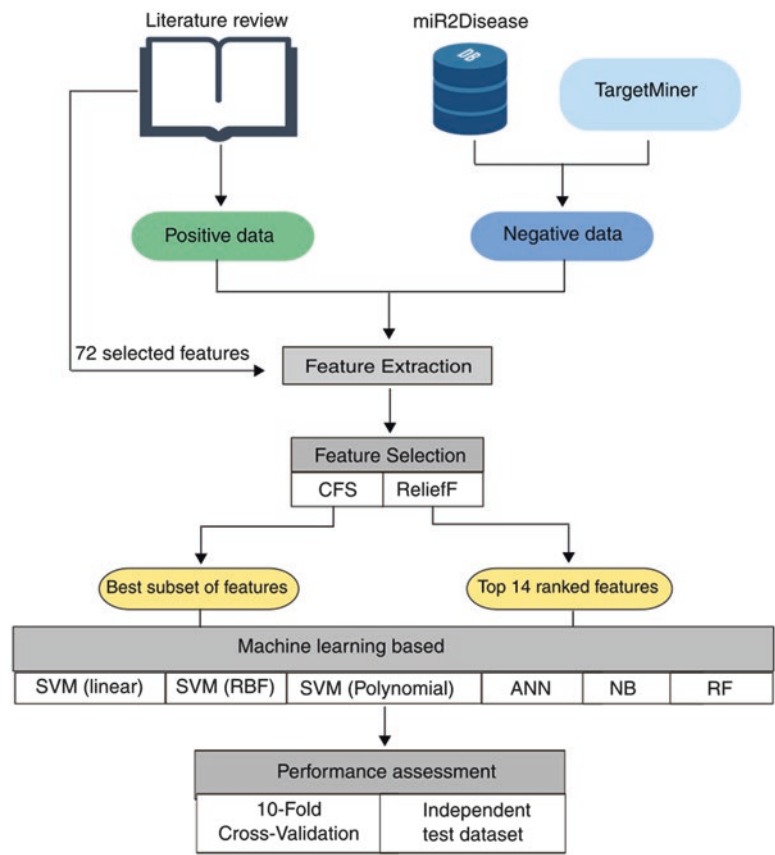


Fig. 2 Implementation of the CRCmiRTar workflow. The positive dataset involves experimentally validated CRC-specific miRNA–mRNA interactions. The negative dataset consists of the tissue-specific negative data used in [55] and validated miRNA–mRNA target interactions from other cancers based on the miR2Disease database. In the next step, 72 features were extracted from positive and negative interaction data. Two different feature selection methods were used to identify the most informative features. Performance of the various classifiers was assessed using tenfold cross-validation and an independent test dataset. The figure was reproduced from [43] with permission from the Royal Society of Chemistry

In addition, CRCmiRTar predicts experimentally validated miRNA target genes with higher accuracy compared to other algorithms. Our results indicate that sequence features, including miRNA and target base composition, confer the highest contribution to the specificity ($SPC = 0.91$) of our model. All features used in our model are listed in Table 1.

Although we have demonstrated the utility of our algorithm for the prediction of miRNA target interactions in CRC, the underlying approach can be adopted for the implementation of other disease or tissue-specific miRNA target prediction algorithms. We conclude from this study that reliable tissue-specific miRNA

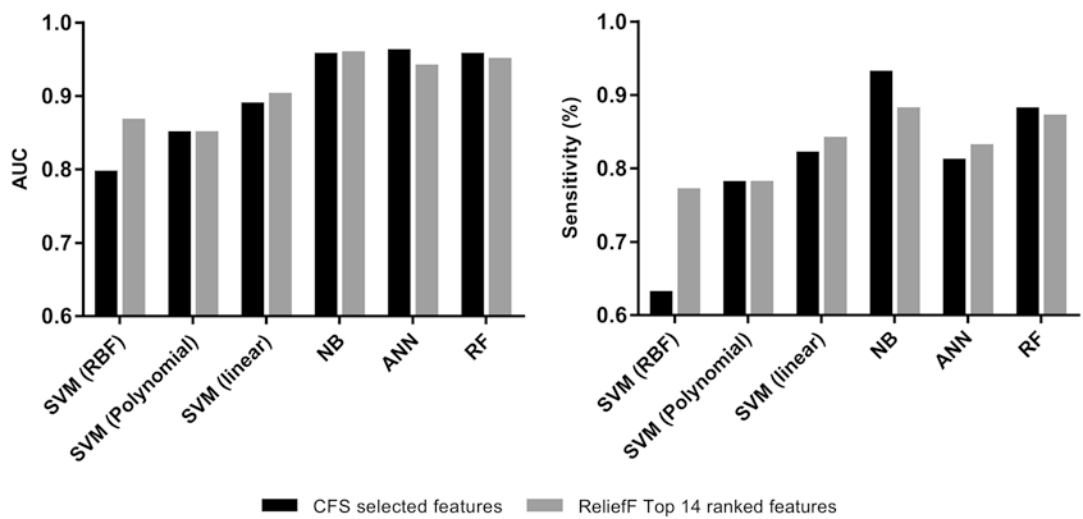


Fig. 3 Evaluation of different classifiers on two categories of selected features using CFS and ReliefF. The two plots illustrate the performances of the different classifiers in the tenfold cross-validation with respect to (a) the AUC; and (b) their sensitivity. SVM (RBF): support vector machine with radial basis function kernel; NB: Naïve Bayes; ANN: Artificial Neural Network; RF: Random Forest; AUC: area under the receiver operating characteristic curve. Reproduced from [43] with permission from the Royal Society of Chemistry

Table 1
Selected features by correlation-based feature selection [43]

Features	Description	Feature type
UU_seed	UU di-nucleotide frequency in seed matching site	Sequence
CG_seed	CG di-nucleotide frequency in seed matching site	Sequence
AU_match	Frequency of AU base pairs in seed region	Structural
UA_match	Frequency of UA base pairs in seed region	Structural
GC_target	GC di-nucleotide frequency in target site	Sequence
GC_match	Frequency of GC base pairs in seed region	Structural
CG_match	Frequency of CG base pairs in seed region	Structural
Seed	Number of base pairings to the miRNA 8-mer seed	Structural
Tail	Number of base parings to the first 8 nucleotides of the miRNA 3' end	Structural
Cons_bp_mir_5p	Number of consecutive base-pairings to the miRNA 5' end with two allowed non-pairing positions	Structural
Binding asymmetry	the ratio between the number of paired bases in the 3p versus the 5p region of the microRNA (considering 8 nucleotides on each side)	Structural
B_tagt_s6	bulges in target sequences of size 6 nt and more	Structural
Pos_3	Position 3	Position
Pos_7	Position 7	Position

target prediction depends on the kind of the data that is used to train a computational model. Therefore, we are confident that with the inclusion of NGS-derived data on miRNA and mRNA expression, isoforms and cross-links further improvements in tissue-specific miRNA target prediction can be achieved.

3 NGS Improves miRNA Target Predictions

Small RNA sequencing (sRNA-seq) makes it possible to determine the presence of all miRNAs in a sample, or to compare miRNAome expression in different cellular or disease conditions. In contrast to microarray and HT-qPCR experiments, deep sequencing has the advantage that even novel sequences or isoforms can be detected. Higher sensitivity and a wider dynamic range are additional advantages of NGS. RNA fragments with single nucleotide variants, edited miRNAs or RNAs with very low copy numbers may be readily detected. Thus, miRNA isoforms and family members can be distinguished by NGS. With respect to miRNA target interactions, the recently developed combination of RNA-seq with CLIP can provide evidence for the presence of predicted miRNA–target interactions, as this technique detects active components of the RISC complex hybridized to mRNA molecules.

In summary, data obtained from sequencing experiments can be used to characterize miRNA expression and function by expression profiling, identification of isoforms, prediction of novel miRNAs, and their downstream mRNA targets. Furthermore, precursor miRNAs as well as sequence modifications can be easily identified by small RNA sequencing [56].

3.1 Bioinformatics Analysis of sRNA-Seq Data

sRNA-seq experiments generate millions of short reads. Prior to addressing specific research questions, several basic data preprocessing steps have to be performed in order to extract the relevant information from these raw data. Basic data preprocessing steps include an initial filtering to exclude low quality reads, and the removal of adaptor sequences from the reads using adapter trimming tools like trimLRPatterns [57]. In the next step, all reads are mapped to a reference genome of the organism under investigation, for example with Bowtie [58]. Thereafter reads that belong to other RNA species are filtered out prior to further downstream analysis. To this end, mapped reads are annotated with the help of annotation files for coding and noncoding RNAs, e.g., from the UCSC table browser. Reads mapping to exonic regions of coding genes will be excluded from further analysis and ncRNAs will be quantified in each library [56].

A number of tools have been developed to process NGS data. miRMOD, for example, is a windows-based standalone tool which detects miRNA modifications in a sRNA-seq dataset, and predicts

the effect on miRNA–target interactions [59]. It has been reported that miRNA sequence modifications may affect miRNA stability and the ability to bind mRNA targets [60]. Several web services and computational pipelines are available for the detection, prediction, and quantification of expressed miRNAs through NGS data. miRDeep was the first stand-alone tool for the detection of both known and novel miRNAs [61]. The algorithm uses Bayesian statistics to score reads based on their fit to the biological model of miRNA biogenesis. Other tools include miRanalyzer [62], miRNAkey [63], miREval [64] and miRExpress [65]. The web server CPSS (a computational platform for the analysis of small RNA deep sequencing data) analyses sRNA-seq data from paired samples for differential miRNA expression, while it can also be used to predict novel miRNAs in individual samples [66]. CPSS allows users to predict target genes for differentially expressed known and novel miRNAs and also to perform functional enrichment analysis. The UEA sRNA workbench is a collection of independent tools for the analysis and visualization of sRNA-seq data, the annotation and prediction of novel miRNAs and for profiling sRNA expression patterns [67]. Similar to the UEA sRNA workbench is the sRNAtoolbox, which provides interconnected tools for sRNAseq expression profiling and functional downstream analyses [68]. MiREval 2.0 is able to simultaneously search up to 100 sequences for novel miRNAs in multiple organisms and uses multiple published bioinformatics approaches to detect miRNAs in sequences of interest from DNA or RNA [64].

3.2 CLIP-seq Enhances Identification of miRNA Targets

CLIP-seq technologies comprise diverse experimental approaches including HITS-CLIP [69], photoactivatable ribonucleoside-enhanced (PAR)-CLIP [70] and cross-linking ligation and sequencing of hybrids (CLASH) [71]. They facilitate the high-throughput identification of putative miRNA targets. Each of these approaches involves ultraviolet cross-linking and immunoprecipitation of AGO-bound miRNA–mRNA complexes, digestion and release using proteinase. Adaptors are ligated to RNA fragments followed by reverse transcription and sequencing.

Individual-nucleotide resolution CLIP (iCLIP) was developed to detect protein–RNA cross-link sites on a genome-wide scale. The high specificity of this approach is achieved by an intramolecular cDNA circularization step that enables the analysis of truncated cDNAs that are usually lost during the standard CLIP library preparation protocol [72].

PAR-CLIP is a modified version of the original CLIP approach based on the introduction of 4-thiouridine (4SU) into transcripts, which causes a T to C conversion and thereby enhances the efficiency of cross-linking and the resolution in protein–RNA

binding sites. While HITS-CLIP and PAR-CLIP provide individual sets of sequencing reads that are either derived from miRNAs or target mRNAs, CLASH has an additional step in which the miRNA and target mRNA sequences are ligated to produce chimeric reads that comprise a respective miRNA sequence along with the targeted mRNA. However, CLASH suffers from a relatively low efficiency of hybrid capture and failure to provide quantitative information regarding the strength of miRNA target interactions. MiRNA cross-linking and immunoprecipitation (miR-CLIP) uses a miRNA modified with psoralen and biotin to capture its targets in cells [73]. The association of these biochemical approaches with NGS will augment knowledge concerning miRNAs, seed pairing, conservation, structure and their tissue-specific function.

CLIP-seq for the large-scale investigation of miRNA target interactions has become popular and as a consequence computational analysis pipelines and resources have been developed. Web resources include CLIPZ [74], starBase [75], and doRiNA [76]. PIPE-CLIP [77], is a recently released Galaxy pipeline designed for genome-wide CLIP-seq data analyses. Other tools include Piranha [78], PARalyzer [79], and miRTarCLIP [80]. The latter was developed by Chou et al. and is an example of a computational approach for the analysis of HITS-CLIP and PAR-CLIP sequencing data towards the identification of miRNA–target interactions [80]. Table 2 provides an overview about current tools and resources for CLIP-seq data analysis.

4 miRNA Target Validation and Resources

4.1 *Experimental Methods to Validate miRNA Target Predictions*

Bioinformatics tools are used to predict potential targets of miRNAs, thereby preselecting candidates for experimental validation. The enormous number of predicted targets, however, challenges experimental scientists to choose candidates for further investigation. Small and large-scale techniques have been used to verify target predictions. The most extensively used small-scale wet lab experimentation techniques are reporter gene assays and gene directed mutagenesis for the identification of exact miRNA binding sites [82]. Western blotting, however, is typically used to confirm protein level changes of miRNA targets [83].

High throughput approaches for the identification of miRNA targets include proteomics methods like SILAC, followed by mass spectrometry, and transcriptomics approaches including microarrays and sequencing-based methods, such as HITS-CLIP [69], and PAR-CLIP [70].

A more detailed description of experimental methods for the identification and validation of interactions between miRNAs and

Table 2
A comparison of several resources for CLIP-seq data analysis

	StarBase	PARalyzer	CLIPZ	Piranha	doRINA	miRTarCLIP	Pipe-clip
Resource type	Database	standalone tool	database/web tool	software tool	database	standalone tool	Galaxy workflow
Data type	HITS/PAR-CLIP/ iCLIP/CLASH	PAR-CLIP	HITS/PAR-CLIP	HITS/PAR- CLIP/iCLIP & RIP-seq	HITS/PAR-CLIP	HITS/PAR-CLIP	HITS/PAR- CLIP/iCLIP
Prediction of miRNA targets	Seed-rule	Seed-rule	EIMMO [81]	–	PicTar	TargetScan, miTarBase	–

their target mRNAs can be found in the other chapters of this book and elsewhere [84, 85]. However, despite all the technical advancements, an integration of experimental data with computational predictions is still required to understand the biological impact of miRNA-mediated gene regulation.

4.2 Resources of miRNA–Target Interactions

Numerous databases have been created to store information on miRNA–target interactions from computational predictions or experimental results. Examples of databases with experimentally verified miRNA–target interactions include: miR2Disease [86], OncomiRDB [87], miRecords [36], miRTarBase [37] and TarBase [88]. Information in these databases is manually curated or derived through text mining as in the case of TarBase and miRTarBase.

miRCancer is a recently released database that follows a rule-based text-mining approach to extract miRNA–cancer associations from the literature [89]. 878 relationships between 236 miRNAs and 79 human cancers have been found in this way. miRWalk is a composite database that hosts data of both predicted and validated miRNA–target interactions [41]. Validated miRNA–target interactions contained in the miRWalk database were derived through text-mining MedLine abstracts and therefore should be used with caution. Interestingly, miRWalk is one of a few databases that allow miRNA target sites searches not only in gene 3′ UTR sequences but also in the 5′ UTR and open reading frame sequences of known genes. The TriplexRNA database contains predictions of cooperatively acting miRNAs and their mutual target mRNAs [90]. The RNA triplexes composed of two miRNAs and a mutual target were derived through a comprehensive workflow involving 2D and 3D structure predictions, molecular dynamics simulations, and kinetic modeling for the prediction of target repression efficiency. Another web-accessible database is TSmiR, which contains interaction maps of miRNA transcription factors and tissue-specific miRNAs from experimentally validated and predicted data [91]. It currently contains 2347 transcription factor–miRNA regulatory relations involving 116 miRNAs and 101 transcription factors in 12 tissues.

The miRSeI, a miRNA–gene association database incorporates information through text-mining and combined the results with existing databases and computational predictions. It provides a web interface for users to retrieve information on miRNA–gene pairs stored in the database [92].

miRGate provides validated and predicted miRNA targets for human, mouse and rat. The most recent release includes a complete and updated dataset of miRNAs and all variants of every gene providing high confidence predictions [93]. In Table 3 some of the most commonly used miRNA web resources are listed.

Table 3
Overview of miRNA web resources

Resource	Description	URL
Databases of validated miRNA targets		
miRTarBase	Experimentally verified miRNA target interactions	http://miRTarBase.mbc.nctu.edu.tw/
TarBase	Experimentally supported animal miRNA targets database	http://www.microrna.gr/tarbase
Databases of predicted and validated miRNA targets		
miRecords	Predicted and validated miRNA–target interactions in animals	http://c1.accurascience.com/miRecords/
miRWalk	Predicted and validated miRNA targets	http://mirwalk.uni-hd.de/
starbase	MiRNA–target interaction networks from Argonaute CLIP-Seq (HITS-CLIP) and degradome sequencing	http://starbase.sysu.edu.cn/
miRGate	a curated database of human, mouse, and rat miRNA–mRNA targets	http://mirgate.bioinfo.cnio.es
miRSEL	Text-mining based identification of miRNA–target interactions	http://services.bio.ifi.lmu.de/mirsel
Databases of disease-related miRNA–target interactions		
miR2Disease	Human miRNA diseases associations	www.miR2Disease.org
OncomiRDB	Experimentally verified oncogenic and tumor-suppressive miRNAs	http://bioinfo.au.tsinghua.edu.cn/oncomirdb/
miRCancer	Literature-derived miRNA–cancer associations	http://mircancer.ecu.edu/
Other miRNA-related databases		
miRò	MiRNA–phenotype associations in humans	http://ferrolab.dmi.unict.it/miro
TriplexRNA	Information on predicted RNA triplexes composed of two cooperatively acting miRNAs and their mutual target mRNAs	www.sbi.uni-rostock.de/triplexrna
TSmiR	Information on interaction maps of transcription factor and tissue-specific miRNAs from experimentally validated and predicted data	http://bioeng.swjtu.edu.cn/TSmiR
mimiRNA	A miRNA expression profiling and classification resource designed to identify functional correlations between miRNAs and their targets	http://mimirna.centenary.org.au
miREval 2.0	A search tool for the detection of novel miRNAs.	http://mimirna.centenary.org.au/mireval/

5 Conclusion

The importance of posttranscriptional gene regulation by miRNAs is well accepted. However, due to the complexity underlying miRNA–mRNA interactions, currently available target prediction algorithms still generate high rates of false positive and false negative predictions. Thus, there is a need to improve the current approaches. Technical support may come from the advancements in NGS technologies. High-throughput sRNA sequencing data provides an important source for understanding the characteristic features of miRNAs, and their function in gene regulation.

In this chapter we emphasized the tissue-specificity of miRNA function, suggesting this may explain some of the uncertainties in miRNA target predictions. To date, this issue is neglected by most algorithms, the only exception being CRCmiRTar, an algorithm that makes reliable predictions for miRNA targets in CRC. CRCmiRTar proves that well-chosen tissue-specific training data is the key for effective performance of miRNA target prediction algorithms. These could be further improved by also incorporating information about APAs and SNPs in the respective tissue under investigation. With the help of advanced NGS techniques, and therefore the ability to profile whole transcriptomes at base-resolution, we now have the opportunity to predict miRNA targets at an isoform level. This in turn allows us to determine miRNA targetomes significant for specific tissues or diseases.

References

1. Wen J, Parker BJ, Jacobsen A, Krogh A (2011) MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA* 17(5):820–834
2. J-i S, Tabunoki H (2011) Comprehensive analysis of human microRNA target networks. *BioData Min* 4:17–17
3. Garzon R, Marcucci G, Croce CM (2010) Targeting microRNAs in cancer: rationale, strategies and challenges. *Nat Rev Drug Discov* 9(10):775–789
4. Schmitz U, Naderi-Meshkin H, Gupta SK, Wolkenhauer O, Vera J (2016) The RNA world in the 21st century—a systems approach to finding non-coding keys to clinical questions. *Brief Bioinform* 17(3):380–392
5. 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
6. Su X, Xing J, Wang Z, Chen L, Cui M, Jiang B (2013) microRNAs and ceRNAs: RNA networks in pathogenesis of cancer. *Chin J Cancer Res* 25(2):235–239
7. Schmitz U, Wolkenhauer O, Vera J (2013) MicroRNA cancer regulation: advanced concepts, bioinformatics and systems biology tools, vol 774. Springer Science & Business Media, Dordrecht
8. Amirkhah R, Schmitz U, Linnebacher M, Wolkenhauer O, Farazmand A (2015) MicroRNA-mRNA interactions in colorectal cancer and their role in tumor progression. *Genes Chromosomes Cancer* 54(3):129–141
9. Clark PM, Loher P, Quann K, Brody J, Londin ER, Rigoutsos I (2014) Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. *Sci Rep* 4:5947
10. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* 310:1817–1821
11. Kowarsch A, Preusse M, Marr C, Theis FJ (2011) miTALOS: Analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA* 17(5):809–819

12. Preusse M, Theis FJ, Mueller NS (2016) miTALOS v2: analyzing tissue specific microRNA function. *PLoS One* 11(3):e0151771
13. Ritchie W, Rajasekhar M, Flamant S, Rasko JEJ (2009) Conserved expression patterns predict microRNA targets. *PLoS Comput Biol* 5(9):e1000513
14. Agarwal V, Bell GW, Nam J-W, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005
15. J-y Z, Wang F, Li Y, X-b Z, Yang L, Wang W, Xu H, D-z L, L-y Z (2015) Five miRNAs considered as molecular targets for predicting esophageal cancer. *Med Sci Monit* 21: 3222–3230
16. Pizzini S, Bisognin A, Mandruzzato S, Biasiolo M, Faccioli A, Perilli L, Rossi E, Esposito G, Rugge M, Pilati P, Mocellin S, Nitti D, Bortoluzzi S, Zanovello P (2013) Impact of microRNAs on regulatory networks and pathways in human colorectal carcinogenesis and development of metastasis. *BMC Genomics* 14:589
17. Li X, Gill R, Cooper NG, Yoo JK, Datta S (2011) Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. *BMC Med Genomics* 4:44
18. Ritchie W, Flamant S, Rasko JE (2010) mim-iRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. *Bioinformatics* 26(2):223–227
19. Wu X, Watson M (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* 25(6):832–833
20. Bhattacharya A, Cui Y (2015) miR2GO: comparative functional analysis for microRNAs. *Bioinformatics* 31(14):2403–2405
21. Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res* 43(W1):W460–W466
22. Ritchie W, Rasko JE (2014) Refining microRNA target predictions: sorting the wheat from the chaff. *Biochem Biophys Res Commun* 445(4):780–784
23. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136(2): 215–233
24. Ritchie W, Flamant S, Rasko JE (2009) Predicting microRNA targets and functions: traps for the unwary. *Nat Methods* 6:397–398
25. Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of MicroRNA-Target Recognition *PLoS Biol* 3 (3), e85.
26. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120(1): 15–20
27. Reczko M, Maragkakis M, Alexiou P, Papadopoulos GL, Hatzigeorgiou AG (2011) Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. *Front Genet* 2:103
28. Legendre M, Ritchie W, Lopez F, Gautheret D (2006) Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput Biol* 2(5):e43
29. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* 10(10):1507–1517
30. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39(10):1278–1284
31. Yue D, Liu H, Huang Y (2009) Survey of computational algorithms for microRNA target prediction. *Curr Genomics* 10(7):478–492
32. Ritchie W, Gao D, Rasko JE (2012) Defining and providing robust controls for microRNA prediction. *Bioinformatics* 28(8):1058–1061
33. Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK (2007) Naive bayes for microRNA target predictions--machine learning for microRNA targets. *Bioinformatics* 23(22): 2987–2992
34. Huang J, Lu J, Ling CX (2003) Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. *Third IEEE International Conference on Data Mining*, 553–556
35. Liu H, Yue D, Chen Y, Gao SJ, Huang Y (2010) Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics* 11:476
36. 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:D105–D110
37. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ, Tsai TR, Ho SY, Jian TY, Wu HY, Chen PR, Lin NC, Huang HT, Yang TL, Pai CY, Tai CS, Chen WL, Huang CY, Liu CC, Weng SL, Liao KW, Hsu WL, Huang HD (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* 44:D239–D247
38. Menor M, Ching T, Zhu X, Garmire D, Garmire LX (2014) mirMark: a site-level and

- UTR-level classifier for miRNA target prediction. *Genome Biol* 15:500
39. Hammell M (2010) Computational methods to identify miRNA targets. *Semin Cell Dev Biol* 21:738–744
 40. Laganà A, Forte S, Giudice A, Arena MR, Puglisi PL, Giugno R, Pulvirenti A, Shasha D, Ferro A (2009) miRò: a miRNA knowledge base. *Database (Oxford)* 2009:bap008
 41. Dweep H, Gretz N (2015) miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat Methods* 12:697
 42. Afonso-Grünz F, Müller S (2015) Principles of miRNA-mRNA interactions: beyond sequence complementarity. *Cell Mol Life Sci* 72(16):3127–3141
 43. Amirkhah R, Farazmand A, Gupta SK, Ahmadi H, Wolkenhauer O, Schmitz U (2015) Naive Bayes classifier predicts functional microRNA target interactions in colorectal cancer. *Mol Biosyst* 11(8):2126–2134
 44. Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT (2006) miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* 7:411
 45. Zheng H, Fu R, Wang JT, Liu Q, Chen H, Jiang SW (2013) Advances in the techniques for the prediction of microRNA targets. *Int J Mol Sci* 14(4):8179–8187
 46. Gumienny R, Zavolan M (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res* 43:9095
 47. Ghoshal A, Shankar R, Bagchi S, Grama A, Chatterji S (2015) MicroRNA target prediction using thermodynamic and sequence curves. *BMC Genomics* 16:999
 48. Grimson A, Farh KK, Johnston WK, Garrett-Engle P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 27(1):91–105
 49. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320:1643–1647
 50. Beuvink I, Kolb FA, Budach W, Garnier A, Lange J, Natt F, Dengler U, Hall J, Filipowicz W, Weiler J (2007) A novel microarray approach reveals new tissue-specific signatures of known and predicted mammalian microRNAs. *Nucleic Acids Res* 35:e52
 51. Bandyopadhyay S, Ghosh D, Mitra R, Zhao Z (2015) MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. *Sci Rep* 5:8004
 52. Ciafrè SA, Galardi S (2013) microRNAs and RNA-binding proteins: a complex network of interactions and reciprocal regulations in cancer. *RNA Biol* 10(6):934–942
 53. Nam JW, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, Yildirim MA, Rodriguez A, Bartel DP (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol Cell* 53(6):1031–1043
 54. Deng N, Puetter A, Zhang K, Johnson K, Zhao Z, Taylor C, Flemington EK, Zhu D (2011) Isoform-level microRNA-155 target prediction using RNA-seq. *Nucleic Acids Res* 39(9):e61
 55. Bandyopadhyay S, Mitra R (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* 25(20):2625–2631
 56. Motameny S, Wolters S, Nurnberg P, Schumacher B (2010) Next generation sequencing of miRNAs - strategies, resources and methods. *Genes (Basel)* 1(1):70–84
 57. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80
 58. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
 59. Kaushik A, Saraf S, Mukherjee SK, Gupta D (2015) miRMOD: a tool for identification and analysis of 5' and 3' miRNA modifications in Next Generation Sequencing small RNA data. *Peer J* 3:e1332
 60. Katoh T, Sakaguchi Y, Miyauchi K, Suzuki T, Kashiwabara S, Baba T (2009) Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes Dev* 23(4):433–438
 61. An J, Lai J, Lehman ML, Nelson CC (2012) miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res* 41(2):727–737
 62. Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* 37:W68–W76
 63. Ronen R, Gan I, Modai S, Sukachev A, Dror G, Halperin E, Shomron N (2010) miRNAkey:

- a software for microRNA deep sequencing analysis. *Bioinformatics* 26(20):2615–2616
64. Gao D, Middleton R, Rasko JE, Ritchie W (2013) miREval 2.0: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics* 29(24):3225–3226
 65. Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS (2009) miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* 10:328
 66. Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, Cooke HJ, Xue Y, Shi Q (2012) CPSS: a computational platform for the analysis of small RNA deep sequencing data. *Bioinformatics* 28:1925–1927
 67. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 28:2059–2061
 68. Rueda A, Barturen G, Lebron R, Gomez-Martin C, Alganza A, Oliver JL, Hackenberg M (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res* 43:W467–W473
 69. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479–486
 70. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141:129–141
 71. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153:654–665
 72. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 17:909–915
 73. Imig J, Brunschweiler A, Brummer A, Guennewig B, Mittal N, Kishore S, Tsikrika P, Gerber AP, Zavolan M, Hall J (2015) miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nat Chem Biol* 11:107–114
 74. Khorshid M, Rodak C, Zavolan M (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res* 39:D245–D252
 75. Li JH, Liu S, Zhou H, Qu LH, Yang JH (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 42:D92–D97
 76. Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A (2015) DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* 43:D160–D167
 77. Chen B, Yun J, Kim MS, Mendell JT, Xie Y (2014) PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol* 15:R18
 78. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, Hannon GJ, Sanford JR, Penalva LO, Smith AD (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* 28:3013–3020
 79. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* 12:R79
 80. Chou CH, Lin FM, Chou MT, Hsu SD, Chang TH, Weng SL, Shrestha S, Hsiao CC, Hung JH, Huang HD (2013) A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. *BMC Genomics* 14:S2
 81. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* 8:69
 82. Jin Y, Chen Z, Liu X, Zhou X (2013) Evaluating the microRNA targeting sites by luciferase reporter gene assay. *Methods Mol Biol* 936:117–127
 83. Gäken J, Mohamedali AM, Jiang J, Malik F, Stangl D, Smith AE, Chronis C, Kulasekararaj AG, Thomas NSB, Farzaneh F, Tavassoli M, Mufti GJ (2012) A functional assay for microRNA target identification and validation. *Nucleic Acids Res* 40:e75–e75
 84. Kuhn DE, Martin MM, Feldman DS, Terry AV Jr, Nuovo GJ, Elton TS (2008) Experimental validation of miRNA targets. *Methods* 44:47–54
 85. Watanabe Y, Tomita M, Kanai A (2007) Computational methods for microRNA target prediction. *Methods Enzymol* 427:65–86
 86. 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:D98–104
 87. Wang D, Gu J, Wang T, Ding Z (2014) OncomiRDB: a database for the experimentally

- verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics* 30:2237–2238
88. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 40:D222–D229
 89. Xie B, Ding Q, Han H, Wu D (2013) miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics* 29:638–644
 90. Schmitz U, Lai X, Winter F, Wolkenhauer O, Vera J, Gupta SK (2014) Cooperative gene regulation by microRNA pairs and their identification using a computational workflow. *Nucleic Acids Res* 42:7539–7552
 91. 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.
 92. Naeem H, Kuffner R, Csaba G, Zimmer R (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics* 11:135
 93. Andrés-León E, Peña DG, Gómez-López G, Pisano DG (2015) miRGate: a curated database of human, mouse and rat miRNA–mRNA targets. *Database* 2015:bav035

Chapter 12

sRNAtoolboxVM: Small RNA Analysis in a Virtual Machine

Cristina Gómez-Martín, Ricardo Lebrón, Antonio Rueda, José L. Oliver, and Michael Hackenberg

Abstract

High-throughput sequencing (HTS) data for small RNAs (noncoding RNA molecules that are 20–250 nucleotides in length) can now be routinely generated by minimally equipped wet laboratories; however, the bottleneck in HTS-based research has shifted now to the analysis of such huge amount of data. One of the reasons is that many analysis types require a Linux environment but computers, system administrators, and bioinformaticians suppose additional costs that often cannot be afforded by small to mid-sized groups or laboratories. Web servers are an alternative that can be used if the data is not subjected to privacy issues (what very often is an important issue with medical data). However, in any case they are less flexible than stand-alone programs limiting the number of workflows and analysis types that can be carried out.

We show in this protocol how virtual machines can be used to overcome those problems and limitations. sRNAtoolboxVM is a virtual machine that can be executed on all common operating systems through virtualization programs like VirtualBox or VMware, providing the user with a high number of preinstalled programs like sRNAbench for small RNA analysis without the need to maintain additional servers and/or operating systems.

Key words Small RNA, Bioinformatics, Next generation sequencing, Virtual machine

1 Introduction

Given the broad interest in small RNA molecules existing in many different fields, the primary and downstream analysis of small RNA sequencing data is nowadays a highly demanded task. Over the last 15 years many basic research has been conducted to understand the functions that microRNAs can play in numerous vital processes like cell differentiation [1] and its role in conferring robustness to biological processes [2]. More recently, microRNAs have exited expectations both as prognostic and diagnostic biomarkers and therapeutic agent in medicine [3] or as transgenes in plant science [4]. Furthermore, they seem to play vital roles in host–pathogen or host–symbiont interactions [5, 6].

Over the last few years, the generation of small RNA sequencing data has become constantly cheaper and accessible to a broader

audience. Consequently the generated data volume is continually rising converting the data analysis as the main task in HTS-based small RNA research. There are several reasons why the data analysis is currently the main bottleneck in HTS-based research. First, many important programs like samtools [7] are only available for Linux operating systems or would need to be compiled by the user in order to make them run on other OS like Windows. This however is frequently not a straight forward process and requires skilled personal. The asset costs and administration of Linux machines together with the need for personal trained in HTS-data analysis raises the cost for HTS-based research enormously bringing especially smaller groups or laboratories into difficulties. Web servers like miRanalyzer [8] can be used alternatively for basic small RNA data analysis; however, the usage of web servers in general is very often prohibitive due to privacy issues.

In this protocol, we show how virtual machines can be a solution to the above outlined problems. Basically a virtual machine (VM) is an emulation of a given computer system. In this way, a Linux OS can be easily executed within a windows environment. Those virtual Linux machines can then be prepared for its usage (like small RNA analysis) by a developing team lifting from the user the burden to administer a separate Linux OS and to install all needed programs. Furthermore, the maintenance is likewise carried out by the developer team and the user needs to download only the newest versions without the need for specialized personal in Linux administration and maintenance.

We show the strength of this concept by means of the virtualization of sRNAtoolbox [9]. This collection of programs for small RNA research that includes programs like sRNAbench [10] allows for many important analysis types like expression profiling from NGS data, prediction of novel microRNAs, detection of other small RNAs, differential expression of different RNA types and several downstream analyses like target gene prediction and functional analysis of those genes (pathways, GO-terms). Furthermore, sRNAtoolboxVM includes helper tools that automatically populate the local database and keep the machine up to date.

In this chapter we show how to run the virtual machine on any OS, how to customize the sRNAtoolboxVM, i.e., how to install the species annotations needed by the user, and how to carry out a basic HTS data analysis.

2 Install sRNAtoolboxVM

2.1 Requirements

- Supported Operating Systems: Windows (7 or higher), Mac OS X (10.9 or higher), Linux, or Solaris.
- **Virtualbox** (5 or higher) and **Virtualbox Extension Pack**: <https://www.virtualbox.org/wiki/Downloads>. Official User Manual: <https://www.virtualbox.org/manual/UserManual.html>.

- Once imported, select the virtual machine and click on Start (green arrow).
- Now the virtual machine is running (Fig. 1c).
- The default user is “srna” and all relevant passwords are also “srna.” The password can be changed by the user if needed. However, note that this is not a security issue as the VM can only be accessed from the host machine.
- All parameters can be changed any time by means of the “Settings” tab (see Fig. 1b) when the VM is off.
- To shutdown Click on ACPI Shutdown.

2.3 Share Data with the Virtual Machine

An important issue is how to pass data back and forth between the host and the virtual machine. There are basically two possibilities.

- Share a folder between the host and the virtual machine. Click on **Settings** (Fig. 1b) and then on Shared Folders tab. The shared folder will appear in the user's home directory of the VM.
- Use a Dropbox client inside the virtual machine which can be easily installed by means of the “Ubuntu Software Centre.”

3 Customize and Populate the Local Database

Some of the installed programs rely on a local database. This database holds for example the bowtie indexes of the genome assemblies, microRNA sequences or other small RNA annotations among other required data. By default, the VM is shipped without this data as otherwise it would grow too big. Therefore the first step will be to populate the database with the species needed by the user. The local database is installed in the following path: “/opt/sRNAtoolboxDB”.

3.1 Populate and Update Tools

The easiest way to populate the local database and to maintain the VM updated (install the most recent versions of the programs) is by means of two small tools installed in the VM. On the desktop, two links can be found, “Populate DB” and “Update DB” (Fig. 1c).

To populate the local database with any of the species contained in our sRNAtoolbox database:

- Double-click on the icon “Populate DB.”
- Select the species that should be downloaded by activating the “check-box”: for the following protocol, please choose “hg38_prim” (this is the most recent, primary human assembly, i.e., without alternative sequences).
- Already installed species will appear as “checked”; those can be removed from the local DB by unchecking the boxes.
- Click on “ok” to start the processes, which will take a while depending on network conditions.

Important: after installing a species, the file `info.txt` on the desktop will contain the information about the names of the bowtie indexes of the genome assemblies that have to be used for the profiling (*see below*) and the names of the available annotations (Fig. 1d).

Update the sRNAtoolbox programs:

- Double-click on the “Update DB” icon.
- First, the program will check if a new version of the entire sRNAtoolboxVM exists. If a new version is available, the corresponding download link will be provided.
- Second, the update program will check if for any of the sRNA-toolbox programs a new version exists. If so, those will be downloaded and the old versions will be replaced.
- A short summary of the whole process is shown to the user including a brief description of the changes.

Important: perform an update after importing sRNAtoolbox VM!

Note that the local sRNAtoolbox database can be populated also manually. Later in this chapter we give one example on how to do this.

3.2 Small RNA Analysis

The following practical protocol for small RNA data analysis is meant for users with only very basic knowledge in Linux and analysis of HTS data. The practical session consists in a number of different analyses which are carried out by running different programs in the Linux command line. Those commands will be highlighted in **grey shaded boxes** and they can be found also in the file “`protocol.txt`” on the desktop. The commands can be directly copy/paste from this file into the terminal which can be opened by clicking on the “Terminal” icon on the desktop. Finally, most of the programs and analysis types are highly parameterized. In the grey shaded “Boxes” located at the end of many sections, we briefly discuss those parameters.

3.3 Retrieving Data

To illustrate the basic features of sRNAtoolboxVM, we use a publically available dataset from mucosal epithelium of the nasopharynx with two different states: nasopharyngeal carcinoma and chronic nasopharyngitis [11]. First we need to retrieve this data from the public repository SRA and to store it into a previously generated folder (in the `protocol.txt` file, this is shown step by step). We recommend to name this folder exactly as proposed in the `protocol.txt` file which will allow copy/paste all following commands.

After having navigated into the data folder (with “`cd`” command), the following commands are executed to download a subset of the whole study which will be used to illustrate the function of sRNAtoolboxVM (note that they can be pasted together into the terminal and will be processed subsequently one after one:

```
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105948/SRR2105948.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105949/SRR2105949.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105950/SRR2105950.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105951/SRR2105951.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105941/SRR2105941.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105940/SRR2105940.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105939/SRR2105939.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/By
Run/sra/SRR/SRR210/SRR2105938/SRR2105938.sra
```

After downloading the data files in SRA format, those have to be converted into fastq format by means of the fastq-dump program of the SRA Toolkit [12]. In order to save space, we will convert them into compressed files. Note that the “*” will apply the program to all file names starting with “SRR”.

```
fastq-dump --gzip SRR*
```

After the conversion process, the original SRA files (those with sra extension) can be removed with the command:

```
rm *.sra
```

3.4 Preprocessing of the Data

A crucial step in the analysis of high-throughput small RNA sequencing data is the detection of the 3' adapter sequence. Many small RNAs (or RNA fragments) have lengths between 21 and 22 nt (mature microRNAs) and 33 nt (tRNA halves), and therefore the adapter or part of it will be sequenced as well at nowadays typical read lengths of 36–100 nt. As a consequence, the preprocessing consists normally of two basic steps: (1) adapter trimming (detect and remove the adapter sequence) and (2) read collapsing (determine the unique reads and assign a read count to them, i.e., the number of times they have been sequenced). The following command will perform the preprocessing generating additionally length distribution files.

```
sRNAbench input=/home/srna/data/nasopharyngeal_carci-
noma/SRR2105938.fastq.gz output=/home/srna/results/
SRR2105938 adapter=TGGAATTCTCGGGTGCCAAGG
graphics=true
```

3.4.1 Explanation of the Command

- **input=“filename path”**: several input formats are accepted including fastq, sra, fasta, and read/count as well as those formats compressed with the gzip algorithm (e.g., “gz” extension).
- **output=“folder”**: the output folder. If this parameter is omitted, then the output will be written into /opt/sRNAtoolboxDB/out/.
- **adapter=“sequence”**: the adapter sequence that should be trimmed from the reads. If this parameter is not used, then the program assumes adapter-trimmed input.
- **graphics=true**: the length distributions are plotted and the graphics are copied into the “graphs” folder (subfolder of the process output folder)

3.4.2 Results of the Command

- The adapter-trimmed reads can be found in the file “reads_orig.fa”. Those reads will be used for the subsequent analyses.
- In the “stat” folder, two files reflecting the read length distribution can be found
 - “readLengthFull.txt”, the length distribution of all raw reads after adapter trimming.
 - “readLengthAnalysis.txt” the length distribution of all adapter-trimmed reads in the analysis (i.e., after quality and length filtering).

Box 1: Preprocessing

Several parameters can influence the results of this step:

- **holdNonAdapter=true**: the reads for which the adapter was not found are **also** used for the analysis. This might be indicated if many RNA molecules longer than the read length can be expected.
- **adapterMM (default 1) and adapterMinLength (default 10)**: manipulate the stringency of adapter detection, i.e., more stringent if the number of mismatches is lower (adapterMM) and the detected adapter length is longer (adapterMinLength).
- **removeBarcode=“number of bases”**: the given number of nucleotides will be clipped of the 5’ end of the read (if a barcode sequence was used to distinguish between samples)
- **remove3pBases=“number of bases”**: after adapter trimming, the given number of nucleotides is trimmed of the 3’ end of the read. This might be important if random adapters are used to avoid sequencing bias.
- **guessAdapter=true**: if the adapter is not known, sRNAbench will try to guess it.

Box 1 (continued)

- **recursiveAdapterTrimming=true**: The adapter is recursively detected and trimmed. If at least **adapterMinLength** bases of the adapter are not found within the read sequence, then the adapter is recursively detected only at the 3' end of the read.
- **minRC="integer"**: a read must have at least the specified read count in order to be used in the analysis. For example **minRC=2** deletes all singleton reads.
- **minReadLength** and **maxReadLength** allow to set the maximum and minimum read length. Reads outside this range are filtered out (default **minReadLength=15**)
- **sep="separator string"**: if the input is in fasta format, by default sRNAbench expects this format: **>ID#300**, being the number after the **"#"** the read count. By means of the **sep=** parameter, other separators can be chosen. Note that this is necessary if other programs are used for adapter trimming and collapsing.

**3.5 Profile
the Expression
of miRBase
MicroRNAs**

First we see how we can determine the expression of known microRNAs. In general, to profile the expression of known RNAs we will require its reference sequences. By default, sRNAbench uses microRNA annotations from miRBase [13]. The reference precursor ("hairpin.fa") and mature ("mature.fa") sequences must be located in the **"libs"** folder of the local sRNAtoolbox database ("/opt/sRNAtoolboxDB/libs"). The VM contains the sequences of release 21. If miRBase releases a new version, both files can be downloaded from here: <http://mirbase.org/ftp.shtml>

```
sRNAbench input=/home/srna/data/nasopharyngeal_carcinoma/SRR2105938.fastq.gz output=/home/srna/results/SRR2105938_miRBase adapter=TGGAATTCTCGGGTGCCAAGGG graphics=true microRNA=hsa
```

**3.5.1 Explanation
of the Command**

- **microRNA="short species name"**: In miRBase, the nomenclature makes reference to the species in the first 3–4 letters. For example, **hsa-mir-21** refers to microRNA 21 in *Homo sapiens* while **mmu-mir-21a** would be the homologous microRNA in *Mus musculus*. This first part of the microRNA name is used to parse out the reference sequences of the species that should be analysed.

**3.5.2 Results
of the Command**

- Several expression files that contain the information shown in Table 1 (the **"*.grouped"** files)
- **"mature_sense.grouped"** holds the expression values of the mature sequences (see Fig. 2a). The format is the same for all expression files and has the columns shown in Table 1.

Table 1
The file format of all expression files (i.e., *.grouped files)

name	The name of the reference sequence.
unique reads	The number of unique reads mapped, i.e., in the case of microRNAs these sequences are isomiRs.
read count	The total number of reads mapped.
read count (mult. map. adj.)	Read count adjusted for multiple mappings, i.e., if a read maps to N different reference sequences (or genome loci), the read count is divided by N.
RPM (lib)	The Read Per Million expression value using all reads mapped to this library for normalization. Note, the RPM value is calculated using the “read count” and not the adjusted number.
RPM (total)	The Read Per Million expression value using all genome mapped reads (genome mode) or all adapter cleaned reads (library mode) for normalization. Note, the RPM value is calculated using the “read count” and not the adjusted number.
chromosomeString	In genome mode, this string reflects the chromosome coordinates and the name of the reference sequence, while in library mode, the name of the reference sequence and the coordinates (within the reference sequence) are shown.

- “mature_sense_SA.grouped”; **the single assignment expression file**. Multiple mapping of reads is a well-known problem in HTS data analysis. MicroRNAs are frequently members of broader families that contain several genes with highly similar mature sequences. sRNAbench addresses this problem in two ways: (1) a simple adjusted expression value is calculated dividing the read count of each read by the number of reference sequences or genome loci to which they map (Fig. 2a) and (2) by determining a “single assignment” expression value, i.e., each read is only assigned to one sequence or loci, i.e., to the one with the highest expression value (for details please *see* ref. [10]). Therefore, for each expression file, like “mature_sense.grouped”, one with “single assignment” expression values are generated.
- “hairpin_sense.grouped”; the expression of the precursor sequences including the mature sequences.
- “hairpin_sense_SA.grouped”; the expression values of this file are based only on those reads that map to the precursor, BUT NOT to the mature sequences (*see* Fig. 2b). This is because in the single assignment, the reads are first assigned to the mature sequences and then to the precursors. For most microRNAs, the number of reads uniquely mapped to the precursor but not to the mature sequence should be close to zero. Therefore, those precursors with high number of uniquely mapped reads could be incorrect microRNAs (false positive annotations) (*see* Fig. 2c).

a)

name	unique reads	read count	read count (mult. map. adj.)	RPM (lib)	RPM (total)	chromosomeString
hsa-miR-143-3p	771	726385	726385	155648.61	38233.82	hsa-mir-143,61,81,+
hsa-miR-182-5p	1516	708389	708389	151792.46	37286.59	hsa-mir-182,23,46,+
hsa-miR-10b-5p	1109	539712	539628.5	115648.62	28408.15	hsa-mir-10b,27,49,+
hsa-miR-21-5p	733	273726	273726	58653.57	14407.77	hsa-mir-21,8,29,+
hsa-miR-92a-3p	1486	251623	131707.17	53917.37	13244.37	hsa-mir-92a-1,48,69,+
hsa-miR-181a-5p	606	250138	125072.5	53599.17	13166.2	hsa-mir-181a-1,24,46,+
hsa-miR-181a-5p	605	250129	125063.5	53597.24	13165.73	hsa-mir-181a-2,39,61,+
hsa-miR-92a-3p	1264	241260	121344.17	51696.81	12698.9	hsa-mir-92a-2,48,69,+
hsa-let-7f-5p	388	233211	122545.02	49972.08	12275.24	hsa-let-7f-2,8,29,+
hsa-let-7f-5p	329	221379	110723.82	47436.74	11652.45	hsa-let-7f-1,7,28,+
hsa-miR-148a-3p	528	141574	141469	30336.25	7451.85	hsa-mir-148a,44,65,+
hsa-miR-27b-3p	414	126218	125898	27045.79	6643.58	hsa-mir-27b,61,81,+
hsa-miR-92b-3p	1296	99778	99760.67	21380.27	5251.89	hsa-mir-92b,61,82,+

b)

name	unique reads	read count(SA)	read count (MA)	RPM (lib)	RPM (total)	chromosomeString
hsa-mir-7641-2	290	4043	4213	863.402395	212.806353	hsa-mir-7641-2,1,53,+
hsa-mir-1303	66	1757	1768	375.215931	92.4810194	hsa-mir-1303,1,86,+
hsa-mir-6087	55	1199	1199	256.052306	63.1102688	hsa-mir-6087,1,49,+
hsa-mir-5096	74	712	837	152.05111	37.4766567	hsa-mir-5096,1,70,+
hsa-mir-7641-1	33	569	944	121.512729	29.9497439	hsa-mir-7641-1,1,61,+
hsa-mir-1273e	63	535	545	114.251863	28.1601283	hsa-mir-1273e,1,102,+
hsa-mir-182	32	477	708866	101.86568	25.1072546	hsa-mir-182,1,110,+
hsa-mir-3960	40	436	515	93.1099293	22.9491887	hsa-mir-3960,1,91,+
hsa-mir-1248	22	375	375	80.0830985	19.7384077	hsa-mir-1248,1,106,+

c)

GTTTGATCTCGGAAGCTAAGC	AGGGTCGGGCTGGTGTAGTACTTGGATGGGAG	4,213(221.8/902.8)	adj:4,087.5	has-mir-7641-2
(.((((.(.((.(.(((.(.(.....)).)))))).)).))).)))))				
ATCTCGGAAGCTAAGCAGGGTCGGGCCCTG		149(7.8/31.9)	adj:149	+
CGGAAGCTAAGCAGGGTCGGGCCCTGGTT		134(7.1/28.7)	adj:134	+
CGGAAGCTAAGCAGGGTCGGGCCCT		124(6.5/26.6)	adj:124	+
	TCGGGCCCTGGTGTAGTACT	115(6.1/24.6)	adj:115	+
TCTCGGAAGCTAAGCAGGGTCGGGCCCT		101(5.3/21.6)	adj:101	+
	CGGGCCCTGGTGTAGTACTTGA	89(4.7/19.1)	adj:89	+
CGGAAGCTAAGCAGGGTCGGGCCCTGGT		87(4.6/18.6)	adj:87	+
	CGGGCCCTGGTGTAGTACT	85(4.5/18.2)	adj:85	+
AAGCAGGGTCGGGCCCTGGTGTAGTACT		82(4.3/17.6)	adj:82	+
	TCGGGCCCTGGTGTAGTACTTGA	82(4.3/17.6)	adj:82	+
GCAGGGTCGGGCCCTGGTGTAGTACTTGA		81(4.3/17.4)	adj:81	+

Fig. 2 sRNA expression and alignment files. **(a)** Excerpt from a typical *mature_sense.grouped* file. The read count and the adjusted read count will have same values if none of the reads maps to several reference sequences like those marked with the red box. The green box shows *hsa-miR-181a-5p* which can be obtained from two different gene copies (181a-1 and 181a-2) and therefore the adjusted read count is approximately half of the read count. **(b)** Excerpt of a typical *hairpin_sense_SA.grouped* file. In the single assignment files (*_SA*), instead of the adjusted read count column, the single assignment read count is shown (read count (SA)). Note that the read count (MA) is the total read count (without adjusting or single assignment). *hsa-mir-6087* (red box) has the same read count for MA (multiple assignment) and SA. This means, that none of the reads mapped to the precursor was assigned to the mature sequence. Marked with the green box we can see that for *mir-182*, 708866 reads are mapped to the precursor sequences and most of them except 477 are assigned to the mature sequences. **(c)** The alignment file for *hsa-mir-7641-2* which shows that none of the aligned reads coincides with the mature sequence (*red* subsequence of the precursor sequence). This might indicate that this is an incorrect microRNA

- “hairpin” folder; which holds two types of files: i) “*.align” files that show all reads mapped to the sequence (Fig. 2c) and “*.countArray”, which gives the percentage of read coverage based on unique reads (UR) and total read count (RC) and the RPM (read per million) expression per base.

3.6 Profile the Expression of “other” MicroRNA references

By default, sRNAbench takes the “mature.fa” and “hairpin.fa” files from miRBase for microRNA detection. However, other microRNA annotations can also be used like for example miRGeneDB [14]. This database proposes a uniform system for the annotation and

Box 2: Profiling of Known MicroRNAs

There are several alignment parameters that can influence the results.

noMM=integer: the number of allowed mismatches (default = 0).

alignType=[v/n]: v (the whole read is aligned), n (only the first L nucleotides are aligned, i.e., the seed).

seed=integer: the length of the seed if seed alignment is chosen (alignType=n).

winUpMiR="integer" and winDownMiR="integer": these parameters define the windows around the reference mature sequence used to assign a read to this mature microRNA. That means that a read is assigned to a mature sequence if it starts at most **winUpMiR** nt upstream and ends at most **winDownMiR** downstream of the reference sequence.

isoMiR=true: the program performs the classification of isoMiRs and generates an expression file only for the canonical sequences (canonical.txt). If isoMiR=true, then **alignType=n** should be set!

nomenclature of miRNA genes together with stricter criterion on what should be considered a microRNA. The microRNA names use a very similar species nomenclature, starting however with upper case letters.

```
sRNAbench input=/home/srna/data/nasopharyngeal_carcinoma/SRR2105938.fastq.gz output=/home/srna/results/SRR2105938_miRGeneDB adapter=TGGAATTCTCGGGTG-CCAAGGG graphics=true microRNA=Hsa mature=miR-GeneDB_mature.fa hairpin=miRGeneDB_pre.fa
```

3.6.1 Explanation of the Command

- **microRNA=Hsa:** the microRNA names in miRGeneDB start with upper-case letters, i.e., we indicate the usage of human microRNAs.
- **mature=miRGeneDB_mature.fa:** sets the name of the file in the “libs” folder that contains the mature (guide and passenger) sequences.
- **hairpin=miRGeneDB_pre.fa:** the name of the file with the precursor sequences.
- **“results.txt”:** summary of the preprocessing and mapping statistics. In this case it can be used to compare some basic features between miRBase and miRGeneDB annotations.

3.6.2 Results of the Command

- Mature miRBase sequences have a total of 4,666,826 mapped reads (readsRCmatureSense=4666826) while miRGeneDB mature sequences are mapped by 4,167,527 reads (readsRCmatureSense=4167527). The miRGeneDB picks up 89% of the total miRBase expression. However, there are 2.5 times more mature sequences in miRBase release 21 (2588) compared to miRGeneDB (1,046).
- The fact that miRBase might contain a considerable number of false positives can also be seen in the number of reads mapped to the precursor but not to the mature sequence. This number is 15,811 in miRBase but only 1329 in miRGeneDB. Like mentioned before, a high number of unique precursor reads which however are not mapped to the mature sequences can indicate false positives like the one shown in Fig. 2c.

3.7 Profile MicroRNA Expression in “Genome Mode”

sRNAbench allows to map directly to the reference sequences (“library mode”) but also to map first to the genome, quantifying the expression by means of a genome annotation in bed/gff format (genome mode). The advantages and disadvantages have been discussed before [10].

```
sRNAbench input=/home/srna/data/nasopharyngeal_carcinoma/SRR2105951.fastq.gz output=/home/srna/results/SRR2105951_miRBase_genome adapter=TGGAATTCTCGGGTGCCAAGGG graphics=true microRNA=hsa species=hg38_prim_mp
```

3.7.1 Explanation of the Command

- species=“name of the bowtie index”: this parameter specifies the name of the bowtie index (of the genome assembly) that should be used to map the reads first against a genome.

3.7.2 Results of the Command

- The same expression files as when using the “library mode.”
- The file “genomeDistribution.txt” in the “stat” folder: This file shows the percentages of unique reads and read count mapped to the genome. Figure 3a shows the read count percentage of “unmapped reads,” “highly redundant reads” (hsa_HR in the graphic; i.e., reads mapped to more than 10 loci by default but see Box 3 on how to change) and mapped reads (“hsa” in the graphic). It can be seen that both, a rather high number of redundant reads and unmapped reads do exist. In the section “Characterize unmapped reads: sRNAblast” we show how to explore this finding in more detail.
- The “reads.fa” file: This file holds all unmapped reads at the end of the analysis. Note that highly redundant reads are also removed, i.e., not contained in this file.

3.7.3 Explanation of the Command

- winUpPre=“number of nt”: the flanking window upstream of the microRNA precursor that will be used to assign a read to the precursor.
- winDownPre=“number of nt”: the flanking window downstream of the microRNA precursor.

3.7.4 Results of the Command

No new files will be generated by these parameters; however, the expression counts will change. Figure 3c shows the alignment of reads to the precursor sequence of hsa-mir-21 plus the 20 nt up- and downstream flanking windows specified in the command line. **Note that the secondary structure is calculated based only on the reference sequence, i.e., without the flanking sequences.**

3.8 Set Up a Multispecies Profiling: Profile EBV Sequences

It was reported that EBV virus infection might be related to nasopharyngeal cancer [11, 16], and therefore it would be interesting to quantify the content of EBV derived molecules in the samples. Although the Epstein-Barr virus is in our local database and could be easily downloaded by means of the “Populate DB” tool, we

Box 3: Genome Mapping

By default, only reads that map at most to ten different loci are used for any further analysis. All other reads are detected as genome mapped reads but are not considered for expression profiling. This behavior can be changed by means of two parameters:

‘**bowtieReportType=-a-m**’: This parameter must be given within single quotation marks on the command line and regulates the bowtie reporting behavior. The default combination **-a -m** indicates that only reads that map at most **m** times to the genome are reported. Note that all type of parameters can be passed to bowtie, for example ‘**-k**’ would be another possibility (see the bowtie manual for more details)

bowtieReportCount=integer: this parameter specifies the corresponding number to the ‘**bowtieReportType**’ parameter. For example,

- ‘**bowtieReportType=-a -m bowtieReportCount=20**’ would pass to bowtie the following parameters ‘**-a -m 20**’
- ‘**bowtieReportType=-a -k bowtieReportCount=5**’ would pass ‘**-a -k 5**’
- If all mappings should be reported, ‘**bowtieReportType=-a**’ ‘**bowtieReportCount=**’ needs to be specified on the command line (with a space after the =)

chunkmbs=integer: the memory reserved for each chunk by bowtie. For longer and highly redundant reads, this number should be increased (default 256).

p=integer: the number of threads used by bowtie and the parallelized parts of sRNAtoolbox programs.

a)

GenBank

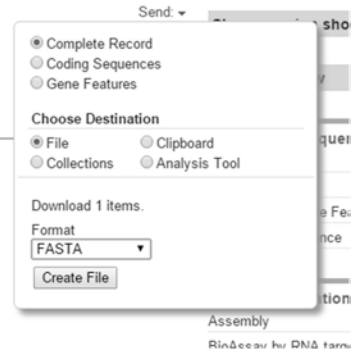
Human herpesvirus 4, complete genome

NCBI Reference Sequence: NC_009334.1

[FASTA](#) [Graphics](#)

Go to:

LOCUS NC_009334 172764 bp DNA circular VRL 26-MAR-2010
 DEFINITION Human herpesvirus 4, complete genome.
 ACCESSION NC_009334
 VERSION NC_009334.1 GI:139424470
 DBLINK BioProject: [PRJNA20959](#)
 KEYWORDS RefSeq.
 SOURCE Human herpesvirus 4 type 2 (Epstein-Barr virus type 2)
 ORGANISM Human herpesvirus 4 type 2
 Viruses; dsDNA viruses, no RNA stage; Herpesvirales; Herpesviridae;
 Gammaherpesvirinae; Lymphocryptovirus.



b)

```
>gi|139424470|ref|NC_009334.1| Human herpesvirus 4, complete genome
AGAATTTGTCTTGCTCTATTACCGTTACTTTTCTTCTTGCCCGTTCTCTTTCTTAGTATGAATCCAGTA
```

↓

```
>NC_009334:ebv Human herpesvirus 4, complete genome
AGAATTTGTCTTGCTCTATTACCGTTACTTTTCTTCTTGCCCGTTCTCTTTCTTAGTATGAATCCAGTA
```

Fig. 4 (a) Download the EBV genome sequence from the NCBI nucleotides database. (b) The fasta headers should be adapted to sRNAbench, by (1) eliminating the “|” and (2) by adding a species tag “:ebv”

show briefly how assemblies and annotations can be added to the local database.

- In <http://www.ncbi.nlm.nih.gov/assembly/> we introduce the term “Epstein-Barr virus” or “Human herpesvirus 4”
- On the search result page, above, in “Global assembly definition” we can access the link “NC_009334.1” http://www.ncbi.nlm.nih.gov/nuccore/NC_009334.1
- Download the genome sequence in fasta format (Fig. 4a) into “/opt/sRNAtoolboxDB/index” (we assume that the file name is NC_009334.fa).
- Edit the header as shown in Fig. 4b. The reason is that the “|” is used as internal separator by sRNAbench and should therefore not be used as sequence names. Furthermore, we add the “:ebv” tag which will be used in the genome distribution statistics.

After obtaining the sequences, we need to generate the genome index and the required sequence object file (this need to be copied into the “seqOBJ” folder):

```
cd /opt/sRNAtoolboxDB/index
bowtie-build NC_009334.fa NC_009334
makeSeqObj NC_009334.fa
mv NC_009334.zip ../seqOBJ/
```


Box 4: Populate the Local Database

General recipe to add annotations to the database:

Genome assemblies:

- Make the bowtie indexes with bowtie-build and place them into the “**index**” folder.
- Make the java sequence object file with makeSeqObj and place the zip file into the “**seqOBJ**” folder.

Sequence annotation: those can be given in fasta format, as bowtie indexes or even in bed or gtf format (in “genome mode“ ONLY). Those are placed always in the “libs” folder.

sRNAhelper: This program can assist the user in setting up the local database for species which are not covered by our sRNAtoolbox database. For example, it can help to extract the sequences of a species (o whole genus, family ect) from RNACentral [30] or genomic tRNA database [31] and format them for use with sRNAbench. Note that these annotations can be also generated using the appropriate helper tools from sRNAtoolbox web server: <http://bioinfo5.ugr.es/srnatoolbox>

Now we can detect the reads that originate from Epstein-Barr viruvs:

```
sRNAbench input=/home/srna/data/nasopharyngeal_carcinoma/SRR2105951.fastq.gz output=/home/srna/results/SRR2105951_miRBase_genome_EBV adapter=TGGAATTCTCGGGTGCCAAGGG graphics=true microRNA=hsa:ebv species=hg38_prim_mp:NC_009334
```

3.8.1 Explanation of the Command

- **microRNA=has:ebv** in order to profile microRNA from the Epstein-Barr virus and human simultaneously we only need to add its short species name (ebv) separated by “:”
- **species=hg38_prim_mp:NC_009334** Add the EBV index. Just like above, in general we can specify as many genomes as we want. The reads are mapped simultaneously to all and only the best mappings are retained.

3.8.2 Results of the Command

- **genomeDistribution.txt:** it can be seen (Fig. 3b) that EBV sequences can be detected, but at very low levels.
- **mature_sense.grouped** and **mature_sense_SA.grouped:** EBV microRNAs can be detected, but at very low levels.

Given the low levels of EBV derived molecules, those cannot explain the high number of unmapped reads that we observed before. We use another tool, sRNAblast in order to obtain more knowledge about the rather high number of unmapped reads.

3.9 Characterize Unmapped Reads: sRNAblast

This program allows using all “BLAST” databases hosted at NCBI to obtain more information about the “unmapped reads” which could reflect contamination, sequencing errors but also correspond to biologically important information like virus or bacterial infections.

```
sRNAblast input=/home/srna/results/SRR2105951_miR-
Base_genome/reads.fa output=/home/srna/results/
SRR2105951_blast maxReads=50
```

3.9.1 Explanation of the Command

- **input=“file”:** sRNAblast accepts the same input files as sRNAbench. In this example we use the “reads.fa” file which holds in genome mode the reads that were not mapped to the genome.
- **maxReads=integer:** number of top reads that will be “blasted” (taken from an expression value ordered list). Blast is a very time intensive procedure. The query is executed on the NCBI servers, and therefore the execution time depends very much on the server load. Therefore, this number should not be too high.

3.9.2 Results of the Command

- **blast.out:** this file contains all reads mapped to a sequence in the database. Only the hits with the highest score are preserved.
- **species.out:** this file contains the read count (RC) and percentage per species. Note that if a read maps to several species, its read count is distributed between the species.
- **speciesSA.out:** the single assignment file of the species. The **species.out** file is sorted, and then each read that maps to several species is assigned only once; to the species with the highest number of mapped reads (Fig. 5).
- **tax.out:** The read distributions per domains (superkingdoms).

The **speciesSA.out** file and its visualization can be seen in the Fig. 5a, b. It can be seen that around 25% of the input read count (50 most expressed, unmapped reads) were assigned to *Acinetobacter baumannii*. This gram-negative bacterium behaves frequently as an opportunistic pathogen in humans, affecting people with weak immune systems. Moreover it is becoming increasingly important as a hospital-derived infection. In this case, it would be important to investigate further if this high number is a consequence of the pathology, or if it might be even a cause.

3.10 Profile Other smallRNAs

The annotations to profile other small RNAs are also downloaded by means of the populate tool and include cDNA and ncRNA from Ensembl, tRNAs from the genomic tRNA database and all RNA types from RNA central:

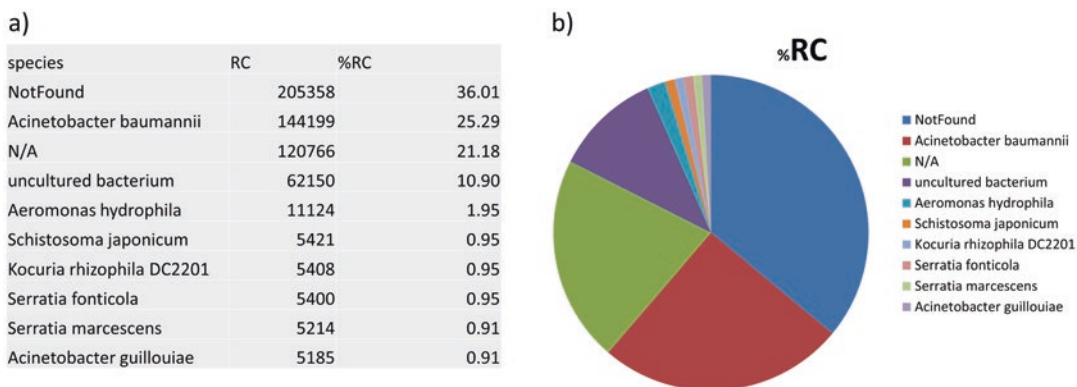


Fig. 5 (a) First lines of the speciesSA.out file and its graphical representation **(b)**. “NotFound” represents those reads that could not be mapped (“blasted”) to the used database. Note that these numbers are only based on the 50 most expressed reads that could not be mapped to the human genome.

Box 5: sRNAblast

Several parameters can be passed to the **blastn** algorithm

blastDB=“**name of the NCBI blast database**”: by default the database “nr” is used, but all NCBI blast databases can be selected.

word_size=“**integer**”: from NCBI manual: *Word size is roughly the minimal length of an identical match an alignment must contain if it is to be found by the algorithm (default 9).*

maxEvalue=“**number**”: the maximum E-value which will be reported in the output

```
sRNAbench    input=/home/srna/data/nasopharyngeal_carcinoma/SRR2105948.fastq.gz    output=/home/srna/results/SRR2105948_libs adapter=TGGAATTCTCGGGTGCCAAGGG
graphics=true microRNA=hsa libs=hg38_ncRNA libs=hg38_RNAcentral libs=hg38_genomic_tRNA libs=hg38_cdna 'libsStringTypes=mature#sense;hairpin#sense|snRNA#sense|snoRNA#sense|tRNA#sense|rRNA#sense|ncRNA#sense|protein_coding#sense;cdna#sense;cds#sense|protein_coding#antisense;cdna#antisense;cds#antisense|piRNA#sense|repeat#sense' 'libsStringNames=miRBase(sense)|snRNA|snoRNA|tRNA|rRNA|ncRNA|cdna(sense)|mRNA(antisense)|piRNA|repeats(sense)'
```

3.10.1 Explanation of the Command

- **libs**=“**library name**”: in library mode the annotations should be given either as fasta files or directly as bowtie indexes (index base names). If only the file name is given (without absolute path), sRNAbench supposes to find the files in “/opt/sRNA-toolDB/libs”, i.e., in the “libs” folder of the local database (“dbPath”). A description of the downloaded annotations can be found in the **info.txt** file on the Desktop.

3.10.2 Results of the Command

- **libsStringTypes=“string”**: this parameter is ONLY important for the summary files. It specifies the categories that should be grouped in the summary files.
- **libsStringNames=“string”**: The names of the categories, i.e., the number of names must fit the number of types.
- **“library name”_ “orientation”.grouped files**: For each library, 4 expression files are produced: 2 for each orientation (sense and antisense) and 1 for each multiple mapping “treatment” (read count adjusted and single assignment). For example **“hg38_ncRNA_sense.grouped”** holds the expression profiles for noncoding RNA annotations obtained from Ensemble mapped in sense direction (with adjusted read counts) while **“hg38_ncRNA_sense_SA.grouped”** contains the single assignment expression values.
- **“mappingStat.txt” (in “stat” folder)**: the summary file of the mapping process (the corresponding graphic can be found in the **“graphs”** folder). *see* Fig. 6a, b.
- **The read length files in the “readLength” folder**: for each category (like tRNA, snoRNA) the length distribution of the reads is generated (see Fig. 6c).

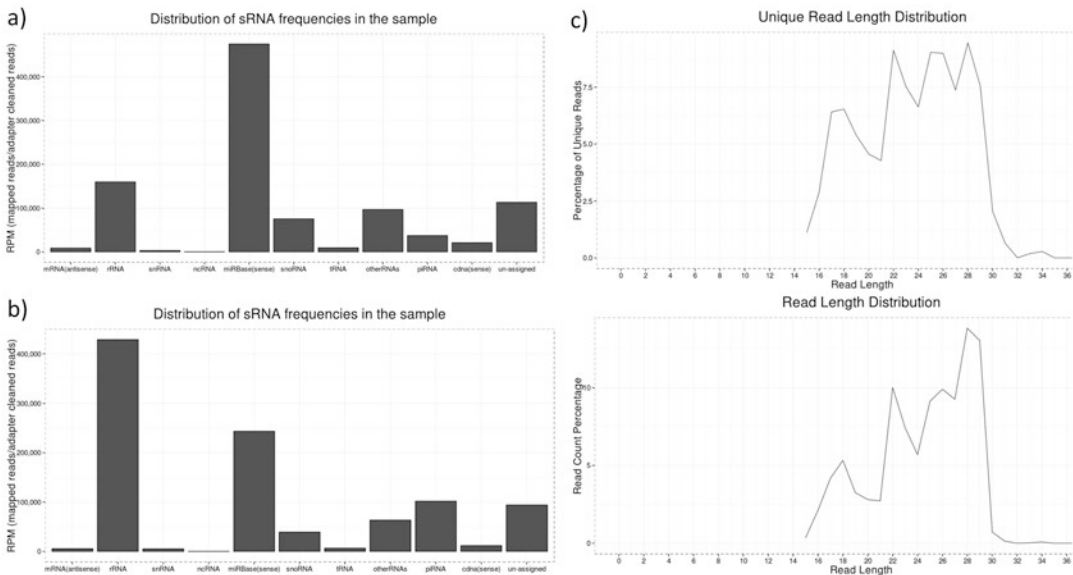


Fig. 6 (a) The mapping statistic as a function of RNA type using “genome mode” and (b) using “library mode.” One clear difference between both distributions is the frequency of rRNA (ribosomal RNA). This is due to the default thresholds for multiple mappings. While in the genome mode, only reads with at most 10 mappings are reported, in the library all mappings are reported. Given that rRNA derived reads map normally to a high number of loci, in genome mode those highly redundant reads are not considered, hence the notable difference in frequency between both modes. (c) The read length distribution for snRNA (small nuclear RNAs) fragments, above for unique reads and below for the read count. It can be seen that fragments from 18 nt to 30 nt exist

- **rnaComposition_readLength.txt in the stat folder:** the RNA composition as a function of read length, i.e., allows answering questions like: which percentage of 33 nt long reads are of tRNA origin?

Box 6: Other Small RNAs

Differences between library and genome mode:

Annotation types in library mode: only bowtie indexes or fasta files can be provided. Fasta files will be first converted to bowtie indexes internally.

Annotation types in genome mode: fasta annotations will be converted into an internal bed format (by mapping of the annotations against the genome); bed and gff/gff3 will be converted into an internal bed format; bowtie indexes are **NOT** converted into bed format, **BUT** the genome mapped reads are mapped again against those indexes.

Up-downstream windows (winUpLibs= and winDownLibs=): these parameters allow adding some flanking regions to the annotation for profiling which is not possible in library mode (note that this is similar to the flanking windows applied to the microRNA precursors shown above).

plotLibs=true: writes out the alignment files for all annotations given by **libs=**

3.11 Differential Expression

To calculate differential expression, we will need to run first all samples (see DE section in **protocol.txt**). Then we can use the sRNAde program to generate the expression matrix calculating the differential expression statistics.

```
sRNAde input=/home/srna/results/ output=/home/srna/
results/diffExpr_libs_genome grpString=SRR2105938_libs_
genome:SRR2105939_libs_genome:SRR2105940_libs_
genome:SRR2105941_libs_genome#SRR2105948_libs_
genome:SRR2105949_libs_genome:SRR2105950_libs_
genome:SRR2105951_libs_genome grpDesc=nasopharyngitis#
carcinoma minRCexpr=5 sampleDesc=nasopharyngitis_38:naso
pharyngitis_39:nasopharyngitis_40:nasopharyngitis_41:carcino
ma_48:carcinoma_49:carcinoma_50:carcinoma_51
```

3.11.1 Explanation of the Command

- **grpString="the grouping instruction":** group string must contain the names of the different sRNAbench output folders in the following format: f1_1:f2_1#f1_2:f2_2 being f1_1 the first folder (sample) of the first group (controls in a case/control study), f2_1 the second folder of the first group, f1_2 the first folder of the second group (cases), etc...

3.11.2 Results of the Command

- **grpDesc="string" (optional):** the groups (different experimental conditions, phenotypes, etc) that exist in the study. In our case we have two groups, individuals with **nasopharyngitis** and with **carcinoma**.
- **sampleDesc="string" (optional):** The names of the samples. The user can give each sample a name which will be used in the heatmaps. Otherwise, the folder names of the individual sRNAbench runs are used.
- **minRCexpr="integer":** the minimum read count that ALL samples in at least one condition (group) must have.
- **mature_sense_5_2.mat:** the expression matrix based on the `mature_sense.grouped` file (indicated by the base name "mature_sense"). This file is used by default for differential expression analysis. The first number of the output file is determined by the **minRCexpr** parameter which was set to 5. The second number indicates the column that was used for the expression value. In this case the read count (second column) which is the required input for edgeR, DESeq and noiSeq.
- **mature_Sense_5_4.mat:** the expression matrix based on the fourth column (starting with 0), i.e., the library normalized RPM expression values.
- **mature_sense_5_2_nasopharyngitis_vs_carcinoma.xlsx:** the results when applying edgeR, DESeq and noiSeq to the expression matrix (`mature_sense_5_2.mat`).
- **mature_sense_5_2_heatmap_perc0.78.png** and **mature_sense_5_2_heatmap_perc0.78_median_normalized.png** (Fig. 7a): the heatmaps based by default on the 50 most expressed microRNAs (or annotations in general). The second is using a median normalization in order to obtain a clearer color code.
- **mature_sense_5_2_TMM_normalized.tsv:** the expression matrix with the TMM normalized read counts (from edgeR).
- **sequencingStat.txt:** a summary of the sequencing, preprocessing, and genome mapping process (Fig. 7b).

3.12 Make Study Summary Files

The sRNAde program can generate different summary files of a whole study. By default, the "sequencingStat.txt" file is

Box 7: Differential Expression

rpmCol="integer" (column of expression file (0-based)): this value should be either 4 (library normalized RPM values) or 5 (total read count normalized RPM values).

rmpCol="expression file": default is "mature_sense.grouped", but any file generated with the **libs=** tag can be specified as well.

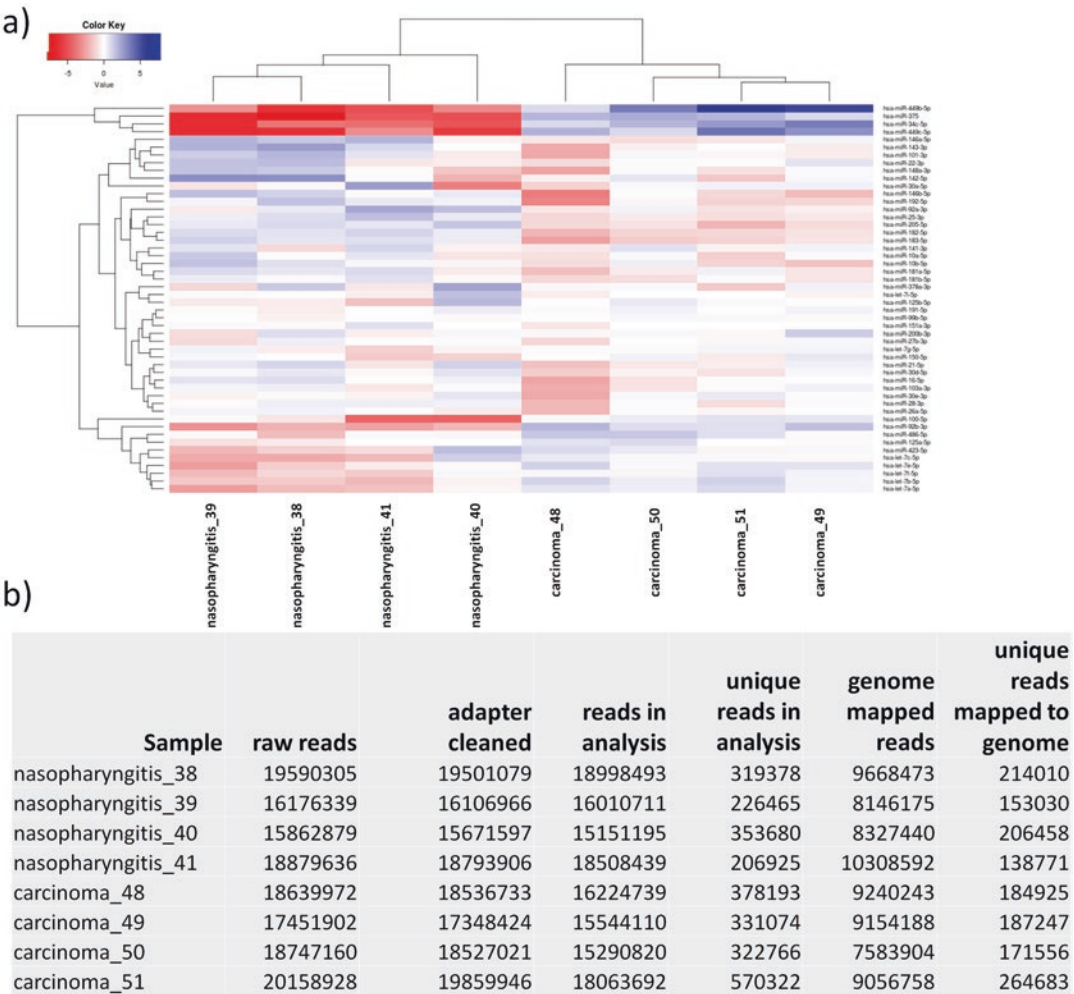


Fig. 7 (a) The heatmap generated by the sRNAde program. It can be seen that both groups can be clearly separated. (b) The study summary of sequencing and mapping generated by sRNAde

Box 7 (continued)

makeSingleAssignDE=true: generates the single assignment expression values.

annotName="name of the category": only needed if makeSingleAssignDE=true. The name of the category must be specified (for example, mature for mature_sense.grouped).

Warning!

Do not use the single assignment files for differential expression (like **diffExprFiles=mature_sense_SA.grouped**). If single assignment should be used, then set **makeSingleAssignDE=true** together with **diffExprFiles=mature_sense.grouped**.

generated. In a very similar way, sRNAde can summarize any file generated by sRNAbench. In the “**stat**” mode sRNAde parses for each sample a given column out of the analysed file, and generates an “expression matrix” type of file. This means that we will have a “key” in the first column, and the “value” for each sample in the other columns.

```
sRNAde input=/home/srna/results/ output=/home/srna/
results/diffExpr_libs_genome_stat grpString=SRR2105938_
libs_genome:SRR2105939_libs_genome:SRR2105940_libs_
genome:SRR2105941_libs_genome#SRR2105948_libs_
genome:SRR2105949_libs_genome:SRR2105950_libs_
genome:SRR2105951_libs_genome grpDesc=nasopharyngitis#
carcinoma minRCexpr=5 sampleDesc=nasopharyngitis_38:naso
pharyngitis_39:nasopharyngitis_40:nasopharyngitis_41:carcino
ma_48:carcinoma_49:carcinoma_50:carcinoma_51
diffExpr=false seqStat=false stat=true statFiles=mappingStat.txt
colData=6 minRCdata=0 folderData=stat
```

3.12.1 Explanation of the Command

- **stat=true:** activates the general study summary function.
- **statFiles=“mappingStat.txt”:** the file that should be summarized.
- **colData=6:** the column that should be parsed out, i.e., the sixth column (0-based) is the RPM value including the unmapped/unassigned reads (unmapped in library mode; genome mapped but not assigned to any annotation in genome mode).
- **minRCData=0:** the minimum expression value. In this case, no limits should be applied to the RPM values.
- **folderData=stat:** the folder where the “**statFiles**” are located. This will be normally either “**stat**” or “**genomeDistribution**”. If the **folderData** parameter is not set, then the “**statFiles**” are expected in the root of the output folder.

3.12.2 Results of the Command

- **mappingStat_0_6.mat:** The matrix summarizing the sixth column of the mappingStat.txt file over all samples. The first column contains the RNA category, and the other columns the % of reads mapped to these categories in the different samples.
- **mappingStat_0_6_1_2.ttest:** the within and between groups statistics using the corresponding “matrix file” (mappingStat_0_6.mat in this case) applying the grouping provided by the user. The output has the following format.
 - name: the name of the category (tRNA, snoRNA in this case).
 - mean_1: the mean value of group 1 (nasopharyngitis in the example).
 - stdev_1: the standard deviation of group 1.

- mean_2: the mean value of group2 (carcinoma in the example).
- stdev_2: the standard deviation of group 2.
- P: the exact p-value using Fisher exact test.
- FDR: the q-value (the corrected p-value).

3.13 Determine Consensus Target Genes

sRNAtoolbox contains two programs to calculate consensus target gene predictions, one for animals and one for plants. These programs are especially useful for cross-species predictions (microRNA of species A vs. transcriptome of species B) or novel microRNAs (novel microRNA of species A vs. transcriptome of species A). In those cases, web servers of precalculated target predictions like TargetScan [17] or experimentally verified interaction databases like TarBase [18] cannot be used. One example would be to estimate the impact of EBV microRNAs in a human cell.

First we need to extract EBV microRNA sequences from the miRBase database. For this we use the sRNAhelper tool which includes several small helper tools.

```
sRNAhelper    input=/opt/sRNAtoolboxDB/libs/mature.fa
output=/home/srna/results/BART1_mir
search=BART1- mode=FA
```

3.13.1 Explanation of the Command

- **search=“the search term”**: all sequences that contain this search term in the name will be extracted.
- **mode=FA**: indicates which of the implemented small tools will be used; “FA” means that certain sequences will be extracted from a fasta file.

```
miRNAconsTargets    /home/srna/results/BART1_mir/
mature.txt          /opt/sRNAtoolboxDB/utr/Homo_sapiens3p_
UTR.fa             /home/srna/results/BART1_targets/      2
‘TS:MIRANDA:PITA’ ‘-s: ; ‘
```

3.13.2 Results of the Command

- **mature.txt**: contains the extracted sequences. The output file is named like the original fasta file used to extract the sequences.

3.13.3 Explanation of the Command

- The input data needs to be provided in a specific order:
 - **microRNA file**: the microRNA sequences in fasta format.
 - **target sequences**: i.e., the 3'UTR sequences in fasta format.
 - **output_directory**: the output directory.
 - **p(threads)**: the number of CPUs that should be used (2 in the example).
 - **program string**: the programs that should be used; TS=TargetSpy, PITA=pita, MIRANDA=miranda. Note that pita is the program with the longest run time.

Table 2
Installed programs

Analysis type	Program(s)	Reference
Sequence alignment	Bowtie, BLAST+	[19, 20]
sRNAtoolbox stand-alone programs	sRNAbench, sRNAde, sRNAblast, miRconsTargets, sRNAhelper	[9, 10]
Vienna package	Prediction and comparison of RNA secondary structures	[21]
Target gene prediction (animals)	TargetSpy, pita, and miranda	[22–24]
Target gene prediction (plants)	psRobot, Tapir	[25, 26]
Differential expression analysis	edgeR, DESeq, and NOIseq	[27–29]
NGS related programs	SRA Toolkit	[12]

- **program parameters:** the parameters that should be passed to the programs. Must be given in the same order than the program string. In this example, only TargetSpy receives a parameter, “-s” which enforces a seed region.

3.13.4 Results of the Command

- **pita.txt; ts.txt and miranda.txt:** the predicted targets of the three algorithms.
- **consensus.txt:** the consensus file giving the microRNA, the target sequence, the number and names of programs that predict this interaction.

Note that for plant target predictions the program miRNA-consTargets_plants.py can be used in a similar way. This program uses Tapir and psRobot for target gene prediction.

All installed programs are listed in Table 2.

References

- Ivey KN, Srivastava D (2010) MicroRNAs as regulators of differentiation and cell fate decisions. *Cell Stem Cell* 7:36–41
- Ebert MS, Sharp PA (2012) Roles for microRNAs in conferring robustness to biological processes. *Cell* 149:515–524
- Iorio MV, Croce CM (2012) MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med* 4:143–159
- Zhou M, Luo H (2013) MicroRNA-mediated gene regulation: potential applications for plant genetic engineering. *Plant Mol Biol* 83:59–75
- Bernal D, Trelis M, Montaner S, Cantalapiedra F, Galiano A, Hackenberg M, Marcilla A (2014) Surface analysis of microcoelium dendriticum. The molecular characterization of exosomes reveals the presence of miRNAs. *J Proteomics* 105:232–241
- Buck AH, Coakley G, Simbari F, McSorley HJ, Quintana JF, Le Bihan T, Kumar S, Abreu-Goodger C, Lear M, Harcus Y, Ceroni A, Babayan SA, Blaxter M, Ivens A, Maizels RM (2014) Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat Commun* 5:5488
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079
- Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM (2011) miRanalyzer: an update on

- the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res* 39:W132–W138
9. Rueda A, Barturen G, Lebrón R, Gómez-Martín C, Alganza Á, Oliver JL, Hackenberg M (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res* 43:W467–W473
 10. Barturen G, Rueda A, Hamberg M, Alganza A, Lebron R (2014) sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods in Next Generation Sequencing* 1:21–31
 11. Cai L, Ye Y, Jiang Q, Chen Y, Lyu X, Li J, Wang S, Liu T, Cai H, Yao K, Li J-L, Li X (2015) Epstein-Barr virus-encoded microRNA BART1 induces tumour metastasis by regulating PTEN-dependent pathways in nasopharyngeal carcinoma. *Nat Commun* 6:7353
 12. The SRA knowledge base, NCBI help manual. Staff SRAS: using the SRA toolkit to convert .sra files into other formats. 2011 Bethesda (MD)
 13. Kozomara A, Griffiths-Jones S (2014) miR-Base: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 42:D68–D73
 14. Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E, Peterson KJ (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet* 49:213–242
 15. Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF (2009) Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics* 25:2298–2301
 16. Chu EA, Wu JM, Tunkel DE, Ishman SL (2008) Nasopharyngeal carcinoma: the role of the Epstein-Barr virus. *Medscape J Med* 10:165
 17. Agarwal V, Bell GW, Nam J-W, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4
 18. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 40:D222–D229
 19. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
 20. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421
 21. Lorenz R, Bernhart SH, Höner Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* 6:26
 22. Sturm M, Hackenberg M, Langenberger D, Frishman D (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics* 11:292
 23. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39:1278–1284
 24. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. *PLoS Biol* 2:e363
 25. Wu H-J, Ma Y-K, Chen T, Wang M, Wang X-J: PsRobot (2012) A web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Res* 40, W22–W28.
 26. Bonnet E, He Y, Billiau K, Van de Peer Y (2010) TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* 26:1566–1568
 27. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140
 28. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106
 29. Tarazona S, García F, Ferrer A, Dopazo J, Conesa A (2012) NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet J* 17:18
 30. The RNAcentral Consortium (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res* 43:D123–D129
 31. Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* 37:D93–D97

An Assessment of the Next Generation of Animal miRNA Target Prediction Algorithms

Thomas Bradley and Simon Moxon

Abstract

The advent of genome-wide next-generation sequencing technologies has revolutionized the genomic and transcriptomic fields. New technologies also present an opportunity for greater discovery and understanding of post-transcriptional processes, in particular, translational inhibition of transcripts by miRBP (microRNA-RNA binding protein) complexes. Not only have novel methodologies been developed for the direct sequencing of RBP-bound RNA, but a new class of miRNA (microRNA) target prediction algorithms trained on this data has emerged. In this article, we will explore and evaluate the next generation of animal miRNA target prediction algorithms, their relationship to more traditional prediction methods, and the implications of such methodologies for the future of miRNA target prediction and miRNA research as a whole.

Key words miRNA, Target prediction, Bioinformatics

1 Introduction

The computational prediction of miRNA targets is a necessary tool for researchers. Estimates suggest that >60% of human transcripts of protein-coding genes are targeted by miRNAs [1]. Each targeted transcript can contain multiple target sites along its entire length, which can bind to different miRNAs [2]. The total number of distinct target sites, which are conserved above background levels, in just the human 3'UTR space alone has been estimated to be >45,000 [1]. Traditional methods for the verification of candidate target sites, such as use of the luciferase assay, are too low-throughput to feasibly screen all candidate miRNA target sites. Higher throughput studies utilizing Ago-immunoprecipitation protocols such as CLIP (cross-linking and immunoprecipitation) [3, 4] and CLIPL (cross-linking immunoprecipitation and ligation) [5–7] have been performed; however, the set of all experiments of this type published are limited by the variety and diversity of the species, tissue types, and developmental time periods assessed. Additionally, alternative cleavage and polyadenylation events

(APA), modulating the presence of target sites on the 3'UTR of transcripts [8], further compound the complexity of gene expression and miRNA targeting, which cannot be fully elucidated through the relatively small numbers of CLIP and CLIPL experiments conducted.

There are many different ways to distinguish miRNA target prediction algorithms, including the different biological and chemical variables by which candidate targets are selected (e.g., conservation, site accessibility), and also the extent to which different algorithms rely on human or machine-guided learning to predict targets. The emphasis of this article, however, is the influence of new experimental validation protocols for the training of novel target prediction methods. Similarly, although miRNA-mediated gene regulation is widespread across many taxonomic groups in the eukaryotic domain, the focus of this article will be on miRNA activity in animals, which is known to possess distinct targeting principles [9].

The relationship between the experimental identification, and the computational prediction of miRNA targets, is not antagonistic, but rather, reciprocal and mutualistic. Information derived from miRNA target identification experiments often informs the design or training of new prediction algorithms. Similarly, computational predictions can be used by experimental biologists to restrict and focus the set of candidate targets to be verified.

The development of newer, high-throughput technologies (CLIP and CLIPL) for target identification has provided computational biologists with relatively large and novel datasets, for which a new class of miRNA target prediction algorithms has emerged, dividing the set of all prediction methods into two principal types: (1) Those prediction algorithms trained on transcriptome-wide CLIP and CLIPL studies and (2) and those prediction algorithms not trained using data derived from these novel experimental methods.

What will follow in this article is a review of the most representative and successful algorithms of the two types identified, an assessment of their respective performances, and a discussion of how the field of miRNA target prediction, has and will continue to be impacted by the arrival of high-throughput target identification data.

2 Traditional Target Prediction Algorithms

Among the first miRNAs to be discovered was *lin-4* in *Caenorhabditis elegans* which was suspected to target the heterochronic *lin-14* gene due to the complementarity between the *lin-4* sequence and the 3' UTR of the *lin-14* transcript [10]. Transcripts from reporter genes including the *lin-14* 3'UTR were shown to be downregulated to a much weaker extent in *lin-4* mutants compared to wild-type [11].

A method often utilized in early research on miRNA target validation was the luciferase reporter assay for detecting miRNA targets by utilizing the natural bioluminescence properties of the luciferase enzyme. Because of the low-throughput nature of the method, it would have been difficult to train models on this data using machine learning principles. However, data obtained from the extensive use of the luciferase assay has helped to establish some of the core principles of miRNA targeting used to develop the earliest target prediction algorithms.

For example, DIANA-microT [12] was one of the earliest miRNA target prediction algorithms produced. DIANA-microT is noteworthy because it explicitly used experimental data generated within the same study in which the method was published to help design the prediction algorithm.

In [12], *C. elegans* miRNA targets experienced single nucleotide mutagenesis multiple times along the entire length of the putative miRNA binding site (~22 nt). The efficacies of these targets upon mutagenesis were then assayed using a dual *Renilla* firefly luciferase assay. A significant enrichment of a mutagenesis effect was seen at the first 7 nucleotides at the 5' region of the miRNA. This supported previous studies establishing the importance of the 5' miRNA seed region. Kiriakidou et al. then used this information to help build their computational model for miRNA target prediction, strongly implicating complementarity to the seed region as being important for binding. The Kiriakidou et al. study is a representative example of how target prediction was achieved (Fig. 1) in the earlier years of miRNA research.

The low-throughput of Luciferase assays limits the number of candidate targets which can be validated within a given time frame. Higher throughput validation of miRNA targets can be achieved through the use of microarrays, by quantifying and comparing gene expression levels in an experimental condition of miRNA knockdown or transfection, with a control condition in which cellular miRNA levels are not perturbed. The generation of larger amounts of data, through the use of microarrays, enabled higher

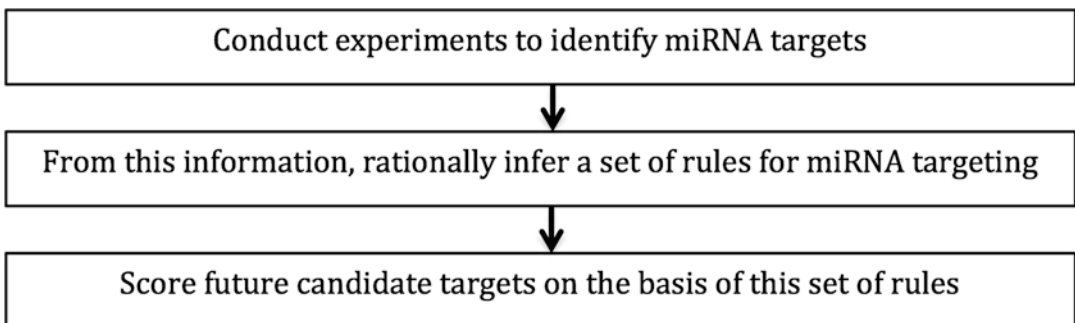


Fig. 1 A typical developmental pipeline for the production of heuristic or rule-based target prediction models. This design is often necessary when low-throughput information is being used, in which there is not enough information available to apply machine learning techniques

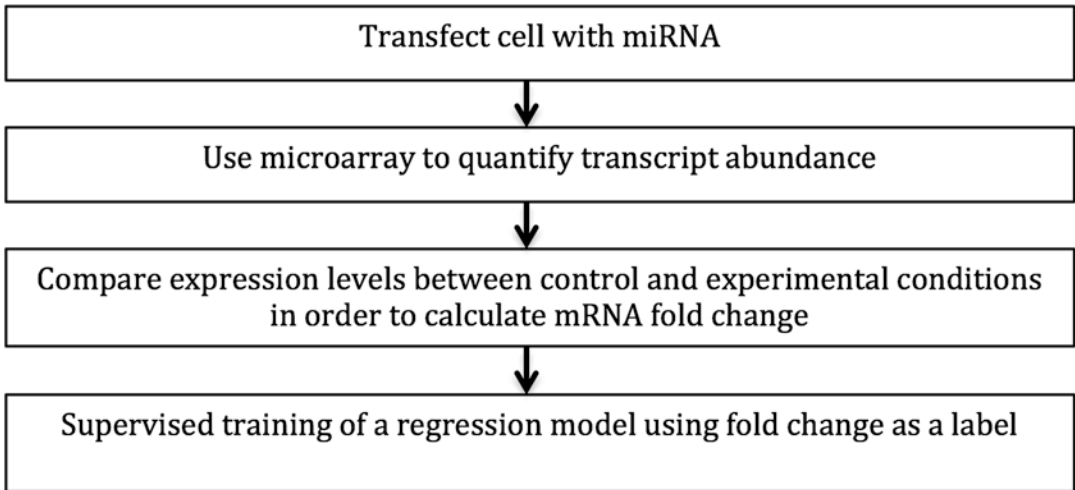


Fig. 2 A typical developmental pipeline for the production of a target prediction model trained on microarray transfection data

confidence inference from experimental data, and also the application of machine learning tools to develop more sophisticated methods for computational prediction of miRNA targets.

An example of a successful target prediction algorithm trained from microarray data is mirSVR [13]. mirSVR is an algorithm that has been developed to process output from the miRanda algorithm developed years earlier [14, 15], producing a miRanda–mirSVR hybrid model with associated scored targets freely accessible on the microrna.org online resource [16].

mirSVR used microarray transfection data to train a support vector regression model. Fold change between the control and experimental condition (i.e., miRNA transfection or knockdown) is used as the response variable in this model.

TargetScan7 [17] is a prediction method trained on similar data, and using a similar training protocol (Fig. 2). TargetScan has existed as many iterations [1, 18, 19] prior to the release of the most recent version. The TargetScan7 model in particular was trained using 74 microarray datasets to produce a multilinear regression model for four different identified types of miRNA seed target site (8mer, 7mer-m8, 7mer-A1, 6mer).

3 NGS-Trained Prediction Methods

The basic CLIP procedures included in all CLIP protocols include the fundamental steps of first UV crosslinking bound RNA molecules to the AGO protein, followed by RNA digestion, immunoprecipitation, RNA isolation, and sequencing of the previously

AGO-bound RNA [3]. CLIP was novel in the respect that it enabled direct, genome-wide discovery of miRNA targets. Implementation of CLIP protocols produced larger datasets than those typically produced from microarray experiments. CLIP data is also, to an extent, less ambiguous than equivalent microarray data, as a direct, physical interaction between the mRNA (messenger RNA) and miRNA is being reported, rather than a change in mRNA expression levels (as is the case for microarray data).

One of the first computational algorithms produced, which considered data derived from next-generation sequencing methods is “MIRZA” [20]. MIRZA is unique among all of the computational methods just discussed in the sense that it is trained exclusively on CLIP data, in particular, PAR-CLIP data [4]. From this PAR-CLIP data, Khorshid et al. developed a biophysical model to help predict miRNA:mRNA interactions. An adaptation of the basic MIRZA algorithm was released in the form of MIRZA-G in which researchers incorporated sequence and structure feature information, as well as evolutionary conservation information in some variants of the model [21].

Another computational method that was trained on CLIP data was the DIANA-microT-CDS method [22], which, like the MIRZA method mentioned above, was trained using PAR-CLIP data [4]. Although the CDS is explicitly mentioned on the name of the algorithm, the computational method can be used to predict target sites in both the coding region (CDS) and 3' UTR of a given transcript. Candidate target sites are scored, and scores are summed across the entire transcript (CDS and 3' UTR) to give a measure of targeting for the entire transcript. The computational algorithm was trained using a combination of stepwise regression and multilinear regression models. Stepwise regression using the Akaike Information Criterion (AIC) was used to select features. These features were combined using the general linear model.

A natural evolution in experimental methodology was the progression from CLIP to CLIPL (cross-linking, immunoprecipitation, and ligation) protocols. In CLIPL, there is an additional ligation step, hybridizing the miRNA and mRNA target region [5]. As a result, cognate mRNA fragments and miRNA molecules can be sequenced together. CLIPL contains all of the advantages of CLIP studies previously mentioned, with the additional benefit of even lower ambiguity identification of targets, due to the sequencing of chimeric mRNA–miRNA sequences.

CLIPL methods have only been available for a few years in reference to the current time of writing [5–7]. Computational methods have also been trained on CLIPL data using a similar training pipeline as seen for CLIP studies (Fig. 3): The latest iteration of MirTarget [23] uses information from both [4] and [5] to train a support vector machine model.

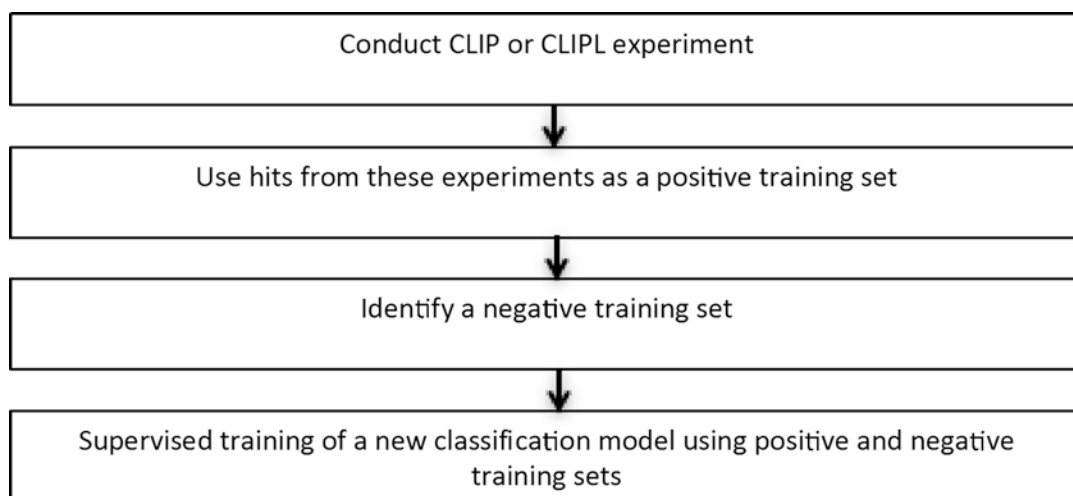


Fig. 3 A typical developmental pipeline for the production of a target prediction model trained on CLIP or CLIP data

4 A Practical Approach

The next generation of miRNA target prediction methods has been described and their performance has been assessed. On a more practical level, consideration of how this new data and information impacts methods or protocols for assessing putative targets must be considered.

Performing CLIP and CLIP data experiments is associated with considerable practical and technical difficulties and uncertainties, including relatively low UV cross-linking and RNA-RNA ligation efficiencies [5]. This problem, coupled with the relatively large number of experimental steps involved in these protocols, means that researchers will very likely need specialist training to perform these experiments. Additionally, the resources required mean that performance of CLIP and CLIP data experiments may not be practical in the short term for a large number of research laboratories.

However, although some researchers may not be able to generate their own data, it should be emphasized that the generation of CLIP and CLIP data can still improve the prediction of candidate miRNA targets. What will follow, is a brief outline of how a researcher can perform target prediction, using previously developed tools and resources, but also in the light of the newly developed CLIP and CLIP data methods:

First, researchers can access, online, publicly accessible databases such as TarBase [24] and StarBase [25, 26], which are databases of experimentally validated miRNA targets. In the case of TarBase, there will be a full description of the method by which the

target was validated, including the more novel CLIP and CLIPL methods as well as many other methods, and also the publication associated with each validated interaction. Starbase differs from TarBase in the respect that it only refers to CLIP and CLIPL methods, and also in the fact that users can combine this information with computational predictions from multiple algorithms.

As well as accessing databases of experimentally validated targets, researchers can provide additional evidence for a miRNA-target interaction using computational target prediction. All of the algorithms mentioned in the previous section, including MIRZA and MiRTarget trained on CLIP and CLIPL data, include databases of computationally predicted targets. In circumstances, in which a researcher wishes to perform target prediction on a candidate target of a novel miRNA sequence, or if the target itself is novel, or indeed, the candidate interaction has not already been assessed and stored on online databases, researchers can download the relevant scripts from TargetScan and microrna.org websites to perform custom target prediction.

Researchers, when examining candidate miRNA-target interactions, usually are considering the candidate interactions with respect to a particular species, tissue, or developmental process. It is important to acknowledge that computational target prediction mostly occurs at a generalized and abstracted level in which tissue or species-specific considerations are often neglected. For example, for any given miRNA-target interaction, either or both the miRNA and the target may not be expressed in a given tissue or interest, as a result, that result would not be physiologically relevant even though it may possess a high prediction score. A well-considered approach for researchers, whether considering CLIP or CLIPL data, or data derived from older methods, would be to identify, e.g., using RNA-Seq data, the expression levels of targets and miRNA both in their species and tissue of interest, to ensure that identified targets are biologically relevant.

To increase the evidence supporting a given interaction—a more active approach can be taken in which cellular expression levels of a given miRNA are perturbed by transfection of exogenous miRNA, or the reduction of endogenous miRNA expression levels through the transfection of antagomir molecules [27] that bind to, and thereby silence complementary miRNA molecules. Differential expression levels (between control and experimental condition) of predicted targets can then be conducted to infer whether or not miRNA perturbation affected the expression levels of targets. A large change in expression levels of the predicted target would provide additional evidence for the efficacy of that target. This method is particularly appropriate for those instances in which researchers wish to examine the effect of a single miRNA on a large number of potential targets.

Differential expression analysis (from microarray or RNA-Seq data) can also be conducted without prior perturbation of miRNA expression levels, in the case in which the researcher is examining the transition of a tissue from one developmental state to another, e.g., from a healthy state to a diseased state; from an adolescent state to a mature adult state, etc. In this case, the mechanism for differential expression of candidate mRNA targets is not an exogenous source of miRNA, but endogenous changes to miRNA levels. Differential expression levels of both miRNA and mRNA can be assessed, and miRNA–mRNA interactions can be inferred when there exists an anti-correlation between the expression of the two molecules through time. This differential expression analysis procedure can also be extended to analysis of CLIP data upon miRNA perturbation—referred to as d-CLIP [28], and also potentially CLIPL studies in the future.

A more direct approach for assessing the efficacy of targets is through the use of luciferase assays described earlier. The direct approach allows the identification or verification of higher-confidence targets than what may be found through the use of differential expression analysis alone. However, as mentioned previously, the luciferase assay is a much lower throughput approach. A potentially fruitful analysis pipeline would be to use the luciferase assay to validate targets most highly differentially expressed when examining microarray or RNA-Seq data, or to validate any particular targets of interest putatively identified by CLIP or CLIPL experiments.

All of the steps mentioned may be implemented in isolation of each other, as a collection of experimental and analytical steps. As mentioned previously, prediction of miRNA target produces a relatively large number of false-positive predictions. As a result, corroborating evidence beyond that which exists from a single experiment or single computational prediction may be necessary for effective miRNA target prediction. In the coming years, implementation of CLIP and CLIPL methods is likely to become easier and cheaper for research laboratories, and more scientists are likely to be trained in these methods. In which case, as well as the steps mentioned above, many research laboratories will be able to produce their own CLIP and CLIPL data.

5 Performance Review

There is often a bias in how algorithm performance is reviewed. For example, the performance of an algorithm is often tested on the same type of data on which the algorithm was trained (e.g. miRNA transfection data), which could disadvantage algorithms in the analysis that were not trained on that particular type of data. Therefore, it is often useful and necessary to examine how a particular algorithm is reviewed and tested using multiple different types of data set to mitigate this bias.

Table 1
A summary of results from [23]. The first column lists the names of the competing algorithms. The second column represents the results of a precision-receiver operator characteristic (pROC) analysis. The third column represents the ranking of results from a cumulative graph representation of fold change. The fourth column represents the results of an analysis reporting the average fold change of the top 300 predicted targets of all algorithms. The fifth column represents the results from an analysis in which the Pearson product moment correlation coefficient was found between target prediction score and gene expression changes. The last column represents the total of all rankings. Further details of the analyses can be found at [23]. The prediction algorithm associated with this paper is highlighted in bold

	pROC analysis	Cumulative distribution	Analysis of top targets	Correlation analysis	Total
MirTarget	1	1	1	1	4
TargetScan7 (context++)	2.5	3	2	2	9.5
DIANA-microT-CDS	4	3	3	3	13
miRanda-mirSVR	2.5	3	4	4	13.5
RNA22	5	5	5	5	20

Table 2
A summary of results from [17]. The first column lists the names of the competing algorithms. The second column represents the results of a cumulative distribution analysis. The third column represents the ranking of results from an R-squared analysis indicating the degree to which prediction algorithms explain variability in microarray transfection data. The last column represents the total of all rankings. Further details of the analyses can be found at [17]. The prediction algorithm associated with this paper is highlighted in bold

	Cumulative distribution	R ² test	Total
TargetScan (context++)	1	1	2
Seed-MIRZA-G-C	2.5	2	4.5
DIANA-microT-CDS	2.5	3	5.5
miRanda-mirSVR	4	4	8

A review of a selection of the algorithms discussed in the previous section has been conducted (*see* Tables 1, 2, and 3). Presented is a collation of data from some of the most recently released articles documenting prediction algorithm performance on a range of different performance metrics. Prediction methods used in this analysis are: MirTarget [23], miRanda-mirSVR [13], TargetScan7 [17], TargetScan.P_{CT} [19], TargetScan6 [18], DIANA-microT-CDS [22], MIRZA [20], Seed-MIRZA-G-C [21], and RNA22 [29].

Table 3

A summary of results from [21]. The first column lists the names of the competing algorithms. The second column represents the rankings of results of an analysis examining the fold change of the top targets of the different prediction algorithms. The third column represents the ranking of results from an analysis reporting the number of functional targets identified by each algorithm. The last column represents the total of all rankings. Further details of the analyses can be found at [21]. The prediction algorithm associated with this paper is highlighted in bold

	Analysis of top targets	No. of functional targets analysis	Total
Seed-MIRZA-G-C	1.5	2	3.5
TargetScan.P _{CT}	1.5	2	3.5
TargetScan6 (context+)	4.5	2	6.5
DIANA-microT-CDS	3	4	7
miRanda-mirSVR	4.5	5	9.5

The different target prediction methods were ranked by analyzing figures featured in the aforementioned published research. A conservative judgement was made that cumulative plot curves that, by inspection, overlapped $\sim \geq 20\%$ of the entire length of the curve were given the same rank, while curves with less or no overlap were given different ranks. A result of this ranking method was to prevent an assignment of different ranks to prediction algorithms with only slightly different performances, and hence similar output distributions. A fractional ranking system is employed.

6 Discussion

The expectation may be that CLIP and CLIPL trained methods would outperform competitors because of the greater quality and size of training data. However, from the results just presented, it can be inferred that no one prediction algorithm seems to dominate the performance rankings in all of the analyses conducted. More importantly, there does not seem to be any obvious difference in the average ranking of algorithms trained on CLIP and CLIPL studies and those not trained on those methods. These results suggest that reported performance is partially dependent on the exact datasets used for testing and the exact testing metric used, and that no algorithm achieves the highest ranking across all data inputs.

Although not discussed at great length in this article, different methods are trained using a different set of features often rationally chosen by the respective researchers, e.g., some methods place a particular emphasis on target site conservation. These idiosyncrasies in feature selection could potentially explain why CLIP and CLIPL trained methods do not consistently outcompete other algorithms.

This ambiguity is compounded by the fact that not all studies test the same algorithms, or the same versions of that algorithm.

Despite these conclusions, some broad inferences can be made. First, the algorithm associated with a given paper consistently achieves the highest or joint highest ranking for analyses reported in that same paper. Second, algorithms that place major importance on the structural stability of binding, e.g., RNA22, usually compare poorly to other methods, with the caveat that these methods are usually older, and hence may not reflect newly procured knowledge of miRNA biology. Neither DIANA-microT-CDS nor miRanda-miRSVR perform best in any analysis although the former outperforms the latter in most analyses. TargetScan performed well in [23] and [21], while MIRZA-SEED-G-C performed well in [17]. In reference to the comparison of CLIP and CLIPL with other methods—those algorithms not trained on CLIP and CLIPL seem to achieve some of the consistently best (TargetScan) and worst (RNA22 and miRanda-miRSVR) result across all analyses. Again, the fact that there is little consistency among method types is informative in itself.

It is important to reiterate, and has been shown in this article, that computational prediction of miRNA targets, while being necessary, is largely dependent on experimental data in the design and training of new prediction algorithms. Therefore, to attempt to explain the inconsistency of performance results of CLIP and CLIPL studies, a discussion of open questions relating to the experimental methods and data is necessary.

Not all experimental data is equal in quality and its usefulness. An ideal experimental dataset would be both high-throughput (i.e., large), would contain data of direct evidence of miRNA activity, and would be free from any experimental confounds and biases.

A sufficient number of studies using CLIP or CLIPL have been conducted for a thorough comparison to be made of sites identified by the newer and more traditional methods. What will follow is a brief assessment of the results that have been unveiled so far using the newer methods, and the implications for computational target prediction.

CLIP and CLIPL methods differ from more traditional methods in several ways. First, there is a remarkable difference in the proportion of target sites located in different regions along the length of the transcript. Compared to previous studies [18, 30], newer methods tend to display a much higher proportion of putative target sites located within the CDS of the transcript compared to older methods, and comparatively fewer sites within the 3' UTR of the transcript [5, 7].

Second, newer methods tend to identify a much higher number of noncanonical target sites [3, 31]. Noncanonical sites are those sites binding to the seed region of the miRNA, but containing a bulge or a mismatch. Furthermore, CLIP and CLIPL methods

have tended to identify a much higher number of sites not binding to the seed region at all, binding to the central region of the miRNA transcript, or rather binding to the 3' region of the transcript [5, 7]. It should also be noted that evidence has been published indicating that noncanonical and nonseed sites have been shown to be ineffective or only marginally effective at repression of mRNA expression levels [13, 17].

Trying to explain the discrepancy between these results and the results from earlier experimental methods is not a trivial problem. There is a discussion of this particular topic in [17]. Agarwal et al. argue that while CLIP and CLIPL methods provide evidence of a miRNA and target interaction, this is not necessarily evidence of any target repression taking place; they hypothesize that the miRNA-argonaute complex may bind noncanonical sites too transiently to mediate effective repression. The researchers here then attribute the larger number of noncanonical binding sites detected to two primary factors: (1) That noncanonical sites, having a lower information content than the canonical site, appear more often and (2) That variable cross-linking and potential ligation efficiencies introduces biases that may favor noncanonical sites.

A potential concern regarding this hypothesis is that evidence from both [5] and [7], which indicate that noncanonical miRNA binding involves a much more extensive range of miRNA nucleotides (across the length of the whole molecule) than previously thought, and also a greater number of nucleotides than previously thought—including many noncanonical sites with 3' compensatory interactions. In these instances, assuming more extensive binding would likely lead to more stable binding, there is no reason to suppose that noncanonical miRISC (miRNA-RNA induced silencing complex) binding would be as transient as suggested. However, not all noncanonical sites possess compensatory interactions. An interesting analysis would be to compare the efficacy of CLIPL identified noncanonical sites differentiated into subsets of those sites containing and not containing compensatory 3' or mid-transcript interactions. The efficacy of these two subsets of target sites could be compared either supporting or not supporting the hypothesis that more transient interactions (i.e., those with less extensive nucleotide binding) are less effective, therefore potentially explaining the inefficacy of noncanonical sites without compensatory binding.

An alternative explanation has been raised [32], which hypothesized that noncanonical sites, due to their limited sequence conservation across species, may act as evolutionary intermediates between non-sites and canonical targets sites, with a selection pressure toward the formation of the canonical site, which could potentially explain their detection in CLIP and CLIPL studies. This is an interesting hypothesis; however, it would still be hard to explain why the hypothesized evolutionary intermediates appeared in such a large number. Thousands of noncanonical sites have been identified, sometimes surpassing the number of canonical

interactions identified within the same experiment [5, 7]. If all of the identified noncanonical sites are evolutionary intermediates, this would indicate a very slow rate of miRNA target evolution. It also may not fully explain why some noncanonical sites do mediate target repression, albeit with 3' compensatory interactions. Therefore, suggesting that noncanonical sites are precursors to canonical sites, may not be a satisfactory explanation for the role of noncanonical sites in miRNA targeting.

Another potential explanation for the proposed nonfunctionality of noncanonical sites is the method by which site functionality is assessed. One of the strongest arguments against the efficacy of the noncanonical site [17] assesses the functionality of candidate sites using the microarray protocol described earlier. A key assumption in the use of this experimental protocol is that the effects of miRNA targeting can be gauged by quantifying transcript expression levels. However, it is known that miRNA can target and act on a transcript, without a subsequent change in gene expression levels in the case of translational inhibition—in which case the miRNA inhibits translation of its target, without causing the decay of that target [33]. It can be hypothesized that translational inhibition is a large, and perhaps even the primary mechanism by which miRNA effects change on transcripts via the use of noncanonical sites. In which case, the methods of benchmarking target site efficacy identical or similar as to that described in [17] would dramatically underestimate the efficacy of those sites. An assessment of the existence and prevalence of translational inhibition with respect to noncanonical sites would help to address this issue, and lead to a more accurate assessment of the efficacy of noncanonical sites with regard to all potential types of miRNA interaction.

There must be a great effort to fully understand and identify why these discrepancies exist—and furthermore, why there appears to be a great number of apparently nonfunctional noncanonical sites. Further research is needed to untangle this problem, in particular to better explore and test the hypotheses that have just been discussed [17, 32]. This general understanding of miRNA targeting is necessary to produce better prediction algorithms, and also to enable greater utilization of the large CLIP and CLIPL datasets already available.

Although great advances have been made in both experimental identification and the computational prediction of miRNA targets in recent years, there are additional ways in which advances can be made.

Primarily, there are outstanding issues, which have been generated from the advent of CLIP and CLIPL methods. As discussed, the identification of the exact role and evolution of noncanonical sites is needed. Another outstanding issue is the low efficiency of both cross-linking and ligation steps of CLIPL protocols. It cannot be reliably assumed that there is a completely uniform distribution of both cross-linking and ligation efficiencies across the entire

miRNA:mRNA input space. It may be entirely possible that either or both cross-linking and ligation introduce biases toward particular inputs (e.g., particular nucleotide enrichments, sequences, UTR lengths, etc.). Complete cross-linking and ligation profiles are needed to better understand this particular issue.

In [5], it is also revealed a high abundance of interactions that were not miRNA:mRNA interactions. In particular, large numbers of both rRNAs (12.7% of all reads) and tRNAs (38.5% of all reads) were pulled down along with AGO1 proteins. It may be the case that the relatively large proportions of rRNA and tRNA interactions are caused by the high abundance of these RNA molecules in the cell. Otherwise, reported interactions could reflect some gene regulatory activity of these molecules. tRNAs are known to be degraded to form tRNA-derived RNA fragments (tRFs) [34], some of which are known to repress mRNA targets [35] and also to associate with Dicer [36] and Argonaute [37] proteins. More research is needed to elucidate the exact nature of Ago and tRNA or tRF interactions.

To better understand target identification, there must also be a better understanding of biological processes that occur downstream of the initial miRISC and mRNA binding event. mRNA cleavage only occurs in a minority of instances [9] in animal species. Similarly, it has been shown that transcripts can be repressed independent of mRNA decay [33]. Many interaction mechanisms between the miRISC complex and targeted transcripts have been hypothesized [38]. While greater knowledge of these processes may not provide a significant amount of information relating to the initial targeting event, it may help explain the great range of efficacies for putative miRNA target sites—including those sites that are either marginally, or completely nonfunctional.

There is also the question of examining the individual read counts of given miRNA:mRNA interactions. Results from the few CLIPL studies that have been published indicate that there is a large range in the extent of chimeric read support for interactions. In [5], most hits are only supported by one chimeric read, while some hits are supported by up to 20 chimeric reads. It is probable that there is a large intersection between this particular topic, and the issue of cross-linking and ligation efficiencies. Sites that may be preferentially cross-linked and ligated will produce a higher number of chimeric reads. You would also expect there to be a higher number of chimeric reads for highly expressed mRNA and miRNA. It could also be the case that sites with a higher number of chimeric reads may bind the miRISC complex for a greater duration than other sites. It is important, but difficult, to elucidate in particular circumstances whether the contribution is because of an experimental bias or an intrinsic property of the interaction. Again, a profile of both cross-linking and ligation efficiencies across a range of inputs is needed to make the necessary distinction.

Computational methods for miRNA target prediction can only ever be as effective as the data used to inform, and in some cases, directly train those algorithms. Therefore, to produce the best computational algorithms requires use of high-quality experimental data, which is well understood by researchers. It was shown in the performance review that prediction algorithms trained on CLIP or CLIPL data showed mixed results when compared to algorithms trained on microarray data. Theoretically, it could be expected for the CLIP and CLIPL algorithms to exhibit better performance because of the higher quality experimental data used to train these prediction methods. The mixed performance of these methods could partially be explained by the fact that most methods are tested using microarray transfection data, which will likely confer a bias to prediction methods trained on microarray data. Additionally, addressing all of the issues raised in the discussion, concerning unanswered questions relating to the experimental data, will enable a clearer understanding of the novel CLIP and CLIPL methods. A greater understanding of methodological issues and core biological processes will hopefully enable researchers to develop a new generation of miRNA target prediction algorithm with distinctly improved performance statistics to what has been seen previously.

Furthermore, computational biologists as well as constantly expanding their knowledge of miRNA biology may also find it helpful to increase their understanding of machine learning methods. A greater knowledge of the theory of machine learning can only help researchers produce more effective algorithms. Principally, it should be understood that what is needed for effective machine learning is clean data, large amounts of data, and well-designed algorithms [39].

Even some of the best performing algorithms available still only account for up to 14% of the variability in some experimental data sets [17], leaving plenty of scope for further improvement in miRNA target prediction. Ultimately, more successful target prediction will reduce the time spent by experimental miRNA biologists validating putative targets, and thereby increase the time spent investigating the precise roles of miRNAs in the cell and in the organism.

7 Conclusion

Great advances have been made in research relating to miRNA targeting in animals. The advent of CLIP and CLIPL studies has enabled for the first time the direct discovery of miRNA targets on a transcriptome-wide scale. The few computational algorithms trained on these methods demonstrate a mixed performance when compared with prediction algorithms trained on microarray data.

CLIP and CLIP-L are still relatively new methods. Many ambiguities and uncertainties relating to the novel datasets produced by these methods have been raised in this article. A thorough understanding of this data, its relationship to older experimental data, and a good understanding of machine learning principles is the key to ensuring the success of the next generation of miRNA target prediction algorithms.

References

1. Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19(1):92–105
2. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136(2):215–233
3. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* 460(7254):479–486
4. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp AC, Munschauer M, Ulrich A (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141(1):129–141
5. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153(3):654–665
6. Grosswendt S, Filipchuk A, Manzano M, Klironomos F, Schilling M, Herzog M, Gottwein E, Rajewsky N (2014) Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol Cell* 54(6):1042–1054
7. Moore MJ, Scheel TK, Luna JM, Park CY, Fak JJ, Nishiuchi E, Rice CM, Darnell RB (2015) miRNA–target chimeras reveal miRNA 3′ [prime]–end pairing as a major determinant of Argonaute target specificity. *Nat Commun* 6
8. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* 320(5883):1643–1647
9. Ameres SL, Zamore PD (2013) Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol* 14(8):475–488
10. Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75(5):843–854
11. Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* 75(5):855–862
12. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A (2004) A combined computational–experimental approach predicts human microRNA targets. *Genes Dev* 18(10):1165–1178
13. 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
14. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2004) MicroRNA targets in *Drosophila*. *Genome Biol* 5(1):R1–R1
15. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. *PLoS Biol* 2(11):e363
16. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res* 36(suppl 1):D149–D153
17. Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005
18. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 27(1):91–105
19. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of *Isy-6* and other microRNAs. *Nat Struct Mol Biol* 18(10):1139–1146
20. Khorshid M, Hausser J, Zavolan M, van Nimwegen E (2013) A biophysical miRNA–mRNA interaction model infers canonical and noncanonical targets. *Nat Methods* 10(3):253–255

21. Gumienny R, Zavolan M (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res* 43(3): 1380–1391
22. Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics* 28(6):771–776
23. Wang X (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-Ligation studies. *Bioinformatics* 32(9): 1316–1322
24. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, Fevgas A (2015) DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions. *Nucleic Acids Res* 43(D1): D153–D159
25. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH (2011) starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res* 39(suppl 1):D202–D209
26. Li JH, Liu S, Zhou H, Qu LH, Yang JH (2013) starBase v2. 0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 42:D92–D97
27. Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M (2005) Silencing of microRNAs in vivo with ‘antagomirs’. *Nature* 438(7068):685–689
28. Wang T, Xie Y, Xiao G (2014) dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biol* 15(1):1
29. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126(6): 1203–1217
30. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP (2008) The impact of microRNAs on protein output. *Nature* 455(7209): 64–71
31. Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB, Leslie CS, Rudensky AY (2012) Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol Cell* 48(5): 760–770
32. Seok H, Ham J, Jang ES, Chi SW (2016) MicroRNA target recognition: insights from transcriptome-wide non-canonical interactions. *Mol Cells* 39(5):375–381
33. Bazzini AA, Lee MT, Giraldez AJ (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* 336(6078): 233–237
34. Lee YS, Shibata Y, Malhotra A, Dutta A (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev* 23(22): 2639–2649
35. Wang Q, Lee I, Ren J, Ajay SS, Lee YS, Bao X (2013) Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection. *Mol Ther* 21(2):368–379
36. Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JW, Green PJ, Barton GJ, Hutvagner G (2009) Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* 15(12):2147–2160
37. Haussecker D, Huang Y, Lau A, Parameswaran P, Fire AZ, Kay MA (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA* 16(4):673–695
38. Morozova N, Zinovyev A, Nonne N, Pritchard LL, Gorban AN, Harel-Bellan A (2012) Kinetic signatures of microRNA modes of action. *RNA* 18(9):1635–1655
39. Forsyth R (1988) Machine learning: Principles and techniques. Chapman & Hall Ltd., New York

Chapter 14

The UEA Small RNA Workbench: A Suite of Computational Tools for Small RNA Analysis

Irina Mohorianu*, Matthew Benedict Stocks*, Christopher Steven Applegate, Leighton Folkes, and Vincent Moulton

Abstract

RNA silencing (RNA interference, RNAi) is a complex, highly conserved mechanism mediated by short, typically 20–24 nt in length, noncoding RNAs known as small RNAs (sRNAs). They act as guides for the sequence-specific transcriptional and posttranscriptional regulation of target mRNAs and play a key role in the fine-tuning of biological processes such as growth, response to stresses, or defense mechanism.

High-throughput sequencing (HTS) technologies are employed to capture the expression levels of sRNA populations. The processing of the resulting big data sets facilitated the computational analysis of the sRNA patterns of variation within biological samples such as time point experiments, tissue series or various treatments. Rapid technological advances enable larger experiments, often with biological replicates leading to a vast amount of raw data. As a result, in this fast-evolving field, the existing methods for sequence characterization and prediction of interaction (regulatory) networks periodically require adapting or in extreme cases, a complete redesign to cope with the data deluge. In addition, the presence of numerous tools focused only on particular steps of HTS analysis hinders the systematic parsing of the results and their interpretation.

The UEA small RNA Workbench (v1-4), described in this chapter, provides a user-friendly, modular, interactive analysis in the form of a suite of computational tools designed to process and mine sRNA datasets for interesting characteristics that can be linked back to the observed phenotypes. First, we show how to preprocess the raw sequencing output and prepare it for downstream analysis. Then we review some quality checks that can be used as a first indication of sources of variability between samples. Next we show how the Workbench can provide a comparison of the effects of different normalization approaches on the distributions of expression, enhanced methods for the identification of differentially expressed transcripts and a summary of their corresponding patterns. Finally we describe individual analysis tools such as PAREsnip, for the analysis of PARE (degradome) data or CoLLide for the identification of sRNA loci based on their expression patterns and the visualization of the results using the software. We illustrate the features of the UEA sRNA Workbench on *Arabidopsis thaliana* and *Homo sapiens* datasets.

Key words Small RNA (sRNA), microRNA (miRNA), High throughput sequencing, UEA sRNA Workbench, Normalization, Differential expression, sRNA loci, Degradome analysis

*The authors wish it to be known that the first two authors should be regarded as joint first authors.

1 Introduction

RNA silencing [1, 2] is defined as the highly conserved, sequence-specific regulation of gene expression triggered by the detection of double-stranded RNA (dsRNA) in the cells [3]. The resulting small RNAs (sRNAs) include the intensely studied microRNAs (miRNAs, 21mers) [4, 5] and other classes of small interfering RNAs (siRNAs) like trans-acting siRNAs (ta-siRNAs, 21mers) [6] and heterochromatin siRNAs (hc-siRNAs, 24mers) [7] in plants or the piwi-interacting siRNAs (piRNAs, 28–30mers) [8] and the repeat associated siRNAs (ra-siRNAs, 21mers) [9] in animals. Depending on the degree of complementarity between the sRNAs and their target messenger RNA (mRNA), the posttranscriptional RNA silencing functions by repressing the mRNA translation (perfect complementarity at the 5' end of the miRNA, also known as seed region [10], poor for the 3' end) or by cleaving the mRNA (high complementarity throughout [11]). The less well-understood transcriptional regulation makes use of siRNAs (e.g., hc-siRNAs organized in compact loci) to induce chromatin-dependent gene silencing.

High throughput sequencing (HTS) technologies [12] have become the de facto standard for the exploration of the sRNA populations within various biological samples, from the investigation of plant traits to clinical diagnosis [13]. Moreover, with the rapid technological advances and improvements in both efficiency and cost, the increased number of replicates, samples within an experiment and achieved sequencing depth often poses a considerable challenge to the molecular biology community.

In this chapter we present the UEA small RNA Workbench [14, 15], a Java based, downloadable suite of tools designed for the analysis of sRNA HTS data in a user-friendly and accessible environment. The project itself was designed as the natural successor to the web based UEA sRNA Toolkit [16]. While web based interfaces are still relatively popular with bioinformatics services, they pose several problems with general maintenance, such as regular updates of reference sequences and can therefore become restrictive for the analysis of datasets of newly sequenced organisms. The Workbench was initialized in 2009 with a basic set of helper programs and analysis tools and has since grown with the addition of several new tools such as PAREsnip [17] and CoLIde [18] as well as enhanced features added to the original set.

The tools can be used to perform a single task over a single or set of sRNA files, e.g., trimming adapter fragments (*see* Subheading 3.1) and searching Parallel Analysis of RNA ends (PARE data) for genome wide targets to an entire set of sRNA data (*see* Subheading 4.3) or can be linked into an analysis pipeline. For example, differential expression analysis should be preceded by a series of quality check and normalization procedures to reduce the number of false positives. This type of analysis is made possible in the Workbench via a

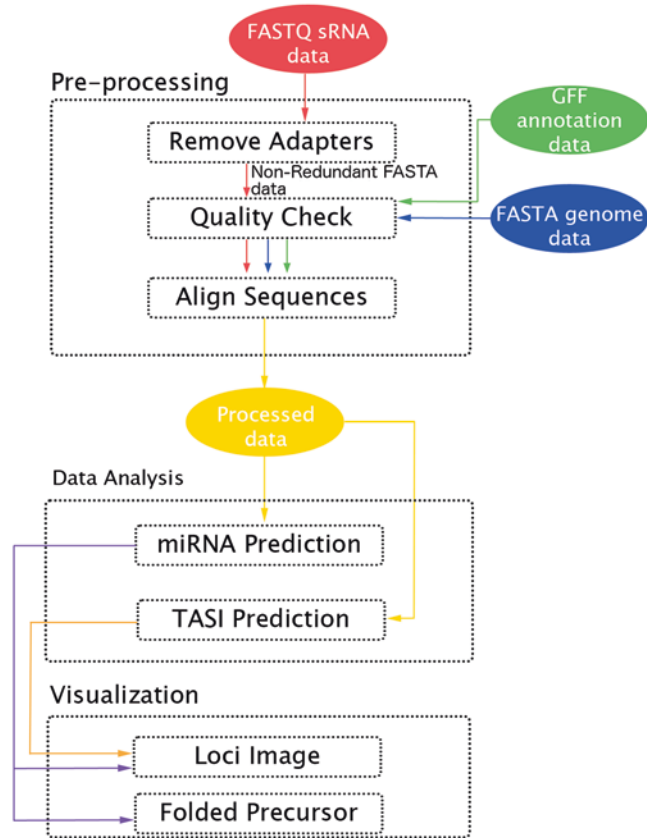


Fig. 1 Example of analysis workflow achieved in the Workbench, The initial preprocessing of the raw data includes the trimming of adapter sequences and general data quality checks. Next, new miRNA and ta-siRNA loci can be identified and visualized. Each stage of the workflow is configurable; the exemplified processing steps can be replaced with other Workbench tools

collection of preprogrammed pipelines that provide a complete analysis flow for an sRNA dataset. A typical workflow schematic for an sRNA data analysis is presented in Fig. 1.

The Workbench can also be accessed from the command line for insertion into existing bioinformatics pipelines and use on remote machines. A complete set of commands and usage instructions can be found at the website <http://srna-Workbench.cmp.uea.ac.uk>. A set of video tutorials on the use of the Workbench can be found at the website <http://srna-Workbench.cmp.uea.ac.uk>

2 Materials

The datasets used to illustrate the use of the Workbench include an experiment in *Arabidopsis thaliana* (GSE57936) investigating the ability of the plant to avoid inappropriate silencing of its own

coding genes by the silencing pathway used to in defense of viral genes [19]. As part of the experiment a wild type (col0) was sequenced (GSM1397834, GSM1397835, GSM1397836). We use the col0 sample taken from GSE57936 to exemplify the use of the TAS prediction tool in Subheading 4.1.2 and miRNA prediction (Subheading 4.1.1) along with the options for visualization of sequence data.

To demonstrate the quality checking and normalization pipeline as well as the differential expression tool we used an experiment on the effect of hypoxic conditions on MCF7 cells in *Homo sapiens* [20] (GSE47532) the experiment is organized into a time series of four time points, each with two biological replicates, Normoxia (N00), Hypoxia at 16 h (H16), Hypoxia at 32 h (H32), and Hypoxia at 48 h (H48). We used the N00 (GSM1151671, GSM1151672, GSM1151673) and H16 (GSM1151674, GSM1151675) sample to investigate differential expression, quality control, and normalization using the Workbench.

In addition to this, sRNA samples taken from [21] (GSM343000) and the degradome from [22] (GSM280226) along with the TAIR10 cDNA 20110103 transcriptome [23, 24] are used to demonstrate the PAREsnip program and its various features.

3 Methods

3.1 Raw Data Processing (Adapter Trimming)

Raw sequencing data is given in FASTQ format. Prior to an analysis, the data must be converted into FASTA format and the sequencing adapter trimmed. The Workbench provides a preprocessing tool called Adapter Removal for this task included in the Helper Tools menu. In the same location sequence alignment tools such as PatMaN [25] or Bowtie [26] can be found as well as a tool for filtering raw sequence data. We now present how the Workbench can be used for the initial preprocessing of raw sequence data.

Most of Illumina sRNA-seq experiments are done as single end, 50 nt reads [27]. Since the majority of known sRNAs are less than 30 nt long [10, 11, 27], the raw reads always contain a fragment of the 3' adapter fragments, which should be removed prior to further analysis (*see* Fig. 2). The 5' adapter is commonly used as sequencing primer and hence is not present in the raw reads.

The input of this tool is the 3' adapter sequence and the original FASTQ file. The adapter sequence is identified and trimmed using an exact string match of the first n nucleotides (default eight). Lower values for n have a higher the chance for allowing incorrect reads through to the output if the matching is achieved by chance. The Workbench can also process 454 data, for which both 5' and 3' adapters must be specified.

In addition to processing datasets produced using the standard protocols, the Workbench also has the ability to trim adapters from

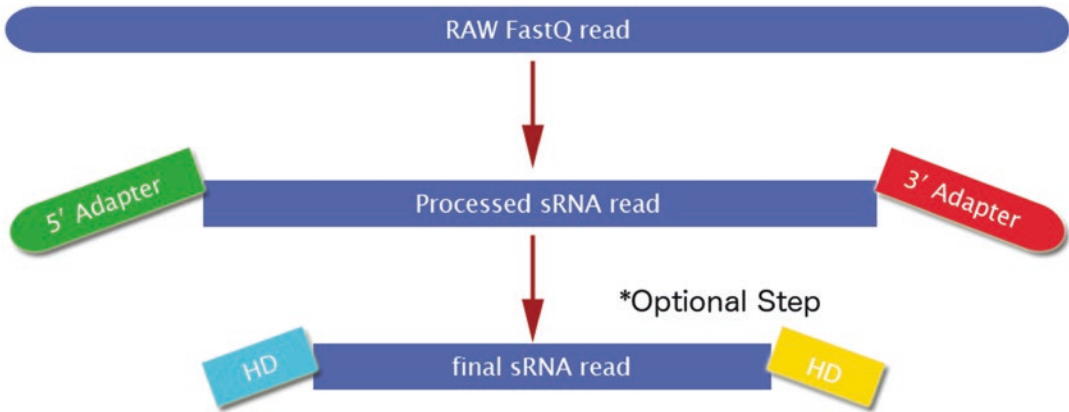


Fig. 2 Adapter trimming for single end sRNA-seq data. The reads processed by HTS instruments contain adapter fragments that must be removed prior to further processing. Illumina instruments use the 5' adapter as a sequencing primer, hence, for this type of sequencing output, only the 3' adapter needs to be trimmed (the area to the immediate right of the red line). Whereas 454 sequence data will contain both adapter fragments and will need 5' and 3' trimming (also available in the Workbench). The Workbench tool is also configured for the processing of HD libraries, which consists in an additional trimming of the 4 nt HD signatures from both the 5' and 3' ends of the read

samples created using the HD protocol [28, 29]. For this, the first and last 4 nt (the HD signatures) are identified and trimmed.

In a sample, some sRNAs are sequenced in multiple copies; however, to optimize subsequent processing the only required information for an sRNA is the nucleotide sequence and its abundance (number of copies in a sample). To prepare the data for further analyses, the Workbench creates a “nonredundant” version of each input file where each unique sequence is presented once accompanied by its raw abundance.

The output of the adapter removal step includes size class distribution plots showing the number of redundant and nonredundant reads obtained for each size class and a table with the raw information used to produce the plots.

After the data has been preprocessed, the Workbench provides a series of approaches for checking the quality of the data. The following section describes how to investigate the quality of an sRNA dataset using the Hypoxia dataset as example.

3.2 Quality Checking

An important, but often overlooked step is to examine the quality and consistency of the samples in a dataset. The rationale is that spurious artefacts introduced with poor(er) quality samples can have an impact on the resulting reports such as prediction of novel miRNAs or calculation of differential expression. The Workbench produces several initial qualitative and quantitative reports on the raw data; this step is reiterated after the normalization of expression levels to prove the efficiency of the selected approach [14].

The quantitative assessment includes the size-class distributions for the nonredundant (unique) and redundant reads and the size-complexity plots (*see* Fig. 3 (i)), defined as the ratio of nonredundant to redundant reads [30].

The next check is the analysis of nucleotide distributions (presented for each size class in the given sample, *see* Fig. 3, subplot (ii)), which can reveal overrepresented sequences or common motifs throughout the dataset. These plots can also be used to reinforce characteristics of sRNA classes such as the T preference at the 5' end of miRNAs (which are a large proportion of the 21mers in plant, and 22mers in animals), or the A preference for heterochromatin siRNAs (24mers, in plants).

In addition the Jaccard similarity index [30] (*see* Fig. 4 (iii)) is calculated for the top N most abundant reads (N is user defined, default 500), to avoid the noise range, which is characterized by spurious variability [18, 30].

A qualitative assessment of consistency of the replication can be made on replicate versus replicate scatter plots and MA plots (Bland–Altman plots) [31] with each point representing the abundance of a sRNA in each replicate (all possible combinations of replicate pairs are assessed individually). Figure 4 (ii) shows an example of replicate-to-replicate MA plots produced using the *H. sapiens* dataset.

Finally, to investigate reproducibility between samples, across size classes, box plots showing the distribution of fold changes, calculated in $\log_2$ scale, can be created (*see* Fig. 4, subplot (v)). Robust datasets depict tight distributions for all size classes, symmetric either on 0 ($\log_2$) or a normalization factor [14].

The MA and scatter plots and the size- differential expression plots can be created for either all sequences, or on a selection of reads, corresponding to specific annotations. The latter can be obtained using a General Feature Format (GFF) file containing annotation information specific to an organism, including transcript identifier, genomic location, and annotation type.

3.3 Filtering

An important step in data preprocessing is to select or filter out sequences based on specific properties or annotations (which can be done using the filter tool in the Workbench). This can improve later analyses by reducing false positives derived from an unbalanced distribution of reads in different annotations (e.g., transfer and ribosomal RNA contamination) or decreasing the drain on various system resources such as memory use or processor time. The characteristics that can be identified and removed include: the removal of reads with low sequence complexity (e.g., consisting of less than three distinct nucleotides or showing repetitive dimer and trimer motifs), and the removal of sequences within a predefined size range (e.g., all sequences shorter than 19 nt or longer than 32 nt). The latter filtering is efficient especially when investigating

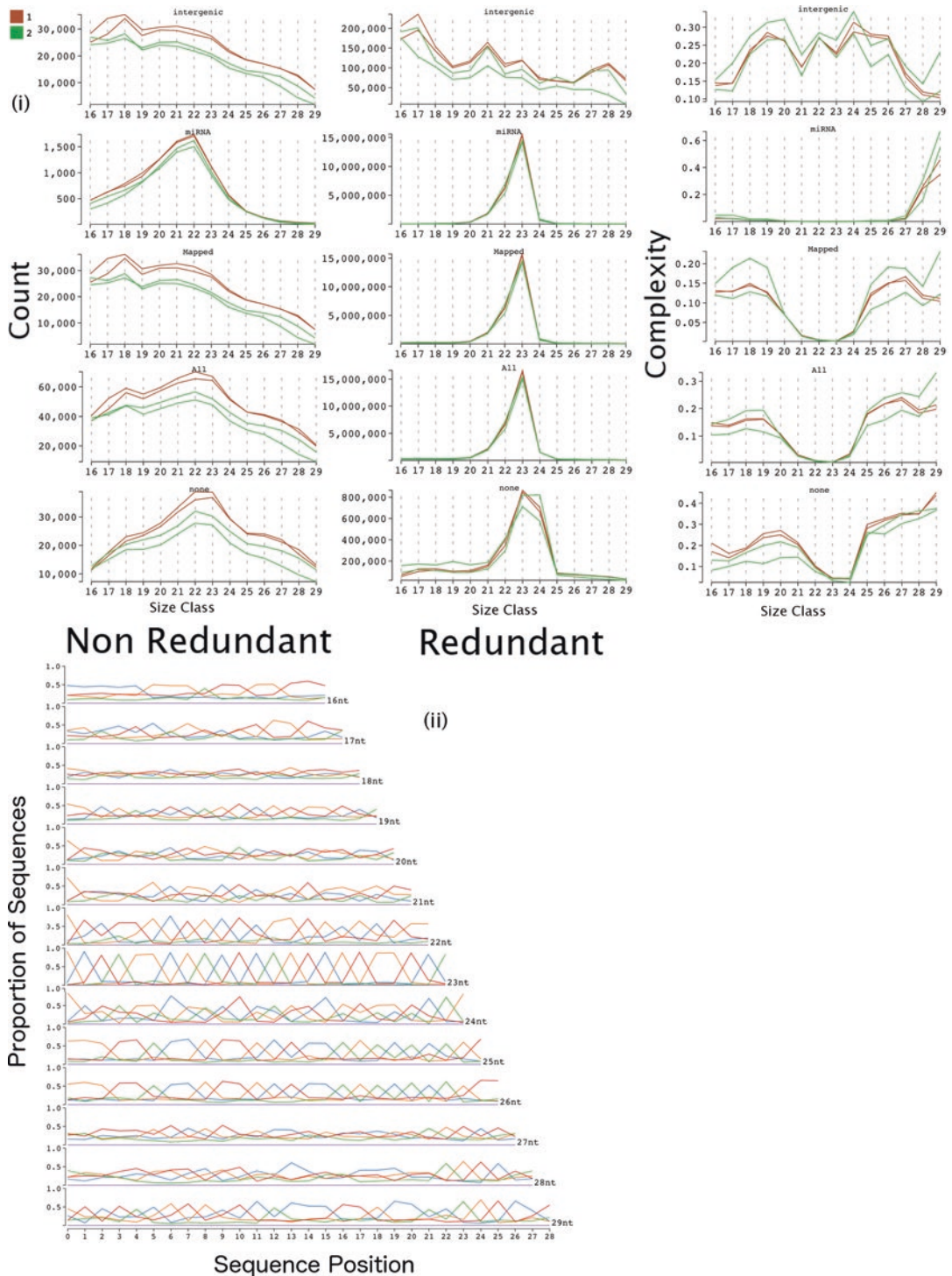


Fig. 3 Size class distributions, complexity and nucleotide composition plots created on an *Homo sapiens* dataset (i) Size Class for redundant and nonredundant reads and complexity distributions for reads matching to intergenic regions and mature miRNAs, genome matching reads, all reads and non-genome matching reads; These are in line with the expected prevalence of miRNA reads in *H. sapiens*. The drop in complexity for 22mers and 23mers reflects the existence of a small number of highly abundant reads (ii) nucleotide composition for reads with lengths between 16 nt and 29 nt. The T-preference for 21mers and 22mers is visible

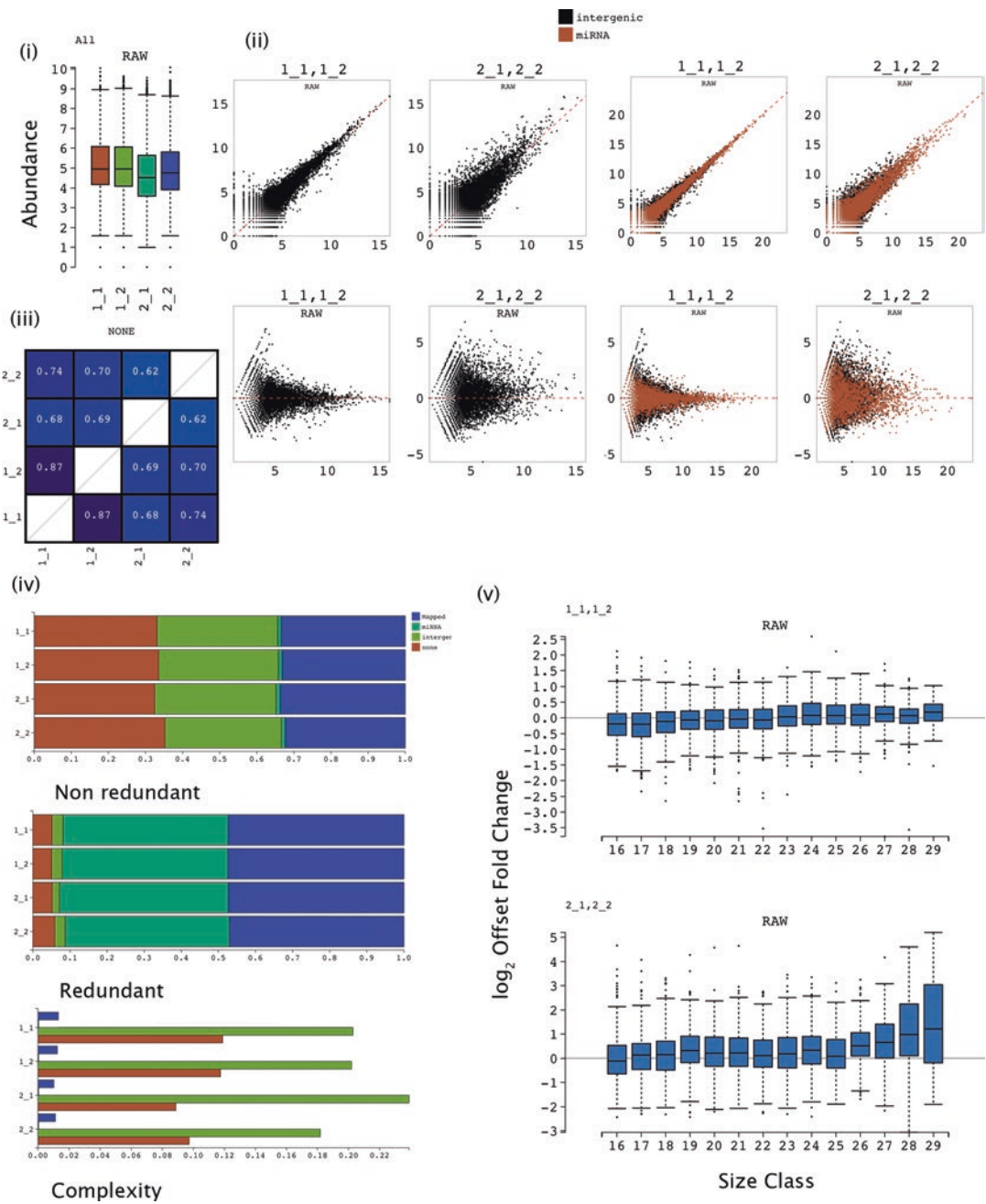


Fig. 4 Quality check plots created on the *Homo sapiens* dataset. (i) Abundance distribution boxplots, for two samples, with two replicates each (the first two boxplots correspond to the first sample, the next correspond to the second sample) (ii) Scatter plots and MA plots for replicate-to-replicate comparisons. Annotations are overlaid; mapped reads are shown in *black*, unmapped in *green* and miRNA in *brown*, in scatter plots the X and Y-axes represent log₂ expression level; in MA plots the X-axis shows the average abundance and the Y-axis indicates fold change (iii) Jaccard similarity index, calculated on the top 500 sequences represented as a heatmap (iv) Mapping Quality plots showing the percentage of mapped, unmapped and annotated reads in redundant and nonredundant form as well as the complexity plots, (v) Size-differential expression represented as boxplots of the offset Fold Change distribution per size class

sequences that fall within a narrow size range such as miRNAs that are between 21 and 23 nt in length. Additional filtering may include discarding sequences with the abundance below a predefined threshold, which may help with the removal of random degradation products (noise) that carry no biological significance.

3.4 Normalization of sRNA Expression Levels

For any comparative sRNA-seq study, an essential step prior to the identification of differentially expressed transcripts is the normalization of expression levels [McCormick, Dillies], i.e., the removal of technical variation while preserving the characteristics of the samples. This step of the analysis should be preferably applied only to samples, which have passed the initial set of quality checks.

Most normalization studies focus on mRNA-seq data [32–34]. Several approaches have been adapted for sRNA-seq data. Six of these were implemented in the Workbench (and are discussed in the next sections), allowing the user to compare the output from different methods and conclude which approach is appropriate for the given dataset.

We propose in [14] that no single normalization is suitable for all datasets due to various causes including technical variation or differences in the sRNA populations which are compared. To help the user choose a suitable normalization a set of quality check plots are produced post-normalization for visual inspection. Based on the assumption that, between treatments, most transcripts should not be differentially expressed, preferred normalization approaches are those which (i) minimize the fold change variation across abundances in the MA plots and (ii) show no differential expression across all size classes in the size-differential expression plots.

3.4.1 Scaling Normalization

Scaling normalizations are based on the identification of a factor, which can bring the total counts (sum of abundances of all reads in a sample), in all samples, to the same total. The only difference between the various scaling methods is how the scaling factor is calculated.

The Workbench provides four approaches for scaling normalizations; (1) Reads Per Million (RPM) for which the scaling factor is the ratio between the sequencing depth of each sample and the a priori fixed total; due to the rapid increase in sequencing depth the fixed total is no longer 1 M, but it is determined as the mean/median of the sequencing depth of the samples forming the dataset [35]; (2) the trimmed mean of M-values (TMM) [36], adapted to suit sRNA-seq data; to robustly summarize the observed M values, both the M and A values are trimmed before calculating a weighted average. Precision (inverse of the variance) weights are used to account for the fact that log fold changes from reads with higher abundance have lower variance (logarithmic scale) compared to those of lower abundance; (3) Upper Quartile normalization is similar to RPM; the scaling factors are calculated using only the

reads with abundance in the upper quartile [37]. Finally, the DESeq2 method [38, 39], which estimates the sample corrections by taking into account both the sequencing depth of each library and the abundance (and variance) of a transcript under each condition. The per-transcript variance is assumed to be a function of the mean that is approximated by an empirical fit to the data. P-values are calculated using a method analogous to a Fisher's exact test, using a 2×2 contingency table, but instead of assuming that the probabilities follow a hypergeometric distribution, they follow a negative binomial distribution parameterized from the mean and the estimated dispersion.

3.4.2 Rank Based Normalization

A version of quantile normalization [40] adapted to sRNA-seq data [14] is also included in the Workbench. It works by imposing the same distribution of ranks on all samples. The sRNA-seq adjustments are: (1) sequences with the same abundance prior to normalization must have equal abundances post-normalization; the adjusted abundance is the average of the normalized abundances after standard quantile normalization; (2) reads with abundance 0 will have abundance 0 post-normalization 0 (regardless of the normalized expression after the standard approach is applied).

3.4.3 Subsampling Normalization

The subsampling normalization technique implemented in the Workbench is an adapted version [41] of the method presented in [42]. The approach consists of two steps: first, the sequences are sampled in decreasing proportions until either the size class distribution of the subsample is significantly different from the original or the proportion equals the lowest sample size; second, the reads are repeatedly subsampled to a fixed total, equal to the lowest sample size and the variability of each subsamples of abundances is tested using bootstrapping. If the resulting variability between them is low (i.e., the number of unique fragments varies by less than 0.01% of the total number of unique fragments in that sample), a random sample is selected and used for the subsequent steps of the analysis.

3.4.4 Normalization Selection

In order to assess which normalization is appropriate for the dataset being investigated, the Workbench recreates all of the quality check plots using the normalized expression values for each normalization technique applied to the data. These plots can be visually inspected and an appropriate approach can be selected based on the MA plot and size-differential expression boxplots. As shown in Fig. 5, users can show the plots side by side with the original raw values and any selected normalization to investigate the effect made on the data.

3.5 Identification of Differentially Expressed Reads/Transcripts

The purpose of multiple sample experiments is to link the sRNA variation in expression level to the observed phenotypes/treatments. This is commonly referred to as differential expression (DE) analysis [27, 33]. To ensure the reliability of this step, it

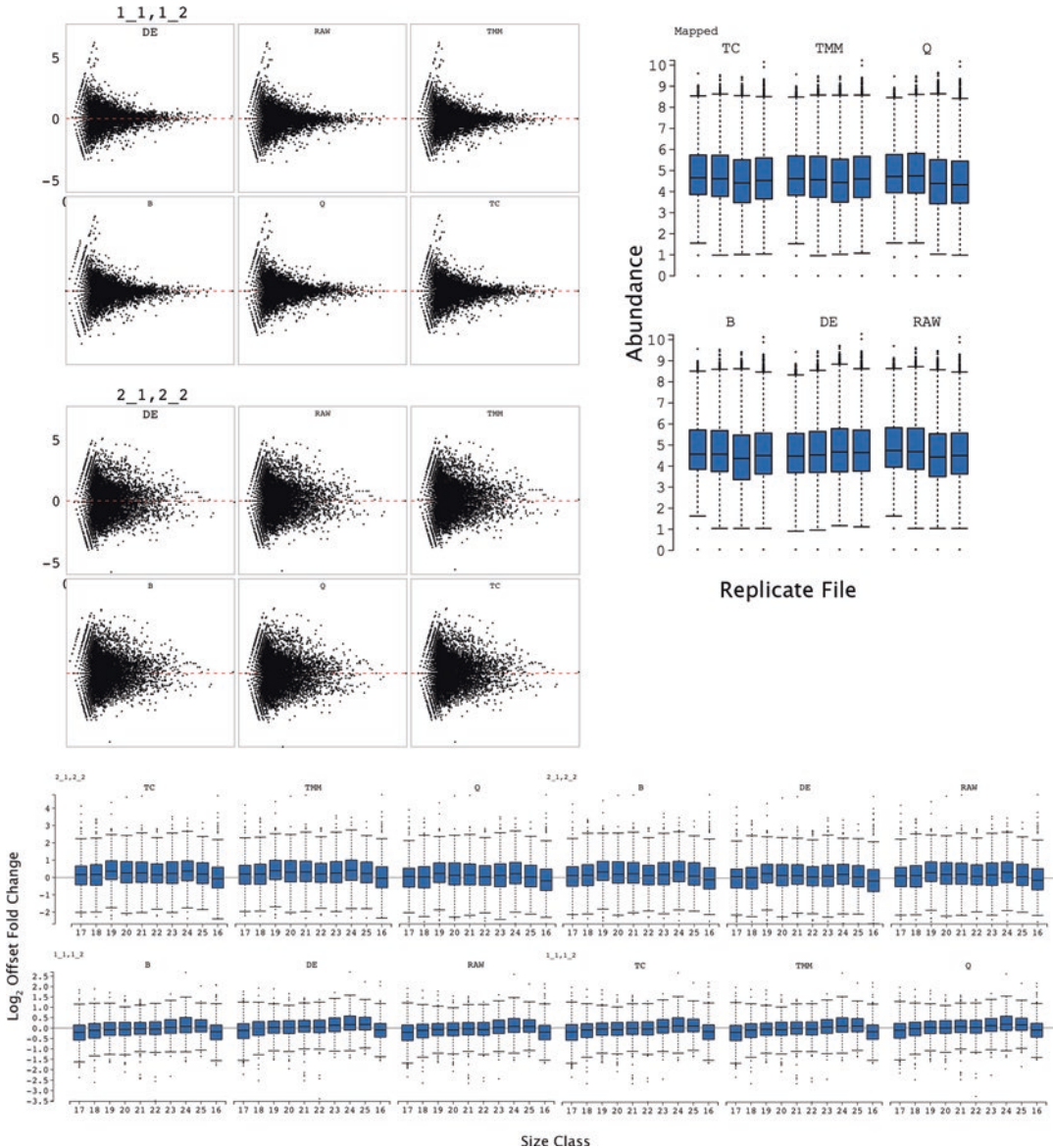


Fig. 5 Normalization Quality Check plots. (i) MA plots showing the DE variation, calculated as FC (Y-axis), against abundance (X-axis) for the different normalization approaches, (ii) abundance distributions; a robust normalization method renders these distributions comparable; (iii) size separated differential expression distributions; a dataset is normalized when all these distributions are centered on 0

is recommended that this analysis follows quality checking and normalization.

The increase in sequencing depth led to an improved quantification of sRNA expression levels, but also an increase in the random, spurious reads which could be classified as noise. This has had an effect on the distribution of differential expression calculated using a fold change approach (e.g., changes in expression

between 1 and 10 have the same fold change as a change between 100 and 1000). The direct consequence, derived from the attempt to focus on the biologically relevant candidates, was the introduction of fold change thresholds (e.g., we accept as DE reads those the highest fold change). However, it was observed that these led to false negatives (i.e., sRNAs which should have been called DE, but were not) in the high abundance range [30] and false positives (i.e., sRNAs which were not DE, but we called DE), in the medium abundance range. To counter the issue of high fold changes observed for low abundance reads, the Workbench uses an improved offset fold change calculation with individual offset values derived from the input dataset. The default threshold of DE is 2 (in logarithmic scale $\log_2(2) = 1$).

3.5.1 Usage of an Offset for Calculating the Differential Expression

After normalizing the expression levels, to avoid the distortion of the distribution of DE caused by large fold changes at low abundances we calculate the fold change using an offset which cancels the spurious variation at noise level. Practically the offset can be interpreted as the noise threshold since the fold change for a sequence with abundance 0 in one sample and abundance equal to the offset in the other sample will be called DE (since

$$\log_2\left(\frac{\text{offset} + \text{offset}}{\text{offset}}\right) = 1).$$

The “noise level” (i.e., the offset) is

calculated separately for each sample of the dataset; however, for the OFC calculation, the same offset must be used for all samples; this is chosen as the maximum for each individual sample in the dataset.

The offset is calculated based on the distribution of strand bias ($SB = |p - 0.5 + n - 0.5|$, where p and n are the proportions of reads matching to the positive or negative strand respectively) in pre-defined overlapping windows across the genome. The strand bias can vary between 0 (there is an equal probability to observe reads from either the positive or the negative strand), and 1 (the reads are exclusively matching to one strand). As shown in [18, 30] the strand bias is high for regions (loci) with few incident reads (e.g., if only one spurious, low abundance read is incident to a window, then the strand bias will be 1). The strand bias decreases for medium abundance regions (the noise range is characterized by random hits on either positive or negative strand). Then it increases again for the high abundance regions, mainly incident with miRNA loci, characterized by few precise regions of high abundance.

The approach implemented in the Workbench is to calculate the Kullback–Leibler (KL) divergence for classes of abundances, [0,5], [5,10], [10,15], etc. For each abundance class we compare the distribution of strand bias (between 0 and 1, using subintervals of equal length) with a random uniform distribution (which is characteristic for random spurious hits, with no biological relevance).

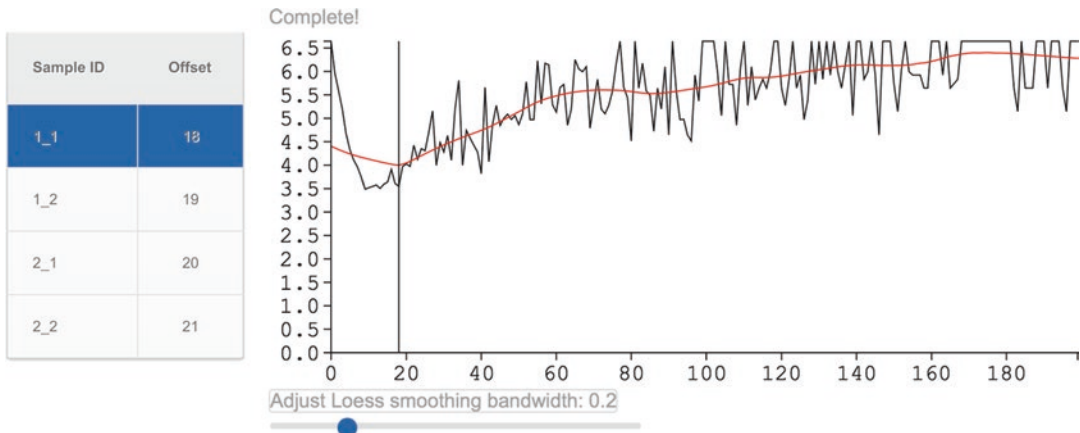


Fig. 6 Identification of an offset for the Hypoxia dataset, using the variation of strand bias, compared with a random uniform distribution using a KL divergence approach. The KL graph is smoothed using the LOESS method (the *red line*) from the Hypoxia dataset. The global minimum used as the estimation of the offset in the differential expression analysis is highlighted using the red arrow

The software then identifies the global minimum of the KL values as the offset for that sample. However, the global minimum is often found in a region with multiple local minima; to counter the effect of noisy measurements (and to avoid selecting an outlier value) a LOESS smoothing curve [43, 44] is calculated based on the KL divergence graph. The users are given the opportunity to manually adjust the smoothing factor of the curve. An example of the KL divergence graph and its resulting LOESS smoothing curve created from the Hypoxia dataset is given in Fig. 6.

3.5.2 Offset Fold Change (OFC) and Variation Patterns

For the DE analysis between two samples the users must specify first which sample will be used as reference and which will be used as “observed”. If a dataset consists of more than two samples, then the users have to specify whether the series are ordered or unordered. For the latter, all pairwise DE calculated are performed by default; however, the users can choose a universal reference sample (e.g., in a mutant series the wild type can be selected as the reference and the mutants as the observed). For the former, only the consecutive comparisons are performed (the first with the second sample, next the second with the third, etc.).

For all pairwise comparisons (regardless if the series are ordered or not), confidence intervals are calculated for each sRNA (or sRNA locus) using the replicate expression levels (in a similar approach to the one used in the CoLide tool described in Subheading 4.1.2 [18, 45]). If only two replicates are available, the confidence intervals are built on the minimum and maximum expression levels over both replicates; if more than two replicates are available, the Workbench uses Chebyshev intervals calculated on the mean and the standard deviation of all replicate expression levels.

If the dataset comprises samples for which only two replicates are available, then for all samples, for consistency, the maximal confidence intervals should be used, because the standard deviation is meaningless for two values.

The direction of DE is then calculated based on the localization of the confidence intervals and the amplitude of DE is calculated on the proximal ends of the confidence intervals that are compared as the $\log_2$ (OFC) (LOFC). The direction of DE is presented as a descriptor; {U} (upregulation), {S} (straight for overlapping confidence intervals) and {D} (downregulation) [45]. Next, for each read, a pattern is defined by concatenating the {U,D,S} descriptors resulted from each comparison [18]. The DE sequences can then be grouped on either individual comparisons or on the whole pattern (as shown in Fig. 10).

Additional information included in the output table consist of the sequence itself, the annotation or potential list of annotations if alignments overlap (if a GFF file was provided during setup), the length of the sequence, the LOFC and the normalized expression levels in each sample.

4 Small RNA Prediction and Analysis

The Workbench offers several options for downstream analysis of sRNA data. The following sections detail the steps of miRNA analyses, ta-siRNA analysis, general sRNA locus prediction and degradome assisted target prediction.

4.1 miRNAs

The biogenesis characteristics of miRNAs (the specific hairpin-like secondary structure) and the development of miRbase [46], a database containing miRNAs, structured on organisms and clades, has facilitated the computational analysis of miRNA loci. In the Workbench, we have implemented three tools for miRNA analyses: miRCat for the prediction of new miRNAs (*see* Subheading 4.1.1), miRProf (*see* Subheading 4.1.2) for determining the expression levels of miRNAs and the RNA annotation tool for creating the secondary structures (*see* Subheading 4.2).

4.1.1 miRCat

The miRNA prediction is handled in the Workbench with the **miRNA Categorization** (miRCat) tool. The input for this tool consists of a single file containing sRNAs in nonredundant form and a reference genome, both in FASTA format. The prediction of new miRNAs is done using a rule-based classification algorithm based on known features of the miRNA biogenesis.

The first stage is to group all genome-matching reads into loci (regions on the genome which act as one unit) employing on the SiLoCo algorithm (*see* Subheading 4.4.1). The parameters for this step are the minimum number of incident sequences to each locus

and the maximum allowed distance between reads within a locus. The read with the highest abundance within a locus, if it has an abundance over the user defined threshold, is then used as the putative mature miRNA. To obtain and assess the secondary structure of the putative miRNA locus regions around the selected read are retrieved from the reference genome and folded using RNAfold [47, 48]. Regions which achieve a minimum free energy (mfe) and adjusted minimum free energy ($\text{mfe}/\text{hairpin length} \times 100$) less than a user-defined threshold progress to the subsequent checks. Next, the number of unpaired bases and GC content of the putative locus are determined and compared to predefined thresholds. Additionally checks include the existence of complex loops, which are only allowed at the request of the user. Finally the randomness of the hairpin sequence is tested using Randfold [49]; the resulting p -value expresses the probability of obtaining a similar secondary structure by chance. If at least one of the regions is classified as a viable precursor sequence, then the one with the best secondary structure (minimum mfe) is presented. Next the miRNA and miRNA* regions are described.

There are 17 parameters used for the classification that can be modified prior to the processing of a dataset in order to adjust the accuracy (sensitivity or specificity) of the results. Next, we describe these parameters, provide the range of meaningful values, and the defaults.

1. *genomeHits*: the maximum number of allowed genome hits for each read, i.e., reads with more hits to the genome than the genomeHits parameter are discarded. [1, 25]; default: 16.
2. *hitDist*: parameter used for the identification of putative loci; the maximum distance between consecutive hits on the genome. [20, 1000]; default: 200.
3. *maxGaps*: the maximum number of consecutive unpaired bases in miRNA:miRNA* region which will determine the size of a inner loop. [1, 5]; default: 3.
4. *maxOverlapPercentage*: a characteristic of miRNA loci is the accumulation of reads in the miRNA:miRNA* region. This threshold sets the minimum proportion of reads incident with the miRNA:miRNA* region. [30, 100]; default = 70.
5. *maxPercentUnpaired*: the maximum percentage of unpaired bases allowed in the secondary structure of the putative miRNA locus. [1 100]; default: 50.
6. *minPaired*: the minimum number of paired bases in the miRNA:miRNA* region. [10, 25]; default: 17.
7. *maxUniqueHits*: the minimum number of unique reads incident to a locus. [1, ∞]; default: 3.
8. *maxSize*: the maximum length of incident reads. [18, 24]; default: 22.

9. *minSize*: the minimum length of incident reads [18,36]; default: 20.
10. *minAbundance*: the minimum abundance of incident reads. [1,∞]; default: 5.
11. *minEnergy*: the threshold for the accepted AMFE of the putative hairpin. default: -25.
12. *minGC*: the threshold of minimum G/C content for the mature miRNA; [1100]; default: 10.
13. *minHairpinLength*: the minimum accepted length of the putative hairpin (expressed in nt) [50,500]; default: 75.
14. *Allowing complex loops*: this is a Boolean parameter (it can be false or true); it allows or rejects hairpins containing complex loops.
15. *Orientation percentage*: a characteristic of miRNA loci is the strong strand bias, as a result of reads being produced from a single stranded RNA that can fold into a tight secondary structure. This parameter is a threshold for the percentage of reads incident to the same strand [1100]; default: 90.
16. *HairpinExtension*: the maximum length of the flanking region used for the identification of the miRNA locus. [40,400]; default: 150.
17. *P-value*: The minimum *p*-value for the precursor as reported using the Randfold program [49].

Some parameters such as orientationPercentage, minEnergy, HairpinExtension have a much wider impact on the results, while other parameters such as minGC, minSize, maxSize have a smaller discriminative power.

The output of this tool presents the characteristics of the miRNA loci (such as location on the genome and strand, miRNA and miRNA* sequences) and the calculated values of the parameters that were used for the classification (*see* Fig. 8). The output table contains the following information: (1) Chromosome, (2) Start coordinate, (3) Stop coordinate, (4) Strand/orientation, (5)

	Sequence	Size	Annotation	Fold Change	Direction	Reference Count	Observed Count	Average Count	Pattern
☐	ACTGTGGGTGTACAGCGCTGA	23	miRNA	1.1309688939420126	UP	2	1	5.538719744931026	U
☐	AGGGTTGGGTGGAGGCTCT	19	miRNA	1.1609157026967907	STRAIGHT	2	1	6.106103519763656	S
☐	CAGCAGCAATTCATTTTGA	21	miRNA	1.0323497281260434	STRAIGHT	2	1	8.7891678947765	S
☒	GGGGGGGGGGGGGGGGGGGG	22	miRNA	1.1499971300777781	UP	2	1	7.770988512037387	U
☐	GTGGGGGAGAGGCTGT	16	miRNA	1.0825578518236252	UP	2	1	7.981866368518787	U
☐	TACACTGTCTGTAAAGATGGCT	24	miRNA	1.3221154468032732	STRAIGHT	2	1	8.516617363994001	S
☐	TGGAAGACTAGTGATTTGTGTT	24	miRNA	1.013136246478907	STRAIGHT	2	1	10.415557344526032	S
☐	AAAGATAAAATTTGAAATCTGGTAGGC	28	intergenic	1.1395861460598442	DOWN	2	1	5.767207164404451	D
☐	AAGACCTGTTTGAGCTG	18	intergenic	1.6322021299353713	UP	2	1	2.695251837606938	U
☐	AAGACCTGTTGAGCTGA	19	intergenic	1.205271380356171	UP	2	1	2.8442704386896764	U

Fig. 7 A snapshot of the differential expression results table found using the Hypoxia dataset. The expression pattern column is indicated with the *red arrow*

Chromosome	Sequence	miRNA*	Hairpin Sequence				miRNA Start Coordinate	miRNA End Coordinate			
Chr1	TTTCTCTACTCTTGACAA	No	TCGGCTCACTGCTCTGCTGCTGTCGACACAGGACGAAGCTCTGCTATGTAATT ACTGAGGTATTGACGTA				28635	28655			
Chr1	TCGACACAGCTTCATCCCC	GGAATGTTGCTGGATCAGG(2) GGAATGTTGCTGGATCAGGA(1)	GTTGACG GCATGTTGCTGGATCAGG ATATTATAGATATACATGCTGTAATGTAAT GAGATCC TCGACACAGCTCTCATCCCC CCAC				78932	78952			
Chr1	TCGGCACTCGGCGACATGCT	No	GCTTCGGTGTCGGGCACTGATATATACATGCTGCTCGACACGGCAGC ATGCTGCGG				117507	117528			
miRNA Start Coordinate		miRNA End Coordinate	Orientation	Abundance	sRNA Length	Hairpin Length	Hairpin Start	Hairpin End	Hairpin Structure		
28635	28655		+ [Positive strand]	1.0	21	148	28528	28675			
78932	78952		- [Negative strand]	421.0	21	111	78926	79036			
Genomic Hits	Minimum Free Energy (MFE)		Adjusted MFE		Hairpin G/C %		miRNA* start coord		miRNA* end coord	miRBase ID	P_Val
1	-68.7		-46.41892		42.56757		-1		mir838	0.009901	
2	-48.7		-43.873875		42.342342		79008		mir165	0.009901	

Fig. 8 An example output table from a miRCat prediction run Col0 replicate 1 of the *Arabidopsis thaliana* dataset. The highlighted column shows a predicted miRNA sequence and all of the relevant statistics used in the prediction. Closer inspection reveals the sequence is found in the miRBase database and is annotated as miRNA 165

miRNA abundance, (6) Predicted mature miRNA sequence, (7) Representative sequence accession ID (miRBase) from input dataset, (8) Length of predicted mature miRNA, (9) Number of genome matches for the mature miRNA, (10) Length of predicted precursor hairpin sequence, (11) G/C % content of hairpin sequence, (12) Minimum free energy (MFE) of predicted hairpin sequence, (13) Adjusted MFE, (14) Putative miRNA* sequences, along with their abundance in the input dataset shown in brackets, (15) Hairpin Sequence (with miRNA sequence highlighted in blue and miRNA*, if present, in red), (16) Secondary structure of the hairpin in Dot-Bracket notation, (17) Hairpin start coordinate, (18) Hairpin end coordinate, (19) miRNA* start coordinate, if present (20) miRNA* end coordinate, if present.

In addition to the description of the putative miRNA loci presented in this table it is informative to visualize some of the features of the predicted loci (*see* Subheading 4.4). Each locus can be piped directly into the RNA annotation tool to display the hairpin structure and highlight the miRNA and miRNA* (if found in the input dataset) (*see* Fig. 9). Alternatively, users can visualize the miRNA locus using the VisSR tool (Subheading 4.4.3).

4.1.2 miRProf

The **miRNA Profiling** (miRProf) tool produces a list of conserved miRNAs, with their respective expression levels, in all the samples of a dataset. The inputs of this tool are either one single or a set of samples containing reads in nonredundant format, i.e., reads with their respective normalized abundances. If files with raw abundances are provided then the expression levels are normalized using the RPM method (described in Subheading 3.4.1) to the sequencing depth of each sample (if no reference genome is provided) or to the redundant number of genome matching reads (if a reference genome is provided). An additional option is the weighting of the expression levels on the number of genomic hits,

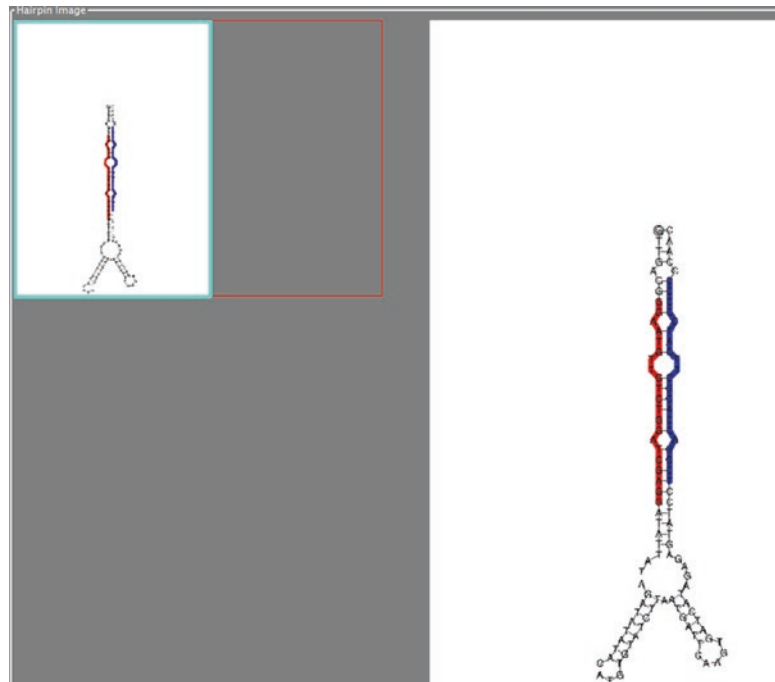


Fig. 9 The secondary structure of *Arabidopsis thaliana* miRNA 165, identified using miRCat on the *Arabidopsis thaliana* col0 dataset, as it is presented in the navigation window of the RNA annotation tool. The sequence was forwarded to the RNA annotation tool using the context menu on the miRCat result set. The miRNA and miRNA* are highlighted respectively in *red* and *blue*, the circled nucleotide denotes the 5' end of the precursor sequence

i.e., the normalized expression level is divided to the number of genomic hits. Next, these reads are aligned to miRBase (which can be downloaded automatically if an active internet connection is available) allowing up to n mismatches (n is a user defined parameter) and no gaps. Only alignments to the positive strand of the miRBase miRNAs are allowed. The output of this tool can be (1) the list of miRNA sequence variants, (2) the collapsed expression levels per miRNA locus, and (3) the grouping of miRNAs conserved in different organisms.

4.2 RNA Annotation

The RNA annotation tool is used to generate the secondary structure of an RNA sequence given as input. It uses the RNAfold and RNAplot from the Vienna suite [47]. The tool receives two types of inputs: (1) Long RNA sequences for which the secondary structure will be generated and (2) Shorter RNA sequences (substrings of the longer input), which will be highlighted on the secondary structure using either a user defined color scheme or randomly generated colors (see Fig. 9). Multiple sequences in each category are given in a comma-separated format. The inputs can also be given as FASTA files.

The main use of the RNA annotation tool is via direct interaction with the miRCat tool as described in Subheading 4.2. Predicted miRNA hairpins and the corresponding miRNA and miRNA* can be directly forwarded to the RNA annotation tool via the context menu of miRCat.

4.3 *ta-siRNA* Prediction Tool

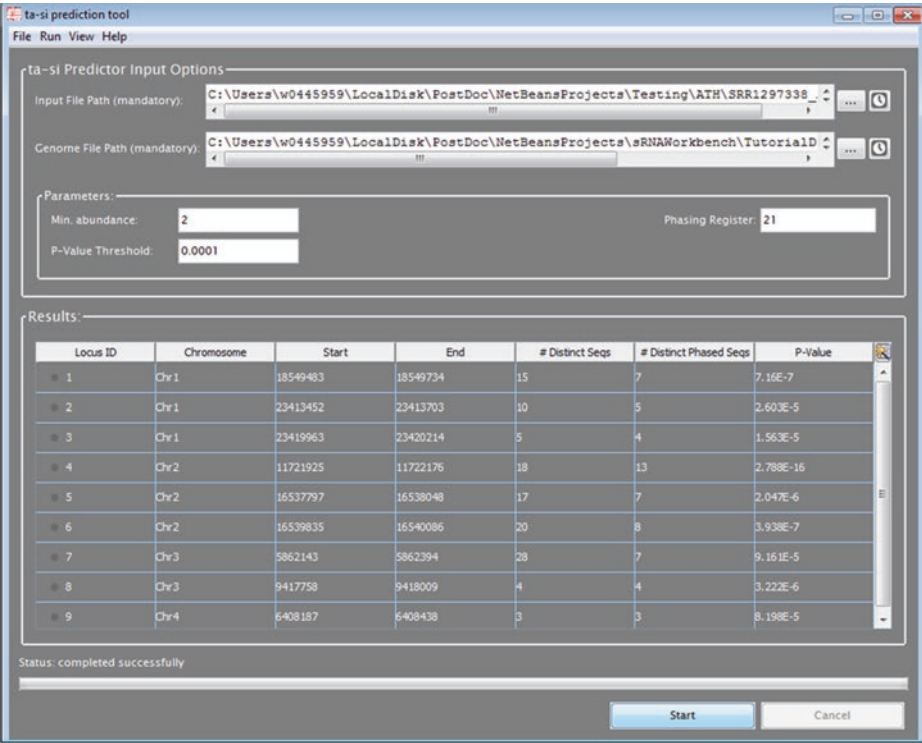
Trans-acting siRNAs (ta-siRNAs) are a variety of plant specific small interfering RNA (siRNA) with similar mode of action to miRNAs, i.e., regulation through posttranscriptional gene silencing [50, 51]. Their characteristic is the phased manner, on a 21 nt register, in which they are produced [52]. Other components on their biogenesis pathway used for validation are the suppressor of gene silencing 3 (SGS3), and the RNA-dependant RNA polymerase 6 (RDR6) [53].

In the Workbench, the current implementation of the ta-siRNA detection tool is based the algorithm described in Chen et al. [54]. The input of this tool consists of a sample, in nonredundant format, and a reference genome.

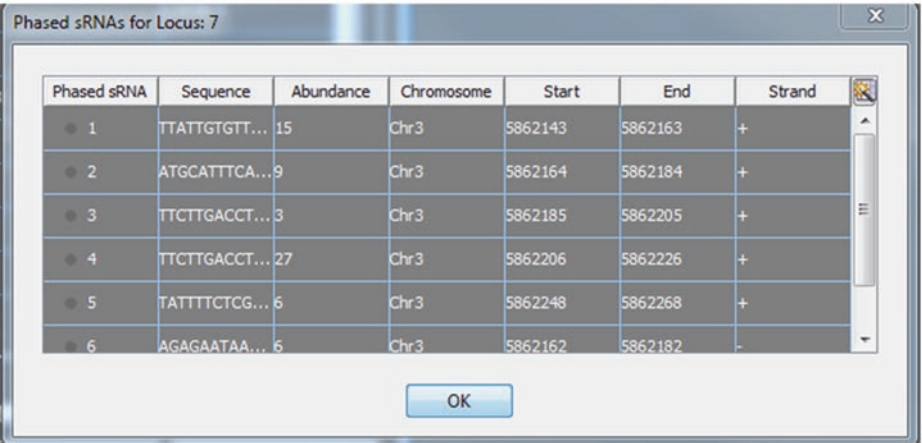
The first step is the alignment of sRNAs to the genome, which is done full length, with no mismatches or gaps allowed. Next, windows of 231 nt are considered, which can accommodate up to 11 phased ta-siRNA duplexes. For the phasing calculation, our implementation differs slightly from the original algorithm as we take into account the length of the sRNAs, only using 21 nt reads for the phasing analysis. Next, the number of reads corresponding to each phasing register is calculated. To determine the probability of the phasing being significant this distribution is compared to a hypergeometric distribution.

The results of this tool are available in the scrollable table presented underneath the input boxes (as shown in Fig. 10 (a)). Each predicted TAS locus is displayed as a row in the table. The columns represent: (1) the chromosome from which the predicted TAS locus originated; (2) the start coordinate of the predicted TAS locus on the chromosome; (3) the end coordinate of the predicted TAS locus; (4) the number of distinct (unique) sRNAs that align to the TAS locus; (5) the number of distinct phased sRNAs that were detected in TAS locus; (6) the P-Value calculated using the Chen algorithm described in [30]. The phased ta-siRNAs for each TAS locus can be displayed in a separate popup dialog by clicking the “Show phased sRNAs” in the context menu for that row in the table, an example is given in Fig. 10 (b).

All TAS loci can also be visualized in VisSR using the “Show in VisSR” menu item located in the View menu. Alternatively, each locus can be sent individually to VisSR by bringing up a context menu on a particular locus and selecting the “Show in VisSR” button. An example of a TAS locus in VisSR is shown in Fig. 11.



(a)



(b)

Fig. 10 Output of the TAS locus prediction tool (a) The TAS prediction results for the analysis on the *Arabidopsis thaliana* dataset; the predicted TAS loci, with information such as the number of predicted phased ta-siRNAs and the corresponding *p*-value are presented. Additional information on a particular locus is obtained by clicking on a row. (b) For the selected TAS locus, the abundance and genomic location of the ta-siRNAs are presented in tabular format

The results produced using the TAS prediction tool can be exported in two types of files (1) a results table detailing all predicted TAS loci, in .csv format, and (2) a list of phased ta-siRNAs for each TAS loci, in .txt format.

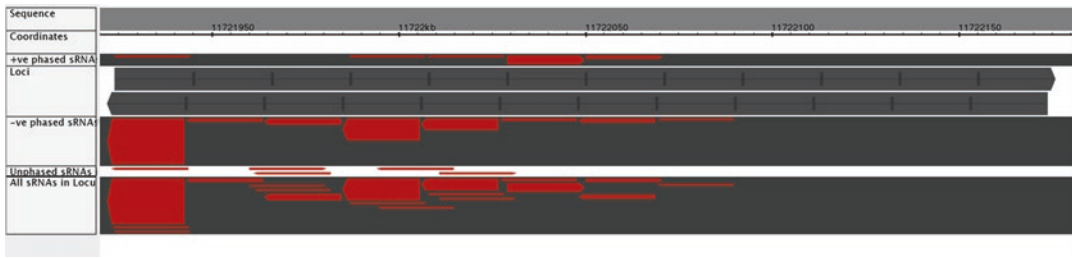


Fig. 11 A visualization of a TAS locus found on the *Arabidopsis thaliana* dataset using the VisSR tool. Each red arrow represents a sequence, the direction of the arrow indicates the strand. The width of the arrow is proportional to its abundance (on logarithmic scale). In the bottom three panels we show (i) the localization of the phased ta-siRNAs, (ii) the localization of the un-phased ta-siRNAs, and (iii) the localization of all incident reads with the TAS locus

4.4 General Locus Prediction

Classes of sRNAs like the miRNAs or ta-siRNA are characterized by a set of distinct and relatively well-understood properties generally derived from their biogenesis. These properties can then be used to determine the region of the genome from which they originated (referred to as a **locus**). For example, for miRNAs, the locus boundaries can be identified using the location of the mature and star sequences [16, 55, 56]. A TAS locus lies between the first and last ta-siRNA which are in phase and the “quality” of the locus can be assessed using the characteristic phased pattern [54]. However, the loci generating other siRNAs, such as heterochromatin siRNAs and piwi siRNAs are not well defined. For the identification of general loci the Workbench proposes two methods SiLoCo [16] and CoLIde [18].

4.4.1 SiLoCo

The **Short interfering Locus Comparison** (SiLoCo) algorithm is a rule-based approach first described in [16, 57]. The user-defined parameters of this method are the number of distinct reads incident to a locus and the maximum allowed distance between consecutive reads allocated to the same locus. The input for this tool is a set of sample file and a reference genome. The output is a table presenting the predicted loci with the normalized abundances in the list of samples. The normalization of the sRNA abundances is done using the RPM method (*see* Subheading 3.4.1), a weighting procedure on the number of genomic hits can also be applied; the expression level of a locus within a sample is the algebraic sum of the abundances of all incident reads. The output from SiLoCo can be piped through the VisSR tool detailed in Subheading 4.4.3, and example of the visualization of a predicted locus can be found in Fig. 7 (a).

The SiLoCo algorithm performs well on small datasets, i.e., datasets with a sequencing depth of less than 3–4 M reads. As the sequencing depth increases the accuracy of the method declines. This is due in part to spurious, random reads which artificially elongate the predicted loci. In extremis, for a high and consistent

coverage of the genome a single locus will be predicted on a chromosome. To alleviate this problem we developed a tool, CoLIde, which uses both genomic location and sRNA expression information to predict loci.

4.4.2 CoLIde

Co-expression based Locus Identification (CoLIde) was introduced as part of the major Version 3 update [18]. It improved the prediction of sRNA loci by using patterns of expression and genomic location. Similar to SiLoCo, the input consists of a set of sample files (in nonredundant format, with normalized expression levels) and a reference genome. In addition, CoLIde is designed to process datasets with or without replicates.

The CoLIde algorithm consists of four stages (a) calculation of confidence intervals, (b) generation of patterns, (c) detection of loci using significance tests on abundance and size class and, as an option, (d) visualization. Next, each stage will be described in detail.

(a) *Calculation of confidence intervals*: For each genome matching sRNA, in each sample, the confidence intervals are constructed using the normalized expression levels. The five equations that can be used to calculate the upper and lower limits of the confidence intervals are presented below. Eqs. 1–4, for determining the Chebyshev intervals, can be used when replicates are available for each sample, whereas if no replicates are present a user-defined percentage is used for determining the upper and lower bound. The former use the mean μ and the standard deviation σ to create intervals capturing 100%, 94%, 67%, and 50% of the replicated measurements

$$CI_{ij} = \left[\min(x_{ijk})_{k=1,r}, \max(x_{ijk})_{k=1,r} \right] \quad (1)$$

$$CI_{ij} = [\mu_{ij} - 2\sigma_{ij}, \mu_{ij} + 2\sigma_{ij}] \quad (2)$$

$$CI_{ij} = [\mu_{ij} - \sigma_{ij}, \mu_{ij} + \sigma_{ij}] \quad (3)$$

$$CI_{ij} = \left[\mu_{ij} - \frac{1}{\sqrt{2}}\sigma_{ij}, \mu_{ij} + \frac{1}{\sqrt{2}}\sigma_{ij} \right] \quad (4)$$

(b) *Generation of patterns and pattern intervals*: Patterns are calculated using the relative location of the confidence intervals (e.g., in an ordered series, the first sample is compared to the second, generating the first letter of the pattern, next the second sample is compared to the third generating the second letter, etc.). The patterns can be {U} up if the right hand side interval's range is completely above the left, {D} down if completely below or [1] straight if one interval overlaps with the other CI (see Fig. 12)

The overlapping (non-empty intersection) reads that share the same pattern are grouped into an initial pattern interval. Then the relative localizations of all pattern intervals are used to calculate

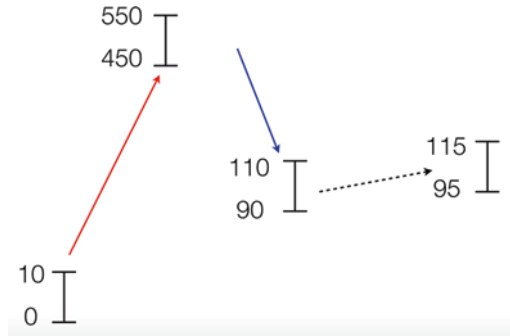


Fig. 12 An example of pattern created on confidence intervals and the use of the {U,D,S} pattern system. The *red arrow* shows an {U} up, *blue* indicates {D} down and the dotted arrow depicts an S straight

the distribution of distances between any two consecutive pattern intervals. Finally, pattern intervals are merged if the distance between them falls below the median of the previously described distribution; these regions are then used as putative loci for further examination.

(c) *Detection of loci using significance tests:* A putative locus is accepted if the overall abundance (sum of expression levels of all incident sRNAs, in all samples) is significant (in a standardized distribution) when compared to the abundances of incident putative loci in its proximity. The abundance significance test is conducted by considering the incident loci with the flanking regions of the locus (500 nt upstream and downstream, respectively).

The accuracy of loci prediction is further improved with the addition of a significance test on the size class distribution of incident reads. Loci derived from a particular biogenesis are characterized by a specific size class (e.g., miRNA loci are enriched in 21mers in plants and 22mers in animals, heterochromatin loci are enriched in 24mers, in plants) whereas random degradation is characterized by a quasi-uniform distribution of read lengths. To check whether the size class distribution of a predicted locus is similar or not to a random distribution, a (need chi symbol here?) X^2 test is used on the most abundant four size classes for each locus (these are determined automatically, but can be modified by the users). To reduce the effect of spurious low-abundance hits an offset version of the X^2 test is used (the default version of the offset is 10), e.g., a locus with four incident reads, the first with abundance 7 and length 21 nt, the second, third, and fourth with abundances 1 and lengths 22, 23, and 24 nt, respectively, would produce a X^2 p -value of 0.012, whereas the offset version of the same test would produce a p -value of 0.53 rendering the locus not significantly enriched in a particular size class.

(d) *Visualization:* In addition to the improved prediction introduced with CoLide [18, 58], a new visualization approach

was added to the visualization tool, VisSR, to allow a clearer representation of dense loci; a complete description is presented in the next Subheading 4.4.3.

The output of CoLIde consists of a table of loci presenting, for each locus: a unique ID indicating the chromosome/scaffold, the start and end coordinates, the locus length, p-value value resulting from the X^2 statistic, the expression series and the absolute differential expression for each locus.

4.4.3 Visualization of Loci—VisSR

The visualization of small RNAs (VisSR tool) was designed as a genome browser based on the GenoViz library [59]. It is used in the Workbench for the visualization of sRNAs aligned to the genome and the production of publication ready images from various data analysis tools, e.g., predictions made using miRCat can be forwarded to VisSR and based on the resulting visual patterns, further characterized.

In VisSR each sRNA is represented as an arrow indicating the position and strand; the thickness of each arrow is proportional with the (normalized) abundance of the sequence, in logarithmic scale; the different colors are used to separate the various size classes and hovering the cursor over an individual read will display a tooltip with all available information such as ID, position, size class, and strand.

Each individual sample loaded into VisSR is displayed in a separate tier, this could include the software that activated the tool, a set of samples or a GFF file that has been included (*see* Fig. 13).

In addition to the standard representation of reads, VisSR also includes an aggregate mode for visualizing loci predicted using the CoLIde or SiLoCo programs proposed as an alternative suitable for loci with numerous incident reads [58].

The genome is first partitioned into windows of fixed length W . sRNAs incident to junction regions are assigned to the window that contains more than 50% of their sequence. The size class distribution of the top four most abundant size classes is then represented as a histogram. The overall abundance of the locus is proportional, on logarithmic scale, with the width of the rectangle corresponding to each window. The strand bias (calculated as described in Subheading 3.5.1, and varying between 0 and 1) is represented as line spanning the length of the window [18, 58].

Figure 14 demonstrates the advantages of the aggregate view for loci with high number of incident reads (the example is based on a 10 time point tomato experiment [30]) by presenting a region using both the traditional approach and the aggregate form.

4.5 Degradome Assisted Target Prediction by PAREsnip

A key step to understanding the function of sRNAs is to identify their messenger RNA (mRNA) targets. In plants, and to a lesser extent in animals [10], small RNAs can silence their mRNA targets through endonucleolytic cleavage [11], resulting in the release of

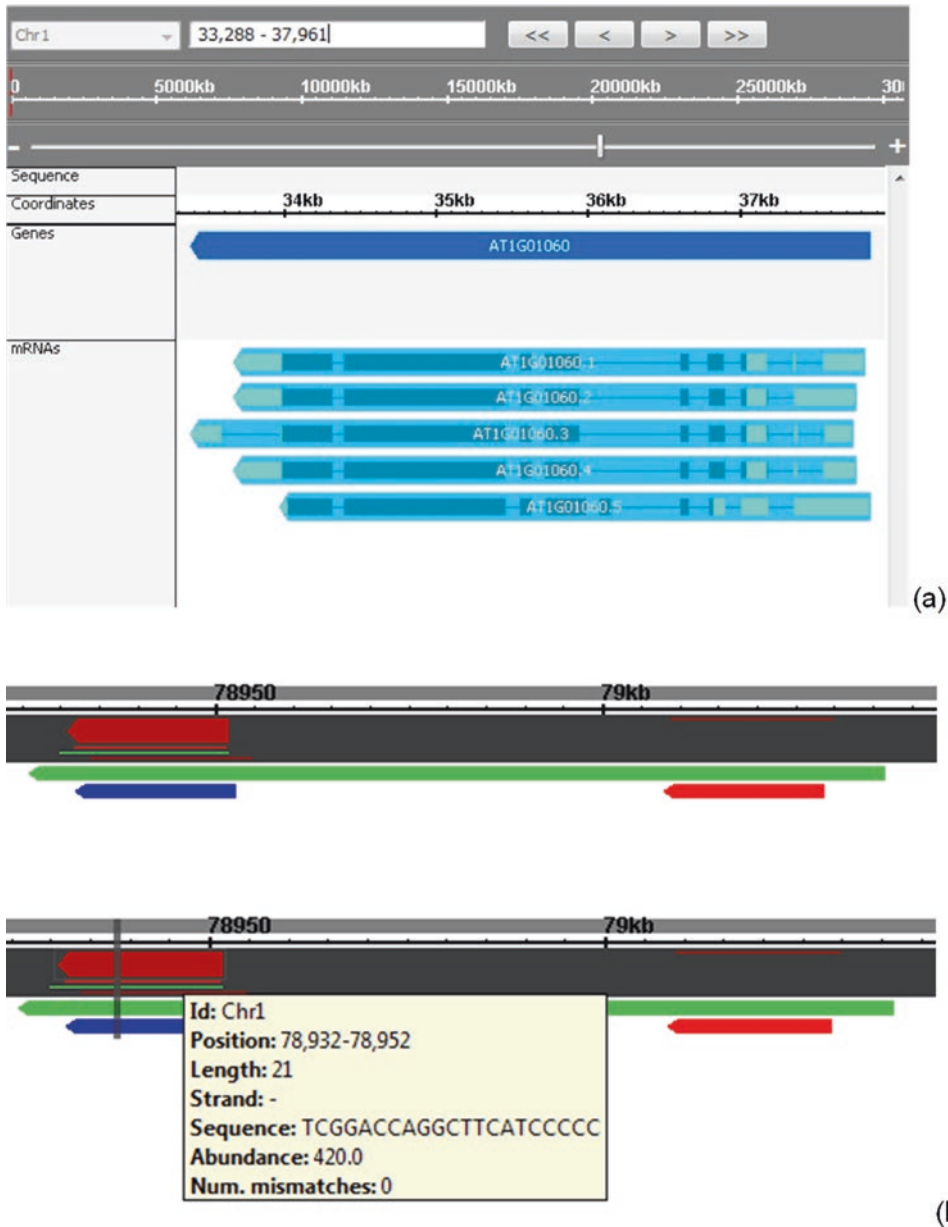


Fig. 13 (a) A VisSR track with annotated miRNA precursor sequences found from a GFF file, placed into a separate track. (b) A predicted miRNA locus found using the miRCat tool and sent directly to the VisSR tool for visual inspection. Hovering over individual reads will give all the available information as a tooltip

millions of mRNA fragments. Such fragments can be captured on a genome wide scale through techniques such as Parallel Analysis of RNA Ends (PARE) [60, 61] using high-throughput sequencing technologies. The resultant degradation profile obtained from a PARE experiment, also known as the degradome, can be used to identify signals of sRNA-mediated cleavage [22, 62, 63].

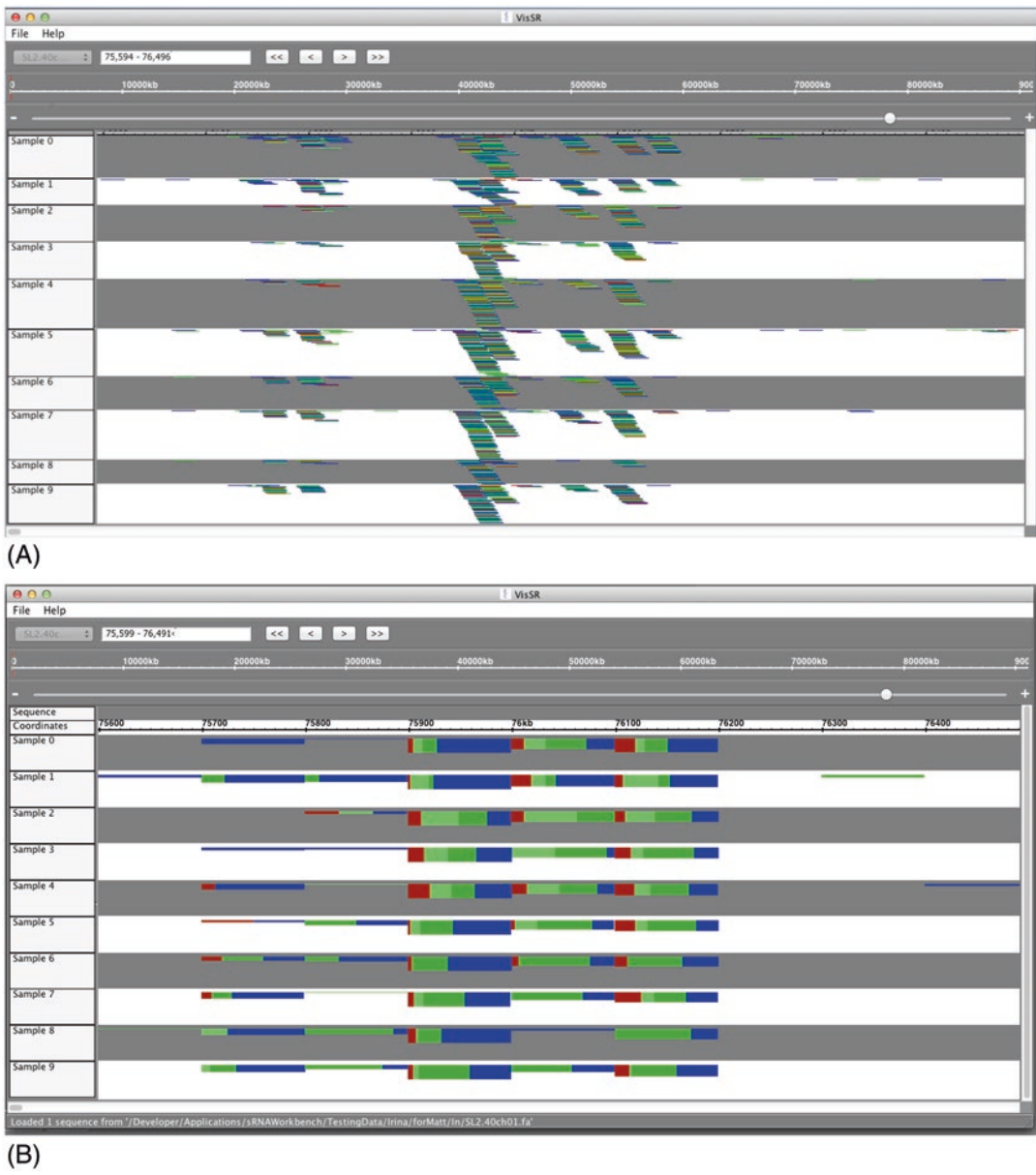


Fig. 14 (a) A standard arrow approach to visualizing aligned sRNAs on a genome (b) the aggregated equivalent view

Computational tools now exist for parsing PARE (degradome) output in order to identify potential signals of cleavage and use this information to assist in predicting sRNA targets. One such tool within the UEA small RNA Workbench is PAREsnip [17].

The PAREsnip tool performs a rapid genome-wide search for binary sRNA/mRNA interactions between all input sRNAs and all input transcripts, supported by the presence of transcript degradation fragments. For a specific organism, PAREsnip requires

the following inputs in redundant FASTA file format; a mRNA dataset (transcriptome), a set of transcript degradation fragments (degradome), and a sRNA dataset (sRNAome). Input sequences must only contain the characters, “A”, “C”, “G”, “T”/“U”, otherwise they are discarded. sRNA sequences are filtered by properties such as length, abundance, and sequence complexity, according to the user-defined parameters; reads that match known tRNAs or rRNAs are discarded. A genome file may also be specified, which when provided, is used to remove sRNA sequences that do not exactly align to the genome.

To identify sRNA/target interactions, the transcript degradation fragments are mapped against the set of input transcripts (with exact matching) to detect potential sRNA cleavage sites. Each site is assigned a category based on either the raw or weighted degradation fragment abundance (on the number of genomic hits, as specified by the user) at the corresponding position on the transcript using the 5-category system defined in [63]. The category system attempts to summarize the data by describing the strength of a potential cleavage signal in relation to other potential cleavage sites within a single transcript. The strongest cleavage signals are classified as Category 0, 1, or 2. Category 3 represents a weak signal and Category 4 classifications generally represent background noise or random degradation.

For each identified cleavage site, the reverse complement of its corresponding transcript, 11-nt downstream and 15-nt upstream from the cleavage site, is compared to the set of sRNAs to find sequences that could potentially target the site. To remove sequence and subsequence redundancy, and to reduce the search-space and memory footprint of the tool, sRNA sequences are encoded into a partitioned 4-way tree where the edges of the tree represent nucleotide bases and the nodes represent traversal paths. mRNA cleavage typically occurs between positions 10 and 11 of a sRNA, and the nucleotides at these positions should be complementary to the corresponding nucleotides on the transcript. Therefore, the tree is traversed from node/position 10. Entry nodes are organized and assigned to one of 16 bins based on the possible dinucleotide permutations at positions 10 and 11 (“AA”, “AC”, “AG”, ..., “TG”, “TT”). When searching for a sRNA that could potentially cleave the transcript, the two nucleotides either side of the cleavage site on the transcript are used to obtain a set of entry nodes from its corresponding node bin. The tree is partitioned such that any path with an entry node in one of the other 15 bins is removed from consideration. For each start node, the tree is then traversed to obtain sRNAs with valid alignments. During the traversal of the tree, previously described rules (Allen et al. 2005; Schwab et al. 2005) accounting for the allowable features within an sRNA/mRNA duplex are enforced. The allowable features include the maximum number of mismatches and gaps within the duplex and their associated positions. If any of the user-configurable rules are

not fulfilled during the tree traversal, the current search is aborted. However, if no rules were broken, a valid alignment was found.

To determine whether the identified valid alignments could have occurred by chance, a p -value is calculated for each alignment as described in [63]. To compute this value, a 26-nt sequence that represents the sRNA binding site is extracted for every position on each transcript with a cleavage signal (11-nt downstream and 15-nt upstream from the cleavage site). This sequence is then assigned to one of five category trees based on its cleavage signal classification (similar to the four-way tree used to encode sRNAs). The sRNA involved in each identified alignment is randomly shuffled (preserving dinucleotide frequency) and then mapped to all the target sites in the corresponding category tree. The output of this mapping procedure has a binary result representing alignment success. This process is repeated a number of times (100 by default), and the p -value represents the number of successes as a fraction of the total number of shuffles. Therefore, a small p -value indicates high confidence that the alignment did not occur by chance.

PAResnip outputs its results in a tabular format (Fig. 15). Each interaction reported includes properties such as signal category, alignment score, sRNA/target duplex, p -value, and sequence abundance. The tool can be used in GUI mode or using the command line. When using the GUI mode of the sRNA Workbench, the user can export the data in CSV format and visualize target plots using the VisSR tool (Fig. 16). For a more comprehensive description of the PAResnip tool, the reader is referred to [17].

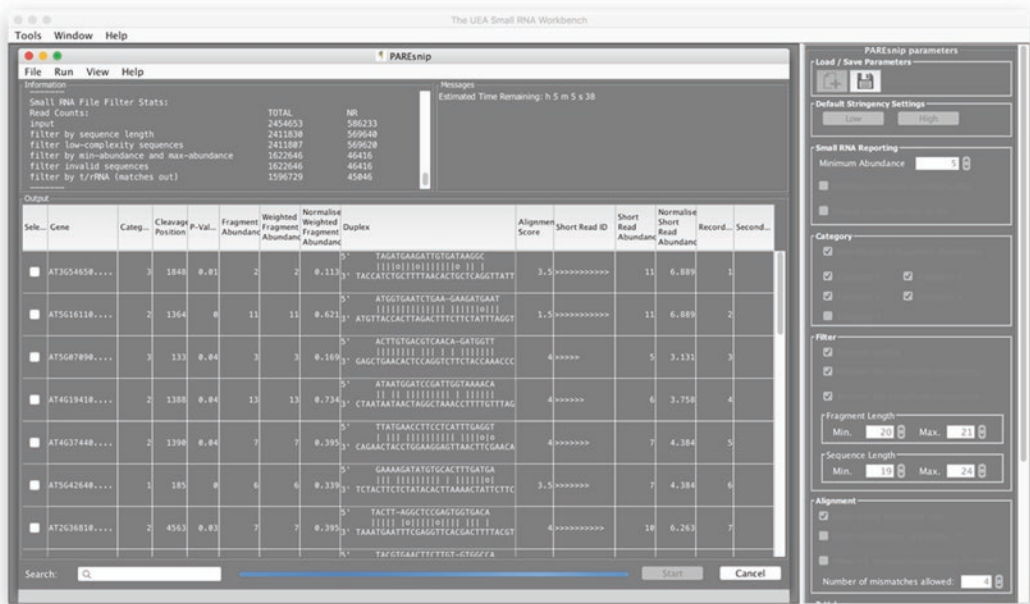


Fig. 15 Screenshot showing an example of the PAResnip tool's output within its graphical user interface while processing *Arabidopsis thaliana* data using high-stringency settings

Category: 0 1 2 3 4
Degradome alignment: Median: —

0 #1 Position:179 Abundance: 227.00(deg) 7(sRNA)
5' GCTCCGAGGATG-TGTTTGA 3' ID:
| |o||o|o||| ||||o|| Score: 4.0
3' GAAGCAAGGGTTTCTACCACAAGCTAAGCTCC 5' p-value: 0.03

3 #2 Position:366 Abundance: 6.00(deg) 10(sRNA)
5' CTTTGTCTACAATTTTGG 3' ID:
|||||o|||o|||o||| Score: 2.0
3' CGAGAAACAAATGTCAAAACCTCAGCATAAC 5' p-value: 0.05

POSITION Transcript length: 1074

Fig. 16 An example of a t-plot generated by VisSR showing degradome supported interactions for transcript AT5G24780.1 when analysing *Arabidopsis thaliana* data using PAREsnip

Acknowledgments

We thank Matthew Beckers, Tamas Dalmay, Frank Schwach, Simon Moxon, Hugh Woollfenden, Helio Pais, and Daniel Mapleson for their contributions to the UEA sRNA Workbench.

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811
2. Ruiz MT, Voinnet O, Baulcombe DC (1998) Initiation and maintenance of virus-induced gene silencing. Plant Cell 10:937–946
3. Meister G, Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. Nature 431:343–349
4. Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15:509–524

5. Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol* 57:19–53
6. Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* 121:207–221
7. Talbert PB, Henikoff S (2006) Spreading of silent chromatin: inaction at a distance. *Nat Rev Genet* 7:793–803
8. Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12:246–258
9. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC (2007) A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* 315:1587–1590
10. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
11. Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell* 136:669–687
12. Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 12:87–98
13. Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92:255–264
14. Beckers M, Mohorianu I, Stocks MB, Applegate C, Dalmay T, Moulton V (2017) An interactive pipeline for quality checking, normalization and differential expression analysis of high throughput small RNA sequencing data. *in preparation*.
15. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 28:2059–2061
16. Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, Moulton V (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 24:2252–2253
17. Folkes L, Moxon S, Woolfenden HC, Stocks MB, Szitty G, Dalmay T, Moulton V (2012) PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic Acids Res* 40:e103
18. Mohorianu I, Stocks MB, Wood J, Dalmay T, Moulton V (2013) CoLId: a bioinformatics tool for CO-expression-based small RNA Loci Identification using high-throughput sequencing data. *RNA Biol* 10:1221–1230
19. Zhang X, Zhu Y, Liu X, Hong X, Xu Y, Zhu P, Shen Y, Wu H, Ji Y, Wen X et al (2015) Plant biology . Suppression of endogenous gene silencing by bidirectional cytoplasmic RNA decay in Arabidopsis. *Science* 348:120–123
20. Camps C, Saini HK, Mole DR, Choudhry H, Reczko M, Guerra-Assuncao JA, Tian YM, Buffa FM, Harris AL, Hatzigeorgiou AG et al (2014) Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia. *Mol Cancer* 13:28
21. Montgomery TA, Yoo SJ, Fahlgren N, Gilbert SD, Howell MD, Sullivan CM, Alexander A, Nguyen G, Allen E, Ahn JH, Carrington JC (2008) AGO1-miR173 complex initiates phased siRNA formation in plants. *Proc Natl Acad Sci U S A* 105:20055–20062
22. German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K, German R et al (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* 26:941–946
23. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* 53:474–485
24. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40:D1202–D1210
25. Prufer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* 24:1530–1531
26. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
27. McCormick KP, Willmann MR, Meyers BC (2011) Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence* 2:2
28. Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* 3:4
29. Xu P, Billmeier M, Mohorianu I, Green D, Fraser WD, Dalmay T (2015) An improved

- protocol for small RNA library construction using High Definition adapters. *Methods in Next Generation Sequencing* 2:2084–7173
30. Mohorianu I, Schwach F, Jing R, Lopez-Gomollon S, Moxon S, Szittya G, Sorefan K, Moulton V, Dalmay T (2011) Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *Plant J* 67:232–246
 31. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J (2000) Comparing methods of clinical measurement: reporting standards for bland and altman analysis. *Anesth Analg* 90:593–602
 32. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J et al (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14:671–683
 33. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 14:R95
 34. Sonesson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91
 35. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628
 36. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11:R25
 37. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94
 38. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106
 39. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550
 40. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193
 41. Mohorianu I, Bretman A, Smith D, Fowler E, Dalmay T, Chapman T (2016) New approaches for analysing RNA-seq data: sampling-based normalization and hierarchical differential expression. *in preparation*.
 42. Li J, Witten DM, Johnstone IM, Tibshirani R (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13:523–538
 43. Cleveland W (1979) Robust locally weighted regression and smoothing scatterplot. *J Am Stat Assoc* 74:829–836
 44. Cleveland W (1981) LOWESS: a program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* 35:54–60
 45. Lopez-Gomollon S, Mohorianu I, Szittya G, Moulton V, Dalmay T (2012) Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions. *Planta* 236:1875–1887
 46. Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 42:D68–D73
 47. Hofacker IL, Lorenz R (2014) Predicting RNA structure: advances and limitations. *Methods Mol Biol* 1086:1–19
 48. Lorenz R, Hofacker IL, Stadler PF (2016) RNA folding with hard and soft constraints. *Algorithms Mol Biol* 11:8
 49. Bonnet E, Wuyts J, Rouze P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* 20:2911–2917
 50. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev* 18:2368–2379
 51. Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gascioli V, Mallory AC, Hilbert JL, Bartel DP, Crete P (2004) Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol Cell* 16:69–79
 52. Fei Q, Xia R, Meyers BC (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell* 25:2400–2415
 53. Yifhar T, Pekker I, Peled D, Friedlander G, Pistunov A, Sabban M, Wachsmann G, Alvarez JP, Amsellem Z, Eshed Y (2012) Failure of the tomato trans-acting short interfering RNA program to regulate AUXIN RESPONSE FACTOR3 and ARF4 underlies the wiry leaf syndrome. *Plant Cell* 24:3575–3589
 54. Chen HM, Li YH, Wu SH (2007) Bioinformatic prediction and experimental validation of a

- microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc Natl Acad Sci U S A* 104:3318–3323
55. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26:407–415
56. Guerra-Assuncao JA, Enright AJ (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* 11:133
57. Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC (2007) miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* 447:1126–1129
58. Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641
59. Helt GA, Nicol JW, Erwin E, Blossom E, Blanchard SG Jr, Chervitz SA, Harmon C, Loraine AE (2009) Genoviz Software Development Kit: Java tool kit for building genomics visualization applications. *BMC Bioinformatics* 10:266
60. German MA, Luo S, Schroth G, Meyers BC, Green PJ (2009) Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc* 4:356–362
61. Zhai J, Arikait S, Simon SA, Kingham BF, Meyers BC (2014) Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. *Methods* 67:84–90
62. Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ (2008) Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol* 18:758–762
63. Addo-Quaye C, Miller W, Axtell MJ (2009) CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25:130–131

Prediction of miRNA–mRNA Interactions Using miRGate

Eduardo Andrés-León, Gonzalo Gómez-López, and David G. Pisano

Abstract

miRGate (<http://mirgate.bioinfo.cnio.es/>) is a freely available database that contains predicted and experimentally validated microRNA–messenger RNA (miRNA–mRNA) target pairs. This resource includes novel predictions from five well-established algorithms, but recalculated from a common and comprehensive sequence dataset. It includes all 3′-UTR sequences of all known genes of the three more widely employed genomes (human, mouse, and rat), and all annotated miRNA sequences from those genomes. Besides, it also contains predictions for all genes in human targeted by miRNA viruses such as Epstein-Barr and Kaposi sarcoma-associated herpes virus.

The approach intends to circumvent one of the main drawbacks in this area, as diverse sequences and gene database versions cause poor overlap among different target prediction methods even with experimentally confirmed targets. As a result, miRGate predictions have been successfully validated using functional assays in several laboratories.

This chapter describes how a user can access target information via miRGate’s web interface. It also shows how automatically access the database through the programmatic interface based on representational state transfer services (REST), using the application programming interface (API) available at <http://mirgate.bioinfo.cnio.es/API>.

Key words MicroRNA, Prediction, 3′-UTR, Validated, API, Target site

1 Introduction

MicroRNAs (miRNAs) are an abundant class of small noncoding RNA molecules of 21–25 nucleotides (nt) that regulate gene expression [1]. miRNAs constrain gene expression via translational repression, mRNA degradation, or transcriptional inhibition [2, 3]. Other mechanisms such as elongation inhibition or ribosome drop-off (premature termination) have also been reported [3].

miRNAs mechanism of regulation is driven by a Watson–Crick complementary binding of their seed sequence (an evolutionary-conserved region of 5–7 nucleotides at the 5′ end of the microRNA) to a sequence in the 3′ UTR of its targeted mRNA [4]. However, other studies confirmed the binding of miRNAs to coding or 5′ UTR regions of known genes [5–7]. Thus, a single miRNA is

putatively able to regulate hundreds of target mRNAs; in fact, it is thought that at least 90% of genes could be subjected to miRNAs regulation [8, 9]. To verify this statement, different experimental approaches has been developed such as SILAC [10], miRNA microarrays [8] and next-generation sequencing (NGS) techniques such as miRNASeq [11]. Current experimental methodologies are complex, expensive and have to deal with tissue specificity, low expression, 3' UTR selection, and miRNA stabilization [12], among others. As a complementary approach, computational target predictors offer an alternative procedure to identify putative miRNA–mRNA targets based on previously described factors, such as: (1) complementarity of the seed sequence with its associated target; [4] (2) evolutionary conservation of the seed sequence (for example, 70% of experimentally validated human targets are conserved, even in nonmammalian species [13]); (3) thermodynamic stability, where favorable energy is determined among the bound and unbound RNA double strand [14]; (4) miRNA cooperation, when an effect in regulation is only detected if several miRNAs are simultaneously bound to the same mRNA [15].

There are many algorithms for target prediction based on the combination of previously commented factors. Among the most used methods, we highlight Targetscan [16], miRanda [17], Pictar [18], PITA [19], RNA22 [20], RNAhybrid [21], TarBase [22], miRTarBase [23], MMIA [24], and MAGIA2 [25]. These tools have been designed to provide putative and/or experimentally confirmed miRNA–target interactions pairs. Many of them highlight the two most important handicaps in the prediction results: the absence of overlap among the different target predictions given by distinct methods, and the low specificity of the predictions once further experimental validation is performed [26].

One possible strategy to reduce those problems rely on integrative approaches based on targets predicted by two or more algorithms. Nevertheless, the abovementioned tools are based on different reference databases and distinct genomics builds (*see* Tables 1 and 2), while the predictions and the overlap among them is highly dependent on the actual sequences used.

Table 1
Comparison of 3' UTRs used sequences among datasets

Algorithm	Genomic build	Coding Genes	Noncoding-genes	Pseudogenes	3' UTR
miRanda	NCBI37 2009	19,778	–	–	34,592
TargetScan	NCBI37 2009	18,414	–	–	30,932
Pita	NCBI36 2006	18,582	–	–	24,086
miRGate	NCBI37 2009	20,805	22,966	14,181	196,501

Table 2
Number of mature miRNAs sequences included in the different datasets

Algorithm	Human miRNAs	Mouse miRNAs	Rat miRNAs	miRBase version
miRanda	1100	717	387	miRBase 15
TargetScan	1433	722	–	miRBase 17
Pita	692	500	–	miRBase 11
PicTar	81	81	81	Rfam 5
miRGate	2680	1983	763	miRBase 20

When tested systematically, the overlap among predictions calculated with the same algorithm but using a different reference dataset, barely reaches 50% [26].

To eliminate these restraints, we designed miRGate to fill the gap between predicted and non-concordant experimentally validated targets. Among miRGate properties, we would like to highlight the following features:

- Common reference dataset: miRGate is build upon a consistent dataset of updated and complete sequences for miRNAs based on miRBase 20 [27] and all 3' UTR sequences for human, mouse, and rat from Ensembl 74 [28].
- Our database incorporates all existing isoforms for each known gene, since each variant could have an exclusive 3' UTR. This comprises all biotypes manually annotated by the Human and Vertebrate Analysis and Annotation group (HAVANA) [29], such as protein coding isoforms, long interference noncoding RNAs (lincRNAs), pseudogenes [30], mitochondrial RNAs and antisense variants, among others. Different gene variants (thus different 3' UTRs) are often expressed in certain cell types and/or development phases, so a distinct set of miRNAs regulate those genes. By including predictions from all existing isoforms, miRGate is able to produce specific miRNA–mRNA predictions (instead of gene–miRNA).
- In order to offer high reliable targets and improve specificity, our methodology implements “genomic agreement” functionality. This integrative approach highlights predictions made by different algorithms occurring in the same 3' UTR genomic position (*see Note 4*).
- In addition, a z-score over all individual prediction is included to rank them. This score is computed by standardizing individual raw scores in each prediction by all predictions stored in the database. When a genomic agreement is found among different targets, the results are combined producing a consensus weighted score as it has been earlier proposed in other areas [31].

- miRGate also collects experimentally verified interactions from four independent public databases: (1) Tarbase [22], (2) miR-Tarbase [23], (3) miRecords [32], and (4) OncomirDB [33].

The integration of genomic agreement, *z*-score ranking, and validated targets increases the overlap among predictions [34], and provides high quality results in independent experimental setups [35, 36].

Finally, due to the potentially massive amount of information that can be simultaneously queried when interrogating full expression datasets, we offer an XML-based REST-API to allow massive queries to be performed directly to the database.

2 Materials

miRGate usage needs at least the following elements:

1. A standard computer with any Internet browser software.
2. A list of miRNAs and/or genes of interest (*see Note 1*).
3. Optionally, for programmatic access only, a programming/scripting language and its corresponding XML library is required. We will use Perl as a example coding language, using the XML::LibXML [37] library included in the Comprehensive Perl Archive Network (CPAN).

3 Methods

In this section we present the recommended procedure to obtain predicted and experimentally validated miRNA–mRNA interactions using the online web page. Next, we explain how to access automatically to the database over the API functionality. Lastly, we summarize the functions of miRGate, based on recently published projects.

miRGate stores information in relational databases based on MySQL (Structure Query Language) database management system (DBMS). Each organism has its own database, which can be accessed in two ways: via a web-based manner (Subheading 3.1), or using REST web services (Subheading 3.2).

3.1 miRGate Web Usage

The web interface is designed to allow users to obtain potential targets from their gene variants and/or miRNAs of interest (Fig. 1). The web page presents an intuitive step-by-step process where users need to supply the organism of interest, and the gene/miRNA names to retrieve predictions from (*see Note 1*). After accessing miRGate at <http://mirgate.bioinfo.cnio.es>, the protocol is as follows:

1. **Choose Organism:** In the query parameters section, choose an organism in the drop-down list.

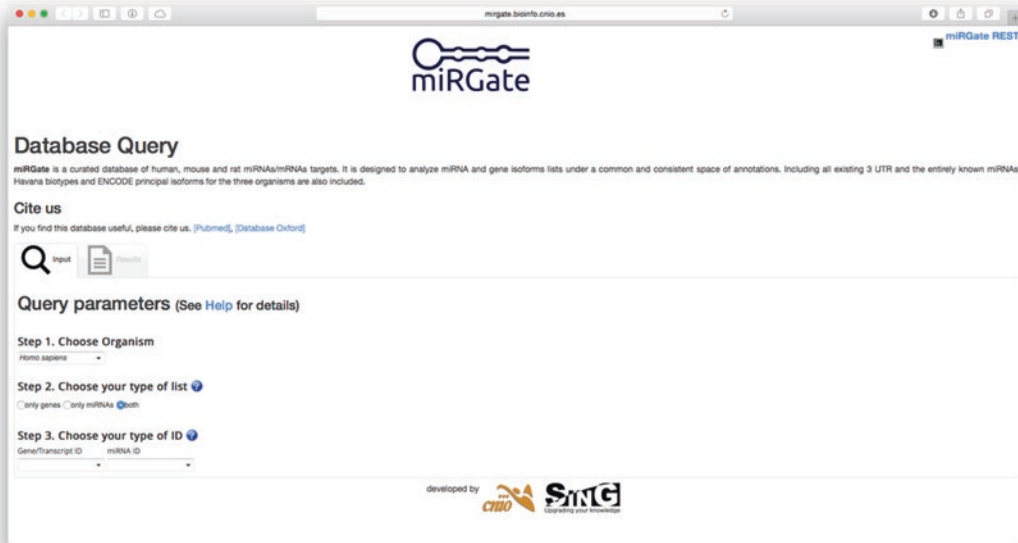


Fig. 1 Snapshot index page

2. **Choose your type of list:** Once the organism has been selected, a second step is available offering three possibilities.
 - (a) Only genes: Use this option to predict miRNAs that could potentially regulate the gene or list of genes included.
 - (b) Only miRNAs. Alternatively, use this option when interested in the genes that could be regulated by specific miRNAs.
 - (c) Both. This option expects two lists (miRNAs and genes) to be specified. This is the preferred option when inverse expression profiles can be supplied (for example, upregulated miRNAs and downregulated mRNAs in the same conditions, obtained from RNASeq or microarrays experiments).
3. **Choose your type of ID:** miRGate accepts different identifiers from different sources and databases. The namespace that best describes the identifiers to be uploaded should be chosen from the drop-down list(s) (*see Note 1*).
4. **Paste your list:** Depending on previous selections, one or two text boxes will appear allowing to upload, write or paste the names of the entities from the list of interest (*see Note 2*). Some entity names of the chosen namespace are shown as examples.
5. **Advanced options (optional):** As an additional step, miRGate offers different filtering choices that allow optional customized queries. Among them, the possibility to filter by

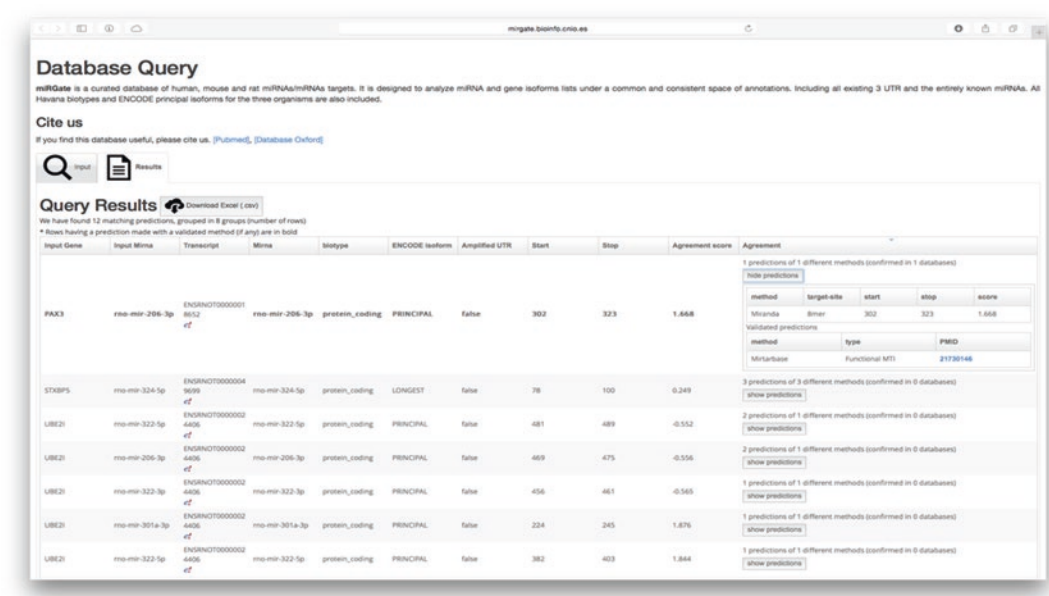


Fig. 2 Result page

ENCODE principal isoforms [38], HAVANA biotypes [29], a minimum genomic agreement value, and/or predicted 3' UTR mRNA sequences (see Note 3) are features exclusive of miRGate. Those filter options are intended to obtain consistent predictions from your favorite prediction methods, as well as experimentally validated interactions (see Note 4).

6. **Query:** Once all options have been filled, the “query” button proceeds to the results page. miRNA–mRNA interactions are presented in tabular form (one interaction per line), and ranked by default using the genomic agreement value, which is calculated among obtained predictions within the start–stop coordinates (Fig. 2). Additional information on biotype, ENCODE isoform and kind of UTR is also shown (see Note 3). Individual predictions for each interaction are shown by pressing the “show predictions” button (see Note 5).
7. **Download Excel (.csv)** allows to download and save the list of interactions in CSV (Comma separated value) format for downstream studies (see Note 6).

3.2 Programmatic Access Through Web Services

Besides the web interface, miRGate includes an XML-based REST-API that allows machine queries to be performed directly against the underlying database. The base URL for the REST API is <http://mirgate.bioinfo.cnio.es/ResT/API/> and it can be tested from a web browser manually constructing the query URLs as shown in Table 3. Full documentation is available at <http://mirgate.bioinfo.cnio.es/API/>.

Table 3
URL examples for API request information

	Query type	API parameters	Example URL
Information	miRNA information	Organism/API method/Identifier	http://mirgate.bioinfo.cnio.es/ResT/API/human/miRNA_info/hsa-miR-329 .
	Gene information		http://mirgate.bioinfo.cnio.es/ResT/API/rat/gene_info/Nfkbib
Predicted targets	miRNA target prediction		http://mirgate.bioinfo.cnio.es/ResT/API/rat/miRNA_predictions/MIMAT0000858
	mRNA target prediction		http://mirgate.bioinfo.cnio.es/ResT/API/rat/gene_predictions/Mapkapk5
	miRNA–mRNA target prediction	Organism/API method/Identifier1/Identifier2/[Method]	http://mirgate.bioinfo.cnio.es/ResT/API/human/miRNA_gene_predictions/hsa-miR-1/TP53
	mRNA–miRNA target prediction		http://mirgate.bioinfo.cnio.es/ResT/API/human/gene_miRNA_predictions/ABCD1/MIMAT0007883/Rnahybrid
Validated targets	miRNA validated predictions	Organism/API method/Identifier	http://mirgate.bioinfo.cnio.es/ResT/API/mouse/miRNA_confirmed/MIMAT0000246
	mRNA validated predictions		http://mirgate.bioinfo.cnio.es/ResT/API/human/gene_confirmed/ABHD11
	miRNA–mRNA validated predictions	Organism/API method/Identifier1/Identifier2/[Method]	http://mirgate.bioinfo.cnio.es/ResT/API/human/miRNA_gene_confirmed/hsa-miR-1/ABHD11
	mRNA–miRNA validated predictions		http://mirgate.bioinfo.cnio.es/ResT/API/mouse/gene_miRNA_confirmed/Gpx7/mmu-miR-122-5p/miRecords

The API lets researchers to include miRGate information in their computational analysis workflows by automatic creation of URLs using any programming language. This API allows:

- **miRNA general information retrieval.** Figure 3a shows detailed XML output obtained from the human miRNA *hsa-miR-329*. It includes miRBase accession names, cleavage information and miRNA sequence among other information.
- **mRNA information recovery.** For instance, Fig. 3b illustrates an XML response containing genomic coordinates, biotypes, various identifiers and 3' UTR sequences of all known variants from rat *NFKB1B* gene name.



Fig. 3 (a) http://mirgate.bioinfo.cnio.es/ResT/API/human/miRNA_info/hsa-miR-329. This XML response contains all genomic information stored in miRGate regarding *hsa-miR-329*. Of interest, we include all miRBase identifiers, Affymetrix probeset names, mature sequence, and cleavage information. (b) http://mirgate.bioinfo.cnio.es/ResT/API/rat/gene_info/Nfkib. This XML response includes all genomic information collected in miRGate regarding *NFKB1B*. Of note, we include all Ensembl identifiers, Affymetrix probeset names, UTR sequence, genomic coordinates, HAVANA biotypes, ENCODE isoforms among others (c) http://mirgate.bioinfo.cnio.es/ResT/API/human/gene_miRNA_predictions/ENST00000218104/MIMAT0007883/Rnahybrid displays a human prediction obtained by the RNAHybrid algorithm calculated with the 3' UTR sequence of the principal ENCODE isoform from the ABCD1 gene and the seed sequence of the hsa-miR-1909-3p miRNA. (d) The following URL: http://mirgate.bioinfo.cnio.es/ResT/API/human/miRNA_gene_confirmed/hsa-miR-1/ABHD11 shows the XML answer from validated databases contained in miRGate. It includes validation methodology name, database name, PubMed Identifiers (if included)

- **miRNA–mRNA target prediction queries.** This includes interaction retrieval for (1) miRNAs where an XML response is returned with all targeted genes stored in the database. (2) mRNAs names, and similarly, all predicted miRNA regulators are retrieved and (3) miRNA/mRNA where only interactions from specified miRNAs and genes are shown. These API methods also allow filtering options, as an example, the Fig. 3c illustrates all RNAHybrid predictions for human gene *ABCD1* and miRNA *hsa-miR-1909-3p*.
- **miRNA–gene validated interaction requests.** miRGate contains links to research articles where the experimental validation of interactions was performed from our miRNAs and/or genes of interest. An example for human *ABHD11* and *hsa-miR-1* is shown in Fig. 3d.

Table 3 shows various sample URLs for general information retrieval of miRNAs and mRNAs. It also includes examples for targets predictions recovery from miRNAs, mRNAs, or both. Finally, several examples are shown to retrieve experimentally validated interactions for miRNAs and/or genes of interest.

3.2.1 Massive Web Service Accession

In previous examples, miRNA data, gene information, predicted and validated targets for just a couple of miRNAs/genes, was explained. But current genomic technologies create large lists of entities of interest; as a result a method to access information for hundred or thousand of entities is needed. In Fig. 4, a script coded

```
#!/usr/bin/perl
$|=1;
use strict;
use XML::LibXML;

my $SourceListingURL = "http://mirgate.bioinfo.cnio.es/ResT/API";

my $parser = XML::LibXML->new();
my $xmldoc = $parser->parse_file($SourceListingURL);

##### Example : retrieve predictions about my genes and miRNAs of interest #####

my $parser = XML::LibXML->new();

my $working_organism="human";
my $genes_of_interest=("TP53","TBCCD1");
my $miRNAs_of_interest=("hsa-miR-1","hsa-miR-605","hsa-miR-195-5p");

#Using the source miRNA_gene_predictions

my $source="miRNA_gene_predictions";
foreach my $gene ($genes_of_interest){
    foreach my $miRNA ($miRNAs_of_interest){
        my $SourceListingURL = "http://mirgate.bioinfo.cnio.es/ResT/API/$working_organism/$source/$miRNA/$gene";
        my $xmldoc = $parser->parse_file($SourceListingURL);
        print "There is/are " . scalar(@{$xmldoc->findnodes('/miRGate/search/results/result')}) . " predictions for $miRNA and $gene\n\n";
        my $scount=1;
        foreach my $predicted_data ($xmldoc->findnodes('/miRGate/search/results/result')) {
            print $scount . " ";
            foreach my $pred_attr ($predicted_data->childNodes()){
                if($pred_attr){
                    print "\t " . $pred_attr->localname . ": " . $pred_attr->textContent . "\n" ;
                }
            }
            $scount++;
            print "\n";
        }
    }
}
```

Fig. 4 Example of API request using Perl programming language. It includes connection details, organism, genes, and miRNAs of interest in order to retrieve predictions stored in the databases without any filter. This and other examples are provided at the main API page at: <http://mirgate.bioinfo.cnio.es/API>

on Perl language is presented. It connects to the database using the API URL (<http://mirgate.bioinfo.cnio.es/ResT/API>). Then, human is selected as organism, *TP53* and *TBCCD1* as genes and *hsa-miR-1*, *hsa-miR-605* and *hsa-miR-195-5p* as miRNAs (any number of genes and miRNAs can be included without any limit).

Once required data is included, the program retrieves directly from the database all predictions belonging to specified genes and miRNAs. An XML response is sent and processed within the script. Finally all information is printed in a human readable manner.

In these examples, several API methods were shown, the complete catalog along with other examples can be consulted at the main API information web page at: <http://mirgate.bioinfo.cnio.es/API/>.

3.3 Applications of miRGate

miRGate predictions have been applied in different experimental studies to provide useful miRNA–mRNA binding sites. In all cases miRGate produced target sites that were robustly confirmed, demonstrating the suitability of this tool to the research community.

To illustrate this, miRGate calculated pairs from miRNAs virus–host target genes from Epstein-Barr and Kaposi sarcoma-associated herpes virus. In recent years, several authors have suggested different triggering mechanisms for cancer development, where pathogen's miRNAs may direct the expression of target genes on those infected cells. Slight evidences are found about virus's miRNAs, as most of other databases focus on intra-organism target predictions. All miRGate predictions showing high genomic agreement values were successfully validated in diffuse large B-cell lymphomas [35, 39, 40] as well as Burkitt Lymphoma samples infected with Epstein-Barr virus miRNAs [41].

The integrative approach incorporated in miRGate has been also employed in other studies [35, 36, 42] as a powerful tool for the analysis of miRNA-binding sites in hereditary breast tumour samples, hyperdiploid multiple myelomas, mantle cell lymphomas, and B-cell lymphomas where expression levels of isoforms and/or miRNAs were measured using distinct techniques. In all cases, the number of false-positive predictions was significantly reduced allowing the experimental corroboration of suitable miRNA–mRNA binding sites.

4 Notes

1. Users need to specify their identifiers of interest using accepted namespaces in miRGate. For miRNAs, all kind of identifiers provided by miRBase are accepted, such as accession numbers, miRNA ID, or mature IDs. Also Affymetrix miRNA probe names can be used. For genes, EnsEMBL transcript identifiers are accepted for isoform selection. Gene Symbols, EnsEMBL gene

identifiers, Affymetrix and Illumina probe names are also recognized. Typically, users will select gene symbols such as *TP53*, *BRCA1* and/or miRBase names for example *mmu-miR-15b-3p*.

2. To avoid processing errors, inserted identifiers (uploaded, pasted, or written), must be new line separated. Importantly, while writing an entity name, a type-ahead function is included to provide a similar and accepted identifier in miRGate.
3. The database stores all known 3' UTR sequences from human, mouse, and rat. When the extent of a given 3' UTR is not yet defined in EnsEMBL, we tag the sequence as “amplified,” and use the 130 bp downstream flanking sequence from the end of the last exon, as this number corresponds to the mode length of all known UTRs sequence size in those organisms. These 3' UTR amplified sequences are not included by default in the results, but can be added using the filtering options.
4. Information collected by experimental databases are obtained via automated text-mining techniques from titles and abstracts articles in PubMed. Each of these interactions is hyperlinked to PubMed by using PubMed identifiers (PMIDs). Of note, those experimentally supported predictions appear highlighted in the results web page using bold fonts.
5. The output shown in the result page displays all records ranked by the genomic agreement value in descendent order. Just by clicking in any column header, results are ordered attended to user selection. For each result, the page shows gene symbol (as specified by the user), miRNA name (as provided), EnsEMBL transcript IDs, miRNA mature name, HAVANA biotype, ENCODE isoform (three possible options: unknown isoform, longest, or principal isoform), type of UTR (amplified or not, *see Note 3*), 3' UTR target coordinates (start and end of the complementary binding site), *z*-score, and agreement. Interactions details, such as algorithm name, target site, coordinates, or interaction score, can be accessed pressing the “show predictions” button as showed in Fig. 2. Of interest, links to EnsEMBL [28] and APADB [43] (database for alternate polyadenylation that provides information of potential loss of miRNA binding sites) are included to obtain external information.
6. All results can be saved in csv format for subsequent analyses once “Download Excel (.csv)” button is pressed. This file contains the same information shown online, but includes four additional columns: (1) Computational Predictions (genomic agreement for the current target site upon selected algorithms), (2) Confirmed Predictions (number of confirmed databases supporting this gene-miRNA binding site), (3) Computational predictions information (details about the predictions), and (4), Confirmed predictions information (details about the validation methods including the PubMed identifier).

Acknowledgments

The authors would like to thank to Rocio Nuñez for critical reading of the manuscript.

References

1. Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75:843–854
2. Shukla GC, Singh J, Barik S (2011) MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. *Mol Cell Pharmacol* 3:83–92
3. Morozova N et al (2012) Kinetic signatures of microRNA modes of action. *RNA* 18: 1635–1655
4. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20
5. Orom UA, Nielsen FC, Lund AH (2008) MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol Cell* 30:460–471
6. Place RF et al (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci U S A* 105:1608–1613
7. Forman JJ, Legesse-Miller A, Collier HA (2008) A search for conserved sequences in coding regions reveals that the *let-7* microRNA targets *Dicer* within its coding sequence. *Proc Natl Acad Sci U S A* 105:14879–14884
8. Lim LP et al (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433: 769–773
9. Wu S et al (2010) Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region. *Oncogene* 29:2302–2308
10. Vinther J et al (2006) Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. *Nucleic Acids Res* 34:e107
11. Langenberger D et al (2010) Identification and classification of small RNAs in transcriptome sequence data. *Pac Symp Biocomput*:80–87
12. Huttenhofer A, Vogel J (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res* 34:635–646
13. Oulas A et al (2015) Prediction of miRNA targets. *Methods Mol Biol* 1269:207–229
14. Min H, Yoon S (2010) Got target? Computational methods for microRNA target prediction and their extension. *Exp Mol Med* 42:233–244
15. Farh KK et al (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* 310:1817–1821
16. Friedman RC et al (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19:92–105
17. Betel D et al (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 11:R90
18. Krek A et al (2005) Combinatorial microRNA target predictions. *Nat Genet* 37:495–500
19. Kertesz M et al (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39:1278–1284
20. Miranda KC et al (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding hetero-duplexes. *Cell* 126:1203–1217
21. Kruger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 34:W451–W454
22. Vergoulis T et al (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 40:D222–D229
23. Hsu SD et al (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 42:D78–D85
24. Nam S et al (2009) MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res* 37:W356–W362
25. Bisognin A et al (2012) MAGIA(2): from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res* 40:W13–W21
26. Ritchie W, Flamant S, Rasko JE (2009) Predicting microRNA targets and functions: traps for the unwary. *Nat Methods* 6: 397–398

27. Kozomara A, Griffiths-Jones S (2014) miR-Base: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 42:D68–D73
28. Flicek P et al (2014) Ensembl 2014. *Nucleic Acids Res* 42:D749–D755
29. Harrow, J (2015) Human and Vertebrate Analysis and Annotation (HAVANA)
30. Polisenio L et al (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465: 1033–1038
31. Gonzalez-Perez A, Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88:440–449
32. Xiao F et al (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 37:D105–D110
33. Wang D et al (2014) OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics* 30:2237–2238
34. Andres-Leon, E., et al., miRGate: a curated database of human, mouse and rat miRNA-mRNA targets. *Database* 2015, bav035.
35. Di Lisio L et al (2010) Mantle cell lymphoma: transcriptional regulation by microRNAs. *Leukemia* 24:1335–1342
36. Bueno MJ et al (2011) Combinatorial effects of microRNAs to suppress the Myc oncogenic pathway. *Blood* 117:6255–6266
37. Fish, S., (2015) XML::LibXML 2.0122 Parser library
38. Harrow J et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22:1760–1774
39. Di Lisio L et al (2012) MicroRNA signatures in B-cell lymphomas. *Blood Cancer J* 2:e57
40. Martin-Perez D et al (2012) Epstein-Barr virus microRNAs repress BCL6 expression in diffuse large B-cell lymphoma. *Leukemia* 26:180–183
41. Ambrosio MR et al (2014) The Epstein Barr-encoded BART-6-3p microRNA affects regulation of cell growth and immuno response in Burkitt lymphoma. *Infect Agent Cancer* 9:12
42. Tanic M et al (2012) Integration of BRCA1-mediated miRNA and mRNA profiles reveals microRNA regulation of TRAF2 and NFkappaB pathway. *Breast Cancer Res Treat* 134:41–51
43. Muller S et al (2014) APADB: a database for alternative polyadenylation and microRNA regulation events. *Database* 2014 bau076

Detection of microRNAs Using Chip-Based QuantStudio 3D Digital PCR

Cristina Borzi, Linda Calzolari, Davide Conte, Gabriella Sozzi*, and Orazio Fortunato*

Abstract

Digital PCR (dPCR) is an innovative approach for detection and quantification of nucleic acid that offers an alternative method to conventional real-time quantitative PCR for absolute quantification. dPCR is a highly precise and sensitive technique that does not require a standard reference, making it a suitable method for the detection of microRNAs. The potential of these small noncoding RNA as biomarkers is on the rise, especially due to their presence in body fluids, making them easily accessible. Nevertheless, the problem of lack of consensus regarding an optimal method for miRNAs normalization compromises their use. Here, we describe an innovative method for the absolute quantification of miRNAs across different types of biological samples using a chip-based platform, the QuantStudio 3D digital PCR.

Key words Digital PCR, Absolute quantification, miRNAs

1 Introduction

The concept of digital PCR (dPCR) was first introduced in 1992 to describe an end-point PCR method useful for absolute quantification [1]. Mainly, this method is based on limiting partition of the sample, which is diluted and subsequently split into hundreds or even millions of separate reaction chambers, which could contain one or no copies of the sequence of interest. Moreover, following PCR analysis, the fraction of negative reactions, according with Poisson's distribution, is utilized to generate an absolute count of the number of target molecules in a genomic or complementary DNA sample [1]. To date, two different types of dPCR instruments are available: one based on the utilization of chips [2–5] and the other based on droplets [6–9].

The precise quantification of nucleic acids is an important feature of dPCR, as it facilitates the measurement of quite small

*Joint senior authors.

percentage differences and, also, rare variants. For these reason, the digital approach is capable of higher sensitivity and precision compared to real-time quantitative PCR (qPCR) [10]. Unlike dPCR, qPCR uses the cycle threshold (CT) as a metric system, and thus CT values for the targets are referenced to an endogenous controls across samples and used for normalization. This calculation system is problematic as expression levels of the endogenous controls may be different across samples [1, 11], particularly true for biological fluids [12, 13]. dPCR does not rely on references or standards, and therefore it overcomes the problematic normalization issue.

Overall, dPCR has many potential applications, including the detection and quantification of low-level pathogens [14], rare genetic sequences [15], copy number variations [16], gene expression in single cells [17], and quantification of circulating miRNAs expression [18, 19].

microRNAs (miRNAs) are small noncoding RNAs of 19–24 nucleotides in length that regulate gene expression, at the post-transcriptional level [20, 21]. Today, it has become clear that expression levels of certain miRNAs are altered in many diseases, including cancer [22]. As such, the potential of these small non-coding RNAs as biomarkers has become obvious. Moreover, miRNAs have other features that make them ideally suited as biomarkers; for instance, miRNAs are present in biological fluids including plasma [23], serum [24], urine [25], and saliva [26], allowing relatively noninvasive sample collection. In addition to their accessibility, miRNAs are highly stable in biological fluids and in collected samples, making them relatively easy to work with.

Here, we describe a new method for the absolute quantification of miRNAs in different type of biological specimens using a chip-based platform, the QuantStudio 3D digital PCR (Fig. 1). This innovative data analysis tool circumvents normalization issues across different specimens while providing highly reproducibility of results.

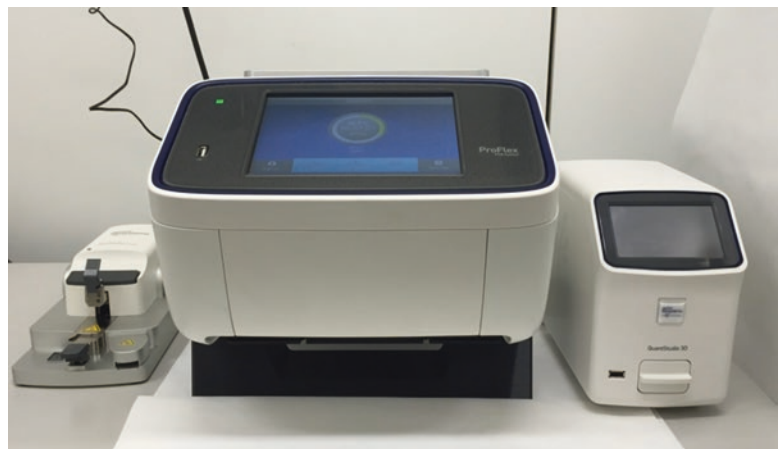


Fig. 1 QuantStudio 3D instruments. The figure show Chip Loader, ProFlex 2× Flat PCR System and QuantStudio™ 3D Digital PCR Instrument

The digital approach provides the isolation of total RNA from samples and miRNAs specific reverse transcription, an optional step of pre-amplification of cDNA, and finally PCR on chip using FAM probes specific for miRNA of interest. In the proposed method, we recommend to add in the mix an exogenous spike-in miRNA (cel-mir-39 conjugated with VIC fluorescence) as internal control for efficiency of the whole reaction (from extraction to PCR), since the fluorescence analysis using only FAM probes does not allow a precise quantification of miRNAs.

2 Materials

2.1 *mirVana miRNA Isolation Kit Components*

1. miRNA Wash Solution 1: Add 21 mL of 100% ethanol to the miRNA Wash Solution 1 and mix well. Store at room temperature (*see Note 1*).
2. Wash Solution 2/3: Add 40 mL of 100% ethanol to the Wash Solution 2/3 and mix well. Store at room temperature (*see Note 2*).
3. Cell disruption buffer: Store at 4 °C.
4. 2× Denaturing Solution: Add 375 µL of 2-mercaptoethanol (not provided with the kit) before use and mix well. Store at 4 °C. Warm up to room temperature before use (*see Note 3*).
5. Acid-Phenol–Chloroform: Store at 4 °C (*see Note 4*).
6. Elution Solution: it could be stored at any temperature (−20 °C, 4 °C or room temperature).
7. Collection tubes and filter cartridges.

2.2 *TaqMan MicroRNA Reverse Transcription Kit Components*

1. The kit contains: 10× RT buffer, dNTP mix w/dTTP (100 mM total), RNase inhibitor (20 U/µL), and MultiScribe™ RT enzyme (50 U/µL) for 200 reactions. Store at −15 to −25 °C.
2. Custom RT Primer Pool: To prepare the RT Primer Pool aliquot 25 µL of selected 5× RT primer into 1.5 mL eppendorf and dry down the tube in a speed vacuum at 50 °C for 1 h. Resuspend primers in 100 µL of nuclease-free H₂O.

2.3 *TaqMan MicroRNA Pre-amplification*

1. TaqMan PreAmp Master Mix (2×).
2. Custom PreAmp Primer Pool.
3. RNase-free water.
4. 1× TE Buffer.

2.4 *Digital PCR*

1. QuantStudio 3D Digital PCR Chip Lid v2.
2. QuantStudio 3D Digital PCR Sample Loading Blade.
3. QuantStudio 3D Digital PCR Chip v2.
4. Immersion Fluid syringe.
5. QuantStudio 3D Digital PCR Master Mix v2.

2.5 Material Not Provided in Kits

1. 2-mercaptoethanol (14.3 M).
2. 100% ethanol, ACS grade or better.
3. Fluorescent-labeled Taqman assays 20 $\times$.
4. *C. elegans* synthetic oligonucleotide cel-mir-39.

2.6 Equipment and Supplies

1. Heat block set to 95–100 °C.
2. Microcentrifuge capable of at least 10,000 $\times g$.
3. Vacuum manifold: to pull solutions through the filter cartridges.
4. RNase-free 1.5 or 0.5 mL polypropylene microfuge tubes, adjustable pipettors, and RNase-free tips.
5. Tissue homogenizer (for solid tissue samples).
6. Gene AMP PCR System 9700 Thermocycler.
7. QuantStudio 3D Digital PCR Thermal Pads.
8. QuantStudio 3D Digital PCR Chip Adapters.
9. ProFlex 2 $\times$ Flat PCR System.
10. QuantStudio 3D Tilt Base.
11. QuantStudio 3D Digital PCR Instrument.
12. QuantStudio 3D Digital PCR Chip Loader.
13. QuantStudio 3D AnalysisSuite Software.

3 Methods
3.1 RNA Isolation

Total RNA extraction from cell or biological fluids/plasma is performed using the mirVana PARIS Kit.

1. Start from 200 μ L of biological fluids/plasma or cell pellet (lysed using an appropriate amount of ice-cold cell disruption buffer). Importantly, add to each sample 5 μ L of cel-mir-39 from a 5 fmol/ μ L stock tube (Qiagen) (*see Note 5*).
2. Add an equal amount of 2 $\times$ Denaturing Solution at room temperature and incubate the mixture on ice for 5 min.
3. Add a volume of Acid-Phenol:Chloroform equal to the total volume of sample plus the 2 $\times$ Denaturing Solution. Vortex to mix and then centrifuge for 5 min at 10,000 $\times g$ at room temperature to separate the mixture into aqueous and organic phases.
4. Carefully remove the upper phase without disturbing the lower phase and transfer it to a fresh tube.
5. Add 1.25 volumes of room temperature 100% ethanol to the aqueous phase and mix thoroughly.
6. For each sample, place a filter cartridge into a collection tube. Pipet 700 μ L of this mixture onto the filter cartridge. Then, centrifuge for 30 sec at 10,000 $\times g$, discard the flow-through,

and repeat until all of the mixture is through the filter. Save the collection tube for the washing steps.

7. Apply 700 μL miRNA wash solution 1 to the filter cartridge and centrifuge for 30 s at $10,000 \times g$. Discard the flow-through and replace the filter cartridge in the same collection tube.
8. Apply 500 μL Wash Solution 2/3 and draw it through the filter cartridge as in the previous step. Repeat with a second 500 μL of Wash Solution 2/3. After discarding the flow-through from the last wash, replace the filter cartridge in the same collection tube and spin the assembly for 1 min at $10,000 \times g$ to remove residual fluid from the filter.
9. Transfer the filter cartridge into a fresh collection tube. Add 50 μL of elution buffer, preheated at 95°C , and centrifuge for 30 s at $10,000 \times g$ to recover the RNA. Collect the eluate and store at -80°C .

3.2 Reverse Transcription Polymerase Chain Reaction (RT-PCR)

A specific multiplex reverse transcription (RT) reaction, using TaqMan MicroRNA Reverse Transcription Kit and Custom TaqMan RT Primer Pool converts selected miRNAs into cDNA.

1. For each sample combine 3 μL or 100 ng of total RNA (*see Note 6*) with 6 μL of Custom TaqMan RT Primer Pool, 0.30 μL of dNTP's with dTTP (100 mM), 3 μL of multiscribe RT enzyme (50 U/ μL), 1.50 μL of RT 10 $\times$ Buffer, 0.19 μL of RNase Inhibitor (20 U/ μL), and 2 μL of ddH₂O.
2. Use these thermal cycling conditions: 16°C for 30 min, 42°C for 30 min, and 85°C for 5 min.
3. Dilute RT product with 180 μL of H₂O (not for preamplification step).

3.3 Pre-amplification Method (Optional)

Pre-amplification of cDNA is an optional step, but it is required for biological fluids/plasma or exosomes (*see Note 6*). Pre-amplification reaction could be performed using TaqMan Pre-amp mastermix and a Custom TaqMan PreAmp Primer Pool with miRNAs of interest.

1. Combine 2.5 μL of cDNA with 12.5 μL of TaqMan Pre-Amp Mastermix 2 $\times$, 3.8 μL of Custom TaqMan PreAmp Primer Pool, and 6.3 μL of ddH₂O.
2. Use these thermal cycling conditions: 95°C for 10 min, 55°C for 2 min, 72°C for 2 min, followed by 12 cycles at 95°C for 15 sec and 60°C for 4 min, and a final hold at 99.9°C for 10 min.
3. Dilute pre-amp product with 175 μL of 0.1 $\times$ TE Buffer.

3.4 Digital PCR on Chip

1. Prepare PCR reaction mix combining 2.25 μL of cDNA (RT or Pre-Amp product) with 3.75 μL nuclease-free H₂O, 7.50 μL of QuantStudio 3D Digital PCR Master Mix, 0.75 μL



Fig. 2 dPCR kit. Materials provided in the Quantstudio 3D kit for digital PCR (Chip, Chip Lid, Loading blade, and immersion oil syringe). On the right a chip assembled correctly

of TaqMan MicroRNA Assay, and 0.75 μL TaqMan MicroRNA Assay (cel-mir-39-VIC) (*see Note 7*).

2. Add sample mix (15 μL) on loading blade and load automatically on a chip using the QuantStudio 3D Digital PCR Chip loader (*see Notes 8 and 9*).
3. After loading, slowly add several drops of the immersion fluid syringe to cover the entire surface.
4. Once loaded a sample, the chip must be filled using the syringe with immersion fluid and sealed using the QuantStudio 3D Digital PCR Chip Lid (*see Notes 10 and 11*) (Fig. 2).
5. Perform PCR reaction on ProFlex™ 2 $\times$ Flat PCR System with the following program: 96 °C for 10 min, followed by 40 Cycles (2 min at 56 °C and 30s at 98 °C) and a final hold at 60 °C for 2 min.

3.5 Chip Analysis

After PCR reaction, the QuantStudio™ 3D Digital PCR 20 K Chips (Thermo Fisher Scientific) must be read and analyzed in order to obtain an absolute quantification of miRNA of interest.

1. Before analysis the chip surface has to be clean, so you have to remove any condensation or immersion fluid using a low-lint wipe.
2. Read chips using the QuantStudio™ 3D Instrument that performs imaging and primary analysis.
3. Perform the secondary analysis through the QuantStudio™ 3D AnalysisSuite Software to yield absolute quantification results from the raw imaging data (*see Note 12*) (Fig. 3).

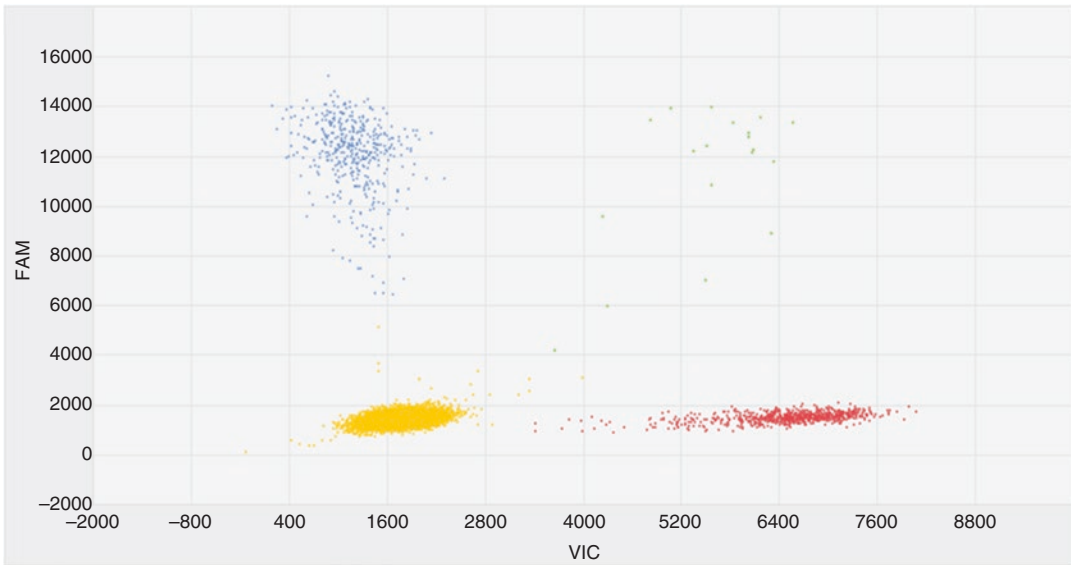


Fig. 3 Representative dPCR plots of the miRNA analysis. The data points in the plot are color-coded according to the following call types: FAM (blue), VIC (red), FAM + VIC (green), and NOT AMPLIFIED (yellow)

4 Notes

1. For storage longer than 1 month, store at 4 °C. Allow to warm up to room temperature before use. This buffer contains guanidinium thiocyanate, which is a potentially hazardous substance: it should be used with appropriate caution.
2. Wash Solution 2/3 will form a crystalline precipitate, due to the excess of EDTA that falls out of solution. We recommend to prepare it 1 day before use.
3. Heat to 37 °C 2× Denaturing Solution before use. 2× Denaturing Solution tends to precipitate at 4 °C. Before use, heat the solution to 37 °C, stirring occasionally, until it is completely in solution. Otherwise, to avoid solidification, store the 2× Denaturing Solution at room temperature for up to 1 month.
4. This reagent contains phenol, which is a poison and an irritant. Use gloves and other personal protection equipment when working with this reagent.
5. Synthetic *C. elegans* miRNA-39 (cel-miR-39) was used as spiked-in control, adding to each plasma sample 5 µL from a 5 fmol/µL stock tube.
6. For miRNA detection in biological fluids you have to use a fixed volume of 3 µL of total RNA because it cannot be precisely quantified using NanoDrop instrument.

7. Determine the optimal dilution when the target is known. Otherwise, it is possible that too many partitions will get multiple copies preventing an accurate estimation of the copies of miRNA.
8. To avoid pipetting errors, prepare a stock solution and include 10% excess for volume loss from pipetting.
9. Always wear gloves when loading and sealing chips and do not touch chip surface.
10. Avoid creating air bubbles when loading samples on blade.
11. To avoid evaporation apply the chip lid and add immersion fluid immediately after loading chip.
12. To get a precise quantification we suggest a threshold of 10.000 filled QT points for quality control of the chip.

Acknowledgments

This work was supported by AIRC (Associazione Italiana per la Ricerca sul Cancro): IG14318 to G.S., and 12162 (Special Program “Innovative Tools for Cancer Risk Assessment and early Diagnosis,” 5 × 1000); Italian Ministry of Health RF-2010-2306232 to G.S.

References

1. Sykes PJ, Neoh SH, Brisco MJ, Hughes E, Condon J, Morley AA (1992) Quantitation of targets for PCR by use of limiting dilution. *Biotechniques* 13:444–449
2. Warren L, Bryder D, Weissman IL, Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc Natl Acad Sci U S A* 103:17807–17812
3. Devonshire AS, Sanders R, Wilkes TM, Taylor MS, Foy CA, Huggett JF (2013) Application of next generation qPCR and sequencing platforms to mRNA biomarker analysis. *Methods* 59:89–100
4. Kreutz JE, Munson T, Huynh T, Shen F, Du W, Ismagilov RF (2011) Theoretical design and analysis of multivolume digital assays with wide dynamic range validated experimentally with microfluidic digital PCR. *Anal Chem* 83:8158–8168
5. Sundberg SO, Wittwer CT, Gao C, Gale BK (2010) Spinning disk platform for microfluidic digital polymerase chain reaction. *Anal Chem* 82:1546–1550
6. Conte D, Verri C, Borzi C, Suatoni P, Pastorino U, Sozzi S, Fortunato O (2015) Novel method to detect microRNAs using chip-based QuantStudio 3D digital PCR. *BMC Genomics* 16:849
7. Hatch AC, Fisher JS, Tovar AR, Hsieh AT, Lin R, Pentoney SL et al (2011) 1-million droplet array with wide-field fluorescence imaging for digital PCR. *Lab Chip* 11:3838–3845
8. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ et al (2011) Highthroughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* 83:8604–8610
9. Li M, Diehl F, Dressman D, Vogelstein B, Kinzler KW (2006) BEAMing up for detection and quantification of rare sequence variants. *Nat Methods* 3:95–97
10. Zhong Q, Bhattacharya S, Kotsopoulos S, Olson J, Taly V, Griffiths AD et al (2011) Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR. *Lab Chip* 11:2167–2174
11. Sanders R, Huggett JF, Bushell CA, Cowen S, Scott DJ, Foy CA (2011) Evaluation of digital PCR for absolute DNA quantification. *Anal Chem* 83:6474–6484
12. Dodd DW, Gagnon KT, Corey DR (2013) Digital quantitation of potential therapeutic target RNAs. *Nucleic Acid Ther* 23:188–194

13. Shen J, Jiang F (2012) Applications of MicroRNAs in the Diagnosis and Prognosis of Lung Cancer. *Expert Opin Med Diagn* 6: 197–207
14. Shen J, Stass SA, Jiang F (2013) MicroRNAs as potential biomarkers in human solid tumors. *Cancer Lett* 329:125–136
15. Gutierrez-Aguirre I, Racki N, Dreo T, Ravnikar M (2015) Droplet digital PCR for absolute quantification of pathogens. *Methods Mol Biol* 1302:331–347
16. Pekin D, Skhiri Y, Baret JC, Le Corre D, Mazutis L, Salem CB et al (2011) Quantitative and sensitive detection of rare mutations using droplet-based microfluidics. *Lab Chip* 11: 2156–2166
17. Marques FZ, Prestes PR, Pinheiro LB, Scurrah K, Emslie KR, Tomaszewski M et al (2014) Measurement of absolute copy number variation reveals association with essential hypertension. *BMC Med Genomics* 7:44
18. Ludlow AT, Robin JD, Sayed M, Litterst CM, Shelton DN, Shay JW et al (2014) Quantitative telomerase enzyme activity determination using droplet digital PCR with single cell resolution. *Nucleic Acids Res* 42:e104
19. Ma J, Li N, Guarnera M, Jiang F (2013) Quantification of plasma miRNAs by digital PCR for cancer diagnosis. *Biomark Insights* 8:127–136
20. Li N, Ma J, Guarnera MA, Fang H, Cai L, Jiang F (2014) Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol* 140:145–150
21. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297
22. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the microprocessor complex. *Nature* 432:231–235
23. Croce CM (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* 10:704–714
24. Shen J, Todd NW, Zhang H, Yu L, Lingxiao X, Mei Y et al (2011) Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab Invest* 91:579–587
25. Hu Z, Chen X, Zhao Y, Tian T, Jin G, Shu Y et al (2010) Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. *J Clin Oncol* 28: 1721–1726
26. Hanke M, Hoefig K, Merz H, Feller AC, Kausch I, Jocham D et al (2010) A robust methodology to study urine microRNA as tumor marker: microRNA-126 and microRNA-182 are related to urinary bladder cancer. *Urol Oncol* 28:655–661

MiRNA Quantitation with Microelectrode Sensors Enabled by Enzymeless Electrochemical Signal Amplification

Tanyu Wang, Gangli Wang, Didier Merlin, and Emilie Viennois

Abstract

Quantification of circulating microRNAs (miRNAs) is of great interest because of their potentials as disease biomarkers. Currently, quantitative reverse transcription polymerase chain reaction (qRT-PCR) and microarray are considered mainstream techniques for miRNA identification and quantitation. However, these techniques are challenged by the low levels and wide dynamic range (from aM to nM) of miRNAs in a physiological sample, as well as the difficulty in the implementation in point-of-care settings. Here, we describe a one-step label-free electrochemical sensing technique by assembling a triple-stem DNA-redox probe structure on a gold microelectrode and introducing a reductant, tris(2-carboxyethyl) phosphine hydrochloride (TCEP) in the detection buffer solution to achieve ultrasensitive miRNAs detection with a detection limit of 0.1 fM.

Key words miRNAs detection, Electrochemical sensor, Gold microelectrode, Nonenzymatic amplification, One-step detection, Label-free sensor, Folding-based sensor

1 Introduction

MiRNAs represent a class of single-strand, short, endogenous nonprotein-coding RNAs with approximately 19–23 nucleotides [1]. In recent years, increasing evidence has shown that miRNAs play important roles in a variety of physiological processes and the dysregulated expression of a specific group of miRNAs can be associated with many diseases [2–4] including different types of cancers [5]. Reports suggest that extracellular miRNAs circulating in the blood are extremely stable and resistance to degradation from RNases activity that make them promising biomarkers for the onset and progression of disease states [6]. However, the particularly small population of miRNAs (~0.01% of the total RNA mass), wide dynamic range (from aM to nM) and similarity among family members present major challenges for quantitative miRNA analysis [7–10]. Current mainstream techniques for miRNAs detection in circulation include microarrays and quantitative reverse

transcription polymerase chain reaction (qRT-PCR). Although microarrays offer high throughput analysis platform, the lack of sensitivity limits its application for quantifying low-level miRNAs [11–12]. The qRT-PCR provides high sensitivity that is particularly favorable for quantifying miRNAs with low levels but the methodology suffers from time-consuming multistep analysis and it only determines relative miRNA abundance with respect to a reference [13–16]. Thus, capable techniques featuring fast, low-cost detection with suitable sensitivity and selectivity for quantitative miRNA analysis remain to be established. Electrochemical techniques, among many others, have been developed to complement current methods and provide additional information as parallel analysis. Gao and Yang reported in 2006 an electrochemical miRNA detection platform with a detection limit of 80 fM [17]. A variety of electrochemical approaches have since been reported over the past decade for miRNA quantification and achieved remarkable sensitivity at lower aM level [18–20], typically with multistep operations.

Here, we detail a one-step label-free electrochemical strategy to quantify miRNAs with high sensitivity and selectivity [21]. The electrochemical sensing platform is on the basis of a gold micro-electrode fabricated in our laboratory, followed by the self-assembly of a triple-stem DNA-redox probe structure via terminal thiol on the surface of the Au electrode (Fig. 1). A multistep surface modification procedure was developed to improve the quality of the self-assembled monolayer (SAM) that would reduce interference signal caused by nonspecific adsorption. The procedure also reduces defects or pinholes in SAM on the electrode surface and

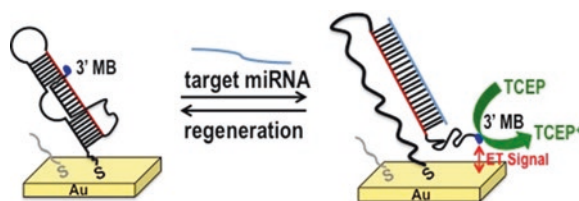


Fig. 1 Principle of current amplification in the signal-on microelectrode sensor. The thiol-MB-DNA probe is designed to have multiple segments including a sequence perfectly complementary to miR-122 (indicated in red) as target, an alkanethiol moiety at the 5' terminus for self-assembly onto the gold surface, and a MB moiety at the 3' terminus as redox signal reporter. In the absence of the target miRNA, the probe will self-hybridize into three distinct, short-base-pair stems that form a discontinuous double helix (*left*). Inert OH terminated alkane (C6) thiols are used to passivate the electrode surface as spacers to support the probe on the surface. The formation of longer double strand upon recognition of target miRNA releases a flexible segment near the MB-labeled 3' terminus and enables electron transfer. Reduction by TCEP in the solution of the electrode-oxidized MB on the surface amplifies the current signal

thus lowers background signals from surface bound redox signal probe alone. In the absence of target miRNAs, the probe is self-hybridized into a tripe-stem loop structure that locks the redox reporter, methylene blue (MB) away from the electrode surface thus minimizes electron transfer. Upon binding with target miRNA sequence, the MB molecule is released and approaches the electrode surface allowing electron transfer that gives an electrochemical signal. A new signal amplification mechanism without the need of a redox enzyme is introduced. By employing a reductant, tris(2-carboxyethyl) phosphine hydrochloride (TCEP) in the detection buffer solution, each redox molecule near the electrode surface upon target recognition is cyclically oxidized at the electrode and reduced by the reductant; thus, the signal is amplified in situ during the detection period. Overall, we have achieved one-step miRNA quantification with a detection limit of 0.1 fM via direct electrochemical readout in 10 min, with a wide detection range from sub-fM to nM.

2 Materials

All solutions were prepared with RNase/DNase-free, ultrapure distilled water and analytical grade reagents (unless indicated otherwise). Prepare all reagents at room temperature and stored at 4 °C (unless indicated otherwise). Diligently follow all waste disposal regulations when disposing of waste materials.

2.1 Electrode Preparation

1. 240-grit, 800-grit, 1200-grit sand papers.
2. Polishing cloth.
3. Aqueous slurry of 0.05 μm diameter alumina particles (*see Note 1*).
4. Electrochemical polishing solution: 0.1 M H_2SO_4 , pH 1.0. Add about 50 mL ultrapure water (prepared with deionized water at a resistivity higher than 18 M Ω cm at 25 °C) to a 200-mL graduated cylinder (*see Note 2*). Add 1.06 mL concentrated H_2SO_4 (95.0–98.0%) to the cylinder. Add water to achieve a total volume of 200 mL. Shake well to mix and store at room temperature.
5. Electrode characterization solution: 1 mM ferrocene, 0.1 M tetrabutylammonium perchlorate (TBAP). (*see Note 3*).

2.2 SAM Preparation

1. High salt buffer: 10 mM KH_2PO_4 , 30 mM Na_2HPO_4 , 1.55 M NaCl, 1 mM MgCl_2 , pH 7.4. Weight 10.2 mg of $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ and transfer to a 50-mL centrifuge tube. Add 50 mL phosphate buffered saline (PBS), 10X, pH 7.4 to the centrifuge tube. Shake well to mix and store at 4 °C (*see Note 4*).

Table 1

Tripe-stem DNA probe (recognition sequence marked in *red*) and signal-stranded RNA sequences (mismatches marked in *blue*) used in the microelectrochemical DNA-sensor

Sequence name	(Abbreviation)	Sequence, listed 5' to 3'
Triple-stem DNA probe	(DNA)	/C6 SS/GGA GTG TTT TTT TTC GTG TTT GTT TTT TCA AACACC ATT GTC ACA CTC CAT TTT TTT TTT TTG ACA ATG /MB/
mmu-miR-122	(miR-122)	UGG AGU GUG ACA AUG GUG UUU G
mmu-miR-122-3p	(miR-122-3p)	AGU GUG AUA AUG GCG UUU
2mismatch-22Base	(miR 22B)	UGG AGU GUG AUA AUG GCG UUU G

2. DNA probe (thiol-MB-DNA) stock solution: 200 μM thiol-MB-DNA (*see Note 5*) (Table 1), pH 7.4. Add 135.7 μL high salt buffer to the tube containing 27.14 nmol of thiol-MB-DNA in dry form (*see Note 6*). Vortex thoroughly and spin down all solution. Aliquot 5 μL into each individual 0.2-mL PCR microtube. Store at -80°C , covered with aluminum foil (*see Note 7*).
3. Reductant (TCEP) solution: 25 mM TCEP. Weight 10.7 mg of TCEP and transfer to a 2-mL PCR tube. Add 1.5 mL high salt buffer to the tube. Vortex thoroughly and spin down all solution. Store at 4°C .
4. Passivation solution: 1 mM mercaptohexanol. Add 1499.8 μL high salt buffer to a 2-mL PCR tube. Add 0.2 μL mercaptohexanol to the tube (*see Note 8*). Mix well and store at 4°C .
5. Hybridization buffer: 0.25 mM KH_2PO_4 , 0.75 mM Na_2HPO_4 , 39 mM NaCl, 10 mM MgCl_2 , pH 7.4. Weight 101.6 mg of $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ and transfer to a 50-mL centrifuge tube. Add 1.25 mL phosphate buffered saline, 10X, pH 7.4 to the centrifuge tube. Add 45 mL RNase/DNase-free, ultrapure distilled water to the centrifuge tube. Shake well to mix. Adjust pH with HCl (*see Note 9*). Make up to 50 mL with RNase/DNase-free, ultrapure distilled water. Store at 4°C (*see Note 4*).
6. Sensor regeneration solution: 4 M guanidine hydrochloride (GHCl). Weight 764.24 mg of GHCl and transfer to a 20-mL glass vial. Add 2 mL RNase/DNase-free, ultrapure distilled water to the vial. Vortex to mix and store at 4°C .

2.3 Electrochemical Measurement

1. Target miRNA stock solution: 100 μM miR-122 (*see Note 5*) (Table 1). Add 578.6 μL RNase/DNase-free, ultrapure distilled water to the tube containing 57.86 nmol of miRNA-122 in dry form (*see Note 6*). Vortex thoroughly and spin down all solution. Aliquot 10 μL into each individual 0.2-mL PCR microtube. Store at -80°C .

2. Mismatched miRNA stock solutions: 100 μM miR-122-3p and 100 μM miR 22B (*see Note 5*) (Table 1). Add 455.2 μL RNase/DNase-free, ultrapure distilled water to the tube containing 45.52 nmol of miRNA-122-3p in dry form (*see Note 6*). Vortex thoroughly and spin down all solution. Aliquot 10 μL into each individual 0.2-mL PCR microtube. Add 336.4 μL RNase/DNase-free, ultrapure distilled water to the tube containing 33.64 nmol of miRNA 22B in dry form and prepare the stock solution in individual microtube as in previous step. Store at -80°C .
3. Electrochemical instrument: Gamry Reference 600 electrochemistry workstation with a Ag/AgCl(s)/KCl(sat) reference electrode and a gold working electrode.
4. 50 mM TCEP solution: Weight 21.4 mg of TCEP and transfer to a 2-mL PCR tube. Add 1.5 mL high salt buffer to the tube. Vortex thoroughly and spin down all solution. Store at 4°C .

3 Methods

Carry out all procedures at room temperature unless otherwise specified. Keep reagents in ice unless otherwise specified.

3.1 Preparation of the Gold Microelectrode

1. Attach a 50 μm -diameter gold wire to a tungsten rod using conductive silver epoxy (*see Note 10*). The gold–tungsten assembly was then inserted into a foam board and placed upright with the gold wire facing up to dry for 4 hours (*see Note 11*).
2. Insert the gold–tungsten assembly into a soda lime glass capillary (*see Note 12*). Place the gold wire near the center of the glass capillary. The tip of the gold wire is about 1 mm inside the glass capillary (*see Note 13*).
3. Fix the glass capillary with gold wire placed inside in position with lab clamps to prevent the capillary from moving during sealing (*see Note 14*). The gold wire was then sealed into the capillary using a natural gas–oxygen flame (*see Note 15*). The other end of the assembly is sealed with a resin epoxy (*see Note 11*).
4. Remove the insulating glass around the Au tip from the end through manual polishing on 240- and 800-grit sand papers (from rough to fine) to expose a gold surface resulting a micro-disk Au electrode with about 50 μm diameter in glass shroud (*see Note 16*). Continue to polish the gold/glass surface to a mirror surface on a 1200-grit sand paper followed by an aqueous slurry of 0.05 μm diameter alumina particles (*see Note 17*).
5. Wash the gold microelectrode by flushing water over the electrode surface. Place the electrode in an ultrasonic cleaner with a mixture of water and methanol (3:1 v/v) for 10 min to

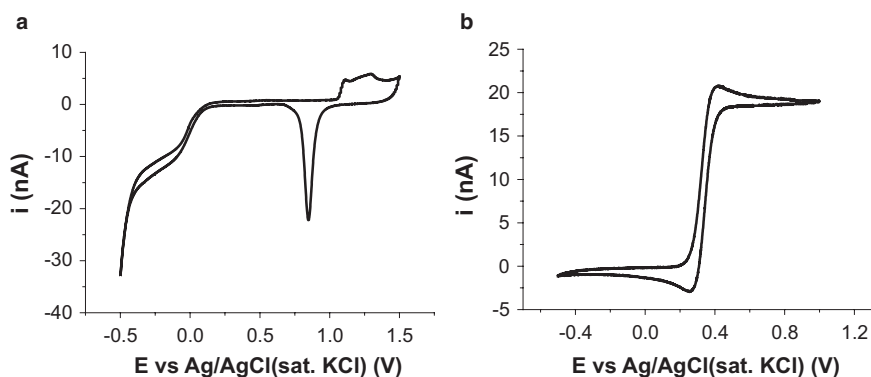


Fig. 2 Characterizations of a 25- μ m radius gold microelectrode in (a) 0.1 M H_2SO_4 at a scan rate of 100 $mV s^{-1}$; (b) 1 mM ferrocene solution at a scan rate of 100 $mV s^{-1}$

remove excess polishing particles. Wash the gold electrode with water and place back in the ultrasonic cleaner with a fresh water-methanol solution for an additional 10 min. Repeat this procedure for 3 times (*see Note 18*).

- Electrochemically polish the gold electrode by scanning the potential from -0.5 to $+1.5$ V in 0.1 M H_2SO_4 at a scan rate of 0.1 $V s^{-1}$ for 10 cycles with cyclic voltammetry technique (Fig. 2a). The cleaned gold electrode was thoroughly washed with water and ethanol and dried under flowing nitrogen.

3.2 DNA Monolayer Assembly

- Defrost the 200 μ M thiol-MB-DNA stock solution by placing the tube containing 5 μ L thiol-MB-DNA kept at $-80^\circ C$ on ice 30 min before experiment.
- Mix 1 μ L of 200 μ M thiol-MB-DNA with 2 μ L of 25 mM TCEP in a 0.2-mL PCR tube. The tube is wrapped with aluminum foil and incubated for 20 min at room temperature for reduction of disulfide bonds and to reduce the MB-moiety of the DAN probe. The color change of the mixed solution from blue to colorless indicates the completion of reduction reaction (*see Note 19*).
- Dilute the mixed solution with 197 μ L of high salt buffer to a total volume of 200 μ L with a final thiol-MB-DNA concentration of 1 μ M. Transfer the solution to a 3.7-mL glass vial.
- Transfer the previously cleaned gold electrode to the diluted and reduced thiol-MB-DNA solution and incubate for 16 h at room temperature in the dark (*see Note 20*).
- Remove excess thiol-MB-DNA physically adsorbed on the electrode surface with a room temperature water rinse for about 30 s (*see Note 21*).
- Transfer 200 μ L of 1 mM mercaptohexanol solution to a 3.7-mL glass vial. Immerse the DNA-probe-modified electrode in the passivation solution for 2 h at room temperature in dark.

7. Remove excess mercaptohexanol physically adsorbed on the electrode surface with a room temperature water rinse for about 1 min (*see Note 21*).
8. Place the surface passivated microelectrode in a 3.7-mL glass vial containing 500 μL of hybridization buffer. The vial was soaked in a water bath at 75 $^{\circ}\text{C}$ for 10 min. Cool down the hybridization solution in the glass vial gradually to room temperature in air.
9. Incubate the microelectrode sensor with 1 mM mercaptohexanol solution as in **step 6** for 1 h for further surface passivation.
10. Remove physically absorbed mercaptohexanol with water rinse as in **step 7**.
11. Store the electrochemical sensor in a 3.7-mL glass vial containing 1 mL of hybridization buffer at room temperature in dark.

3.3 Electrochemical Detection of MiR-122

1. Defrost the 100 μM miR-122 stock solution by placing the tube containing 10 μL miR-122 kept at -80°C on ice 30 min before measurement.
2. Add 2 mL of hybridization buffer to a 20-mL glass vial.
3. Place the Ag/AgCl(s)/KCl(sat) reference electrode in a salt bridge filled with 0.1 M NaClO_4 solution (*see Note 22*). Fix the positions of reference electrode and the gold working electrode with an electrode holder (*see Note 23*) and place in the glass vial containing hybridization buffer (Fig. 3).

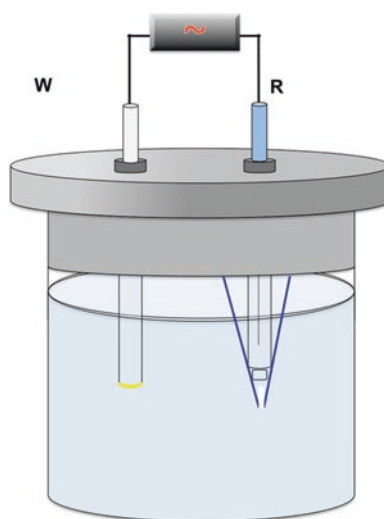


Fig. 3 Electrochemical cell assembly for the detection of miRNAs in hybridization buffer containing a gold working microelectrode and a Ag/AgCl(s)/KCl(sat) reference electrode. The gold electrode surface was modified with thiol-MB-DNA probes. The reference electrode was placed inside a salt bridge filled with 0.1 M NaClO_4

4. Attach electrodes with matching wires of the Gamry electrochemistry workstation. Measure the background signal in hybridization buffer at room temperature using cyclic voltammetry (CV) and square wave voltammetry (SWV) (Fig. 4). CV was scanned from -0.7 to $+0.2$ V at a scan rate of 5 V s^{-1} for 3 cycles versus Ag/AgCl reference. SWV was scanned from -0.6 to $+0.1$ V at a 100 Hz frequency with a 50 mV amplitude signal and 1 mV step versus Ag/AgCl reference.
5. Add miR-122 solution to the hybridization buffer to reach denoted final concentration in the buffer (*see* **Note 24**) (Table 2). Measure the signal using CV and SWV as in **step 4**.

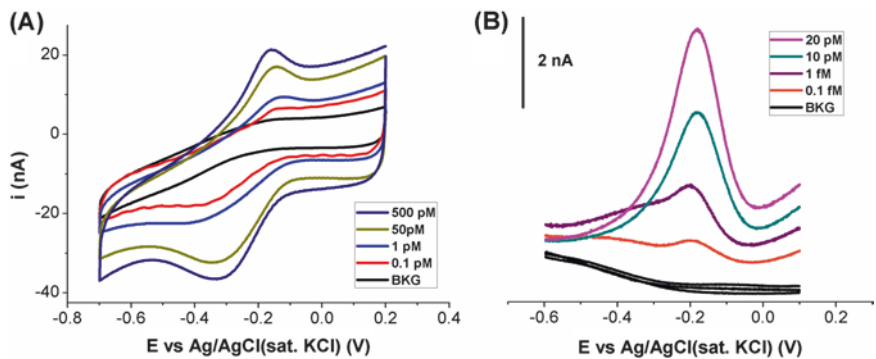


Fig. 4 Sensor responses in the absence and at different miR-122 concentrations. (a) CVs and (b) SWVs at representative concentrations. CV data were recorded at a scan rate of 5 V s^{-1} at 10 min after the addition of target. SWV frequency is 100 Hz. The sensor was regenerated (overlapping black curves in (b) as background demonstrating the excellent sensor recovery) following each measurement by a simple wash with 4 M GHCl. The background (BKG) was recorded with $160 \text{ }\mu\text{M}$ TCEP in hybridization buffer in the absence of target miRNAs

Table 2
Volume and concentration of miR-122 working solutions used in detection of target from 0.1 fM to 0.5 nM

Final [miR-122]	$V_{\text{working solution}}$ added (μL)	miR-122 working solution
0.5 nM	1	1 μM
0.1 nM	0.2	1 μM
10 pM	0.2	0.1 μM
1 pM	0.2	10 nM
0.1 pM	0.2	1 nM
10 fM	1	20 pM
1 fM	1	2 pM
0.1 fM	1	0.2 pM

6. Add 6.4 μL of 50 mM TCEP to the hybridization buffer to achieve a final concentration of 160 μM of TCEP in the buffer (*see Note 25*). Incubate for 10 min at room temperature. Measure the signal using CV and SWV as in previous step (*see Note 26*) (Fig. 4).
7. Rinse the electrochemical sensor with water for 1 min followed by a quick rinse in 4 M GHCl for 30 s (*see Note 27*). Incubate the sensor in 1 mL hybridization buffer at room temperature to fully restore sensor function.

3.4 Measurement with Mismatched MiRNA Sequences

1. Defrost the 100 μM miR-122-3p stock solution by placing the tube containing 10 μL miR-122-3p kept at -80°C on ice 30 min before measurement.
2. Defrost the miR 22 B stock solution as in **step 1**.
3. Add 999 μL RNase/DNase-free, ultrapure distilled water to a 2-mL PCR tube. Add 1 μL of the 100 μM miR-122-3p stock solution to the tube containing water to a final concentration of 0.1 μM .
4. Prepare diluted miR 22B to a final concentration of 0.1 μM as in **step 3**.
5. Set-up experiment as in **steps 2–4** in Subheading 3.3.
6. Add 4 μL of 0.1 μM miR-122-3p to the hybridization buffer to a final concentration of 200 pM.
7. Add 6.4 μL of 50 mM TCEP to the hybridization buffer to achieve a final concentration of 160 μM of TCEP in the buffer.
8. Incubate for 30 min at room temperature. Measure the signal using CV and SWV as in **step 4** in Subheading 3.3 (*see Note 28*).
9. Regenerate the electrochemical sensor as in **step 7** in Subheading 3.3.
10. Measure background signal as in **step 5**.
11. Add 4 μL of 0.1 μM miR 22B to the hybridization buffer to a final concentration of 200 pM.
12. Measure sensor response as in **steps 7–8** (*see Note 28*).

3.5 Diluted Mouse Serum Measurement

1. Transfer 2 μL mouse serum to a 0.2-mL PCR microtube.
2. Heat the serum aliquot at 95°C for 15 min and cool down to room temperature using a PCR thermocycler (Eppendorf, Hamburg, Germany) (*see Note 29*).
3. Spin down to collect all solutions.
4. Add 20 μL of hybridization buffer to a 3.7-mL glass vial.
5. Place a Ag/AgCl wire quasi-reference electrode and the gold working electrode with an electrode holder and place in the glass vial containing hybridization buffer (*see Note 30*).

6. Measure the background signal in hybridization buffer at room temperature using CV and SWV as in **step 4** in Subheading 3.3.
7. Add the serum aliquot after heating treatment to the hybridization buffer.
8. Incubate for 10 min at room temperature and measure signal as in **step 4** in Subheading 3.3.

4 Notes

1. Wear a mask when handling 0.05 μm diameter alumina particles. To avoid exposing alumina particles to coworkers, wet the polishing cloth with water and transfer the particles on the wet cloth. Slowly add more water to the alumina particles to make a slurry.
2. Always add the acid to water, in that order. Pour slowly with mixing to prevent rapid heat generation and spattering of the mixture.
3. The ferrocene solution can be stored at room temperature for repeated use in electrode characterization for about 1 week. Tightly close the vial containing the ferrocene/ACN with the matching plastic cap and seal with Parafilm to avoid solvent evaporation. In our laboratory we use the same solution for characterization of microelectrodes fabricated from the same batch.
4. Check the pH of the buffer with a pH meter regularly. Use a low concentration of NaOH (10 mM) to slowly bring the pH back to 7.4 if the pH is shifted due to CO_2 dissolution etc. over time. The buffer can be stored at 4 $^{\circ}\text{C}$ until finish. However, in our laboratory we make the buffer fresh about every month.
5. The commercially synthesized oligonucleotides are shipped and received as lyophilized product. The dried DNA probe is very easy to resuspend in an aqueous solution. It is also possible for the dried oligonucleotide to become dislodged from the tube during shipping. Thus, it is very important to spin down every oligonucleotide prior to opening the tube for resuspension. In our laboratory, we resuspend dry oligonucleotides to a storage stock concentration of 200 μM for DNA probes (100 μM for miRNAs) and then dilute a portion of this to create working stock solutions. The resuspended stock solution can be frozen at -80°C in small aliquots for day-to-day use to minimize the number of freeze-thaw cycles.
6. Before the resuspension of the oligonucleotides, it is recommended to clean the working area with 30% bleach. We wipe the surface of the bench top and the tip portion of the pipette. It is also recommended to use sterile pipette tips and PCR

microtubes. This is a precautionary practice for synthetic miRNAs that are not stabilized as those from real biological samples.

7. For long-term storage, the divided aliquots can be frozen at -80°C in dark for over 1 year. The thiol-MB-DNA probe contains MB as redox label thus need to be covered with aluminum foil to avoid light.
8. Mercaptohexanol is in liquid form and has a strong smell that may cause respiratory irritation. Always handle the thiolate in a well-ventilated fume hood.
9. The hybridization buffer is basic (slightly over pH 8) prior to pH adjustment. Concentrated HCl (0.1 M) can be used initially to narrow the gap from the starting pH to the required pH. From then on it would be better to use a lower concentration HCl (10 mM) with lower ionic strength to avoid a sudden drop in pH below the required pH.
10. Place a long gold wire on a piece of A4 paper. Cut the gold wire with a blade to several short pieces of about 1 cm. Brush a thin layer of silver epoxy around the tip of tungsten rod. Use the silver paste on tungsten to pick up the short gold wire along the tungsten rod.
11. For silver conductive epoxy, curing times and overall conductivity depend primarily on temperature. Room temperature curing is achievable in 4 h at above 24°C . For faster curing times, heat the bond to between $65\text{--}120^{\circ}\text{C}$ for 5–10 min and allow to cool. For resin epoxy, the heat-producing reaction begins after mixing the resin and the hardener. At room temperature, no reaction will be noticeable for about 20 min (no hardening observed). A drying time of 8–10 hours is required at room temperature. Faster curing is possible by placing the bond in oven for 3–4 hours at 80°C . When curing in oven, lay the electrode with glass capillary flat on a glass petri dish and rotate the electrode every hour to ensure even coating and complete sealing of the epoxy on the electrode and capillary.
12. To fix the tungsten rod with gold wire in positions inside the glass capillary, slightly bend the tungsten rod in the middle section before the insertion.
13. If the gold wire cannot be placed in the center of the capillary, pull the tungsten/gold wire assembly out of the capillary and use finger to gently straight up the gold wire. Wear gloves when touching the gold wire to avoid contamination.
14. Secure the glass capillary with a lab clamp on a ring stand above the position of a Meker burner. Place a microscope in front of the ring stand and focus on the tip portion of the glass capillary to monitor the sealing process.
15. The inner flame with higher temperature can be used to melt the glass at the initial stage. As soon as the glass starts to melt,

switch to the outer flame with lower temperature to avoid burning and bubble formation. Stop the heating immediately when the tip of the gold wire is less than 1 mm to the outer membrane of the sealed glass.

16. The microelectrode is made of glass capillary thus require caution when handling. Hold the electrode upright and toward the base (gold side) during polishing. Remove the backing and affix a piece of Buehler sand paper to the bench. Add a small amount of water on to the sand paper. Polish the electrode against the sand paper by drawing 8. The electrode surface may not be perpendicular to the sealed glass membrane. Find the position of electrode surface under microscope and keep the electrode surface parallel to the surface of sand paper when polishing. Check with a conductivity meter frequently to see if the gold has been exposed while polishing with the 240-grit sand paper. Wash the electrode with water and switch to the 800-grit sand paper after the gold surface is exposed to air.
17. Alumina particles tend to stick to the electrode surface and are hard to clean with simple washing. Therefore, use a small amount of alumina particles and apply less pressure when polishing against the alumina slurry.
18. Depending on the position of the gold wire during sealing and the different angles in polishing, the actual exposed gold area could vary. It is recommended to use 1 mM ferrocene solution in ACN to characterize the electrode surface prior to electrochemical polishing. CV is scanned from -0.4 to $+1.0$ V at a scan rate of 100 mV s^{-1} for 3 cycles versus Ag/AgCl reference (Fig. 2b). The following equation can be used to find out the actual electrode area: $i = 4nFD_0C_0^*r$ where i (A) is the steady state current, n is the number of electron transfer ($n = 1$), F is Faraday constant, D_0 is the diffusion coefficient of ferrocene in ACN ($2.69 \times 10^5 \text{ cm}^2 \text{ s}^{-1}$), C_0^* (mol cm^{-3}) is the concentration of ferrocene, and r (cm) is the radius of the electrode.
19. Depend on the molecular environment of MB and the lab conditions, the TCEP concentration can vary from 5 to 50 mM to completely reduce the disulfide bond and the MB molecule within an appropriate time (typically within 30 min). It is recommended to start with a lower concentration of TCEP (5 mM) for reduction. If the solution color does not turn colorless within 30 min, gradually increase the TCEP concentration (e.g., 10 and 25 mM) used for reduction until identify the suitable concentration.
20. Use a hole puncher with proper size to make a hole on the plastic cap that comes with the 3.7-mL glass vial. The electrode needs to fit snugly in the hole and can be fixed at a position. Make sure the electrode surface is not in touch with the bottom of glass vial during incubation.

21. Flow water over the electrode surface with a water bottle. It is not recommended to directly flush water to the electrode surface.
22. The Ag/AgCl(s)/KCl(sat) reference electrode requires pre-conditioning to reach equilibrium by incubation in 3 M KCl overnight after prolonged storage. It is recommended to keep the reference electrode in 3 M KCl when not in use. We place the reference electrode in a homemade salt bridge during electrochemical measurements to maintain charge balance and prevent cross-contamination. The salt bridge is fabricated with a 1-mL pipette tip. Cut a tiny piece toward the tip of the pipette tip to increase the opening very slightly. Roll a small piece of Kimwipes and insert it in the pipette tip. Use a hard stick to push it all the way inside the pipette tip and cut the excess portion sticking out of the small opening. Fill with 0.1 M NaClO₄ to make sure it is not leaking heavily.
23. Use a hole puncher with proper size to make two holes on the white plastic cap that comes with the 20-mL glass vial. Both the salt bridge holding the reference electrode and the gold working electrode need to fit snugly in the holes and can be fixed at a position. Place the two electrodes close to each other but not in touch and make sure the electrode surface is not in touch with the bottom of glass vial.
24. Dilute a portion of the 100 μ M stock solution to an appropriate working concentration. To reach a final miR-122 concentration in the fM–pM range, multiple dilutions of stock solution are required. An example of the dilutions is shown in Table 2 but can be adjusted based on specific needs.
25. The appropriate TCEP concentration used in the hybridization buffer to amplify signal was established. CV responses of a DNA-modified electrode were recorded at different TCEP concentrations: 40, 80, 120, and 160 μ M at a scan rate of 5 V s⁻¹. The 160 μ M is the highest TCEP concentration that would not interfere with measurements within the detection potential window of MB.
26. The change of detection frequency in SWV has impacts on the DNA sensor response [22] due to its sensitivity nature to electrode reaction rates. We measured the SWV signal for our sensor at 100, 200, 500, and 1000 Hz.
27. Hold the electrode with one hand and immerse it in the GHCl solution. Swirl the solution for 30 s. The GHCl was kept in fridge at 4 °C and can be used repeatedly for sensor regeneration. The GHCl solution does not need to be brought to room temperature for electrode rinse. In our laboratory, we make the GHCl solution fresh about every 2 weeks.

28. To affirm the efficacy of the exact sensor. 0.4 μL of the 0.1 μM miR-122 working solution was added into the same hybridization buffer immediately after the electrochemical measurements of the mismatched miRNAs. CV and SWV were recorded after a 10-min incubation with miR-122.
29. Circulating miRNAs are either stabilized by proteins or encapsulated in vesicles in serum/blood [1]. Therefore, the heating treatment of the serum sample is required to release miRNAs from membrane structure of microvesicles and associated proteins.
30. The Ag/AgCl quasi-reference electrode is made with ion exchange reaction. Bathing the Ag wire in a saturated solution of FeCl_3 to form the AgCl layer on the Ag for overnight. Take out the AgCl covered Ag wire from bathing solution with tweezers and wash with water. Gently polish one end of the wire with 1200-grit sand paper to expose portions of Ag to create Ag/AgCl. Polish the other end with the sand paper to completely expose Ag for conductivity. Tilt the 3.7-mL glass vial containing 20 μL hybridization buffer during measurement to ensure that both electrodes are immersed in the solution.

Acknowledgment

This work was supported by grants from the Department of Veterans Affairs (Merit Award: BX002526 to D.M.) and the National Institute of Health of Diabetes and Digestive and Kidney by the Grants RO1-DK-071594 and RO1-DK-064711 (to D.M.).

References

1. Dong H, Lei J, Ding L, Wen Y, Ju H, Zhang X (2013) MicroRNA: function, detection, and bioanalysis. *Chem Rev* 113:6207–6233
2. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20
3. Cullen BR (2009) Viral and cellular messenger RNA targets of viral microRNAs. *Nature* 457:421–425
4. Laterza OF, Lim L, Garrett-Engele PW, Vlasakova K, Muniappa N, Tanaka WK, Johnson JM, Sina JF, Fare TL, Sistare FD, Glaab WE (2009) Plasma MicroRNAs as sensitive and specific biomarkers of tissue injury. *Clin Chem* 55:1977–1983
5. Hayes J, Peruzzi PP, Lawler S (2014) MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol Med* 20:460–469
6. Wang K, Zhang S, Marzolf B, Troisch P, Brightman A, Hu Z (2009) Circulating microRNAs, potential biomarkers for drug-induced liver injury. *Proc Natl Acad Sci U S A* 106:4402–4407
7. Michel V, Yuan Z, Ramsubir S, Bakovic M (2006) Choline transport for phospholipid synthesis. *Exp Biol Med* 231:490–504
8. Klein D (2002) Quantification using real-time PCR technology: applications and limitations. *Trends Mol Med* 8:257–260
9. Lim LP (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev* 17:991–1008
10. Liang Y, Ridzon D, Wong L, Chen C (2007) Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics* 8:166
11. Liu CG, Calin GA, Meloon B, Gamllet N, Seignani C, Ferracin M, Dumitru CD,

- Shimizu M, Zupo S, Dono M, Alder H, Bullrich F, Negrini M, Croce CM (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc Natl Acad Sci U S A* 101:9740–9744
12. Lodes MJ, Caraballo M, Suci D, Munro S, Kumar A, Anderson B (2009) Detection of cancer with serum miRNAs on an oligonucleotide microarray. *PLoS One* 4:e6229
13. Asaga S, Kuo C, Nguyen T, Terpenning M, Giuliano AE, Hoon DSB (2011) Direct serum assay for MicroRNA-21 concentrations in early and advanced breast cancer. *Clin Chem* 57:84–91
14. Zhi F, Cao X, Xie X, Wang B, Dong W, Gu W, Ling Y, Wang R, Yang Y, Liu Y (2013) Identification of circulating MicroRNAs as potential biomarkers for detecting acute myeloid leukemia. *PLoS One* 8:e56718
15. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, Ambros VR, Israel MA (2007) Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. *Cancer Res* 67:2456–2468
16. Zhang J, Li Z, Wang H, Wang Y, Jia H, Yan J (2011) Ultrasensitive quantification of mature microRNAs by real-time PCR based on ligation of a ribonucleotide-modified DNA probe. *Chem Commun* 47:9465–9467
17. Gao ZQ, Yang ZC (2006) Detection of microRNAs using electrocatalytic nanoparticle tags. *Anal Chem* 78:1470–1477
18. Labib M, Khan N, Ghobadloo SM, Cheng J, Pezacki JP, Berezovski MV (2013) Three-mode electrochemical sensing of ultralow MicroRNA levels. *J Am Chem Soc* 135:3027–3038
19. Labib M, Ghobadloo SM, Khan N, Kolpashchikov DM, Berezovski MV (2013) Four-way junction formation promoting ultrasensitive electrochemical detection of MicroRNA. *Anal Chem* 85:9422–9427
20. Labib M, Khan N, Berezovski MV (2015) Protein electrocatalysis for direct sensing of circulating MicroRNAs. *Anal Chem* 87:1395–1403
21. Wang T, Viennois E, Merlin D, Wang G (2015) Microelectrode miRNA sensors enabled by enzyme less electrochemical signal amplification. *Anal Chem* 87:8173–8180
22. White RJ, Plaxco KW (2010) Exploiting binding-induced changes in probe flexibility for the optimization of electrochemical biosensors. *Anal Chem* 82:73–76

A Robust Protocol to Quantify Circulating Cancer Biomarker MicroRNAs

Emma Bell*, Hannah L. Watson*, Shivani Bailey,
Matthew J. Murray†, and Nicholas Coleman†

Abstract

Reverse-transcriptase quantitative PCR (RT-qPCR) is a widely used method for quantifying microRNAs (miRNAs) in cells and tissues. However, the quantification of miRNAs in the circulation presents specific challenges. Here, we describe an optimized protocol using a small amount of input material to assess serum sample quality and quantify levels of a panel of up to 20 miRNAs. This is achieved by multiplexing Taqman miRNA stem-loop primers in the reverse transcription step. An additional multiplexed pre-amplification step is used to increase the sensitivity of the final quantification step, which is carried out using standard Taqman qPCR methodology.

Key words MicroRNA, Serum, Plasma, Cerebrospinal fluid, Pre-amplification, RT-qPCR†Joint senior authors.

1 Introduction

The accurate quantification of cancer-specific microRNAs (miRNAs) in body fluids represents a specific challenge to investigators [1] on account of their relatively low abundance in the circulation and potential variation in levels due to preanalytical parameters, e.g., collection tube type, differences in storage and processing, patient condition, etc. Here, we describe a RT-qPCR method that allows the sensitive detection of multiple miRNAs, using a small amount of input RNA from samples that are first validated by a robust quality control (QC) screen. For serum, this screen consists of RT-qPCR quantification of two endogenous markers to assess hemolysis (namely hsa-miR-451a and hsa-miR-23a-3p), an endogenous housekeeping miRNA (hsa-miR-30b-5p) and an exogenous,

*Joint first authors.

†Joint senior authors.

nonhuman spike-in control miRNA [*Caenorhabditis elegans* (cel-) miR-39-3p] [2, 3].

Hemolysis results in the release of erythrocyte-specific miRNAs from red blood cells into the serum, causing substantially higher circulating levels of these miRNAs and distortion of the overall miRNA profile [4–6]. Therefore, hemolyzed samples must be identified prior to further testing. A set of guidelines defining serum hemolysis criteria has been described previously [3] and is incorporated in the QC screen detailed in this protocol. Specifically, levels of the erythrocyte-specific miRNA, hsa-miR-451a, are compared with those of the stable endogenous housekeeping miRNA, hsa-miR-23a-3p. Where the delta Cq miR-23a-3p – miR-451a is over the designated value of eight [2, 3], the sample is considered to be substantially hemolyzed. Ideally, a further serum sample should be sourced and/or RNA extraction repeated, and a further QC screen run. If no such sample is available, then any subsequent data obtained from the sample with a delta Cq value >8 are to be interpreted with caution.

The QC screen also includes an assessment of sample integrity. Firstly, a measure of sample RNA extraction efficiency is incorporated. Poor RNA extraction efficiency can result in the complete loss of low abundance miRNAs from the sample, which may in turn reduce the sensitivity of the test. In order to measure this RNA extraction efficiency, a synthetic nonhuman spike-in, cel-miR-39-3p is added during RNA extraction which is then probed for during the QC RT-qPCR steps [7]. Secondly, levels of the endogenous housekeeping miRNA, hsa-miR-30b-5p, provide an accurate surrogate for overall levels of endogenous miRNA in the sample [2, 8, 9], with high Cq values indicating a degraded sample of low integrity. In addition to their roles in the QC screen, cel-miR-39-3p [2, 10] and hsa-miR-30b-5p [2] are also routinely used during the subsequent analysis of multiplex data for normalization of both variability in RNA extraction and pre-amplification efficiency, as well as endogenous miRNA levels, respectively.

Once sample quality has been verified, it is possible to proceed to quantify specific miRNAs of interest. In standard singleplex RT-qPCR, identical Taqman miRNA stem-loop primers hybridize to their target miRNAs, which are then reverse transcribed by the Multiscribe reverse transcriptase enzyme [11]. The cDNA products of this miRNA are then quantified using standard Taqman qPCR methods, demonstrated to discriminate effectively between miRNAs of only one nucleotide difference (e.g., members of the *let-7* family) and which have a dynamic range of seven orders of magnitude [11]. However, performing singleplex reverse transcription reactions for each miRNA under interrogation (for both test and normalization purposes) uses large amounts of RNA and reagents, increasing costs. This protocol describes a multiplex approach to maximize the number of miRNAs measured from a

single volume of starting material. This is achieved by combining multiple different stem-loop primers in the reverse transcription step and subsequently pre-amplifying the cDNA products prior to the final singleplex qPCR step [12, 13]. Following our description of this approach [14], this strategy has been successfully adopted in several other studies examining circulating miRNAs [2, 15–17].

In principle, the methods described can be used for miRNA quantification in other body fluids, for example cerebrospinal fluid (CSF) [2]. However, the quality control steps and housekeeping miRNAs detailed in this protocol have been specifically developed and validated for use on serum samples. Recent studies have found that housekeeping miRNAs should be validated for each disease and body fluid [18, 19]; therefore, we recommend optimizing a housekeeping miRNA panel and QC screening methods for each sample type.

2 Materials

All reagents and plastic-ware must be sterile and nuclease free. RNase-free working practices should be implemented, e.g., cleaning with RNase neutralizing agents and use of filter pipette tips. Preparation of solutions should be carried out at room temperature unless otherwise stated. Frozen reagents should be thawed on ice. Reagents should be kept on ice where appropriate.

2.1 Serum Isolation from Whole Blood

1. Appropriate volume tube with clotting activator and gel for serum separation, e.g. Sarstedt S-monovette Z-gel tube.
2. 2 ml RNase-free cryovials.

2.2 RNA Extraction

1. Qiagen miRNeasy serum/plasma kit: Qiazol lysis reagent, RNeasy MinElute spin columns in 2.0 ml collection tubes, buffer RWT, buffer RPE, 2.0 ml collection tubes, 2 ml collection tubes, RNase-free water. Prior to first use, dilute buffer RWT and RPE with two and four volumes of 96–100% ethanol respectively.
2. cel-miR-39-3p stock solution: Qiagen miRNeasy serum/plasma spike-in control, nuclease-free water. Centrifuge the lyophilized miRNeasy serum/plasma spike-in control for 30 s at $12,000 \times g$, add 300 μ l nuclease-free water, and then vortex to give a 2×10^{10} copy/ μ l solution (*see Note 1*).
3. cel-miR-39-3p working solution: Roche MS2 bacteriophage RNA, nuclease-free water, cel-miR-39-3p stock solution (*see above*). Prepare a solution of 10 ng/ μ l MS2 bacteriophage RNA in nuclease-free water. Add 4 μ l of cel-miR-39-3p stock solution to 16 μ l of the 10 ng/ μ l MS2 RNA solution and then briefly vortex. Transfer 2 μ l of this mix to a further 48 μ l of the 10 ng/ μ l MS2 RNA solution and then briefly vortex, giving a

1.6×10^8 copies/ μ l working solution sufficient to process 12 samples (*see* **Note 1**).

4. QIAzol-MS2-cel-miR-39-3p mix: Roche MS2 bacteriophage RNA, nuclease-free water, Qiagen QIAzol lysis reagent (from kit), cel-miR-39-3p working solution (*see* above). Prepare a solution of 0.8 μ g/ μ l MS2 bacteriophage RNA in nuclease-free water. Make a stock solution consisting of 1000 μ l QIAzol lysis reagent, 1.56 μ l of 0.8 μ g/ μ l MS2 bacteriophage RNA, and 3.5 μ l cel-miR-39-3p working solution per sample to be tested, plus an additional reaction to allow for pipetting error. Invert to mix.
5. Chloroform.
6. 100% ethanol.
7. 80% ethanol made with nuclease-free water.

2.3 Quality Control

Reverse

Transcription Step

1. Nuclease-free water.
2. Life Technologies TaqMan MicroRNA Reverse Transcription Kit: 10 $\times$ RT buffer, dNTP mix w/dTTP (100 M), RNase inhibitor (20 U/ μ L), MultiScribe RT enzyme (50 U/ μ L).
3. Life Technologies TaqMan microRNA Assay primers 5 $\times$ (for serum, we recommend using cel-miR-39-3p, hsa-miR-30b-5p, hsa-miR-23a-3p, and hsa-miR-451a as the quality control miRNAs. For details *see* Table 1).
4. RNA sample.
5. 0.2 ml RNase-free PCR wells (use most appropriate form, i.e., tube, strip, plate).
6. PCR well seals.
7. Thermal cycler capable of the program detailed in Table 2.

Table 1

Recommended Taqman miRNA assays: quality control and housekeeping miRNAs

miRNA	Role in reaction	miRBase accession number	5' to 3' nucleotide sequence
cel-miR-39-3p	Nonhuman spike-in	MIMAT0000010	UCACCGGGUGUAAAUCAGCUUG
hsa-miR-30b-5p	Housekeeping miRNA	MIMAT0000420	UGUAAACAUCCUACACUCAGCU
hsa-miR-30c-5p	Housekeeping miRNA	MIMAT0000244	UGUAAACAUCCUACACUCUCAGC
hsa-miR-191-5p	Housekeeping miRNA	MIMAT0000440	CAACGGAAUCCCAAAGCAGCUG
hsa-miR-23a-3p	Hemolysis assessment	MIMAT0000078	AUCACAUUGCCAGGGAUUUCC
hsa-miR-451a	Hemolysis assessment	MIMAT0001631	AAACCGUUACCAUACUGAGUU

Table 2
Thermal cycler program for reverse transcription

Temperature (°C)	Time (min)
16	30
42	30
85	5
4	∞

2.4 Custom MicroRNA Panel Multiplex Reverse Transcription Step

1. Nuclease-free water.
2. Life Technologies TaqMan MicroRNA Reverse Transcription Kit: 10× RT buffer, dNTP mix w/dTTP (100 M), RNase inhibitor (20 U/μL), MultiScribe RT enzyme (50 U/μL).
3. Custom multiplex primer mix: Multiple selected Life Technologies TaqMan MicroRNA Assay primer 5×. In advance, select and order assays for the custom miRNA biomarker panel (*see Note 2*). Each multiplex RT reaction uses a total of 3 μl of primer made up of equal quantities of primers for each target miRNA. Therefore, to calculate the volume of each Taqman primer needed per multiplex RT reaction, the total 3 μl volume is divided by the number of miRNAs to be tested, e.g., 3 μl/20 miRNAs = 0.15 μl of each Taqman primer. This is then scaled up by a factor of the number of reactions needed (*see Note 3*), i.e., ten reactions would require 1.5 μl of each Taqman primer, which are pipette-mixed to give a total volume of 30 μl of multiplex primer mix (*see Table 3*).
4. 0.2 ml RNase-free PCR tubes (tube, strip, or plate format).
5. RNA sample.
6. PCR well seals.
7. Thermal cycler capable of the program detailed in Table 2.

2.5 Multiplex Pre-amplification Step

1. Life Technologies 2X Taqman Pre-amp Master Mix.
2. Custom multiplex probe mix: Multiple selected Life Technologies TaqMan MicroRNA Assay probe 20×, nuclease-free Tris-EDTA (10 mM Tris, 1 mM EDTA) pH 8.0. Taqman MicroRNA Assay probes 20× should be protected from the light. Each multiplex pre-amplification (pre-amp) reaction uses a total of 12.5 μl of custom multiplex probe mix (*see Note 3*), made up of equal quantities of each probe, each diluted to 0.2x. Therefore, combine equal volumes of each Taqman probe and pipette gently to mix. Using nuclease free Tris-EDTA (10 mM Tris, 1 mM EDTA) pH 8.0, dilute

Table 3
Reaction volumes for multiplex reverse transcription mastermix

Reagent	Volume for one reaction (μl)	e.g. volume for ten reactions (μl)
Nuclease-free water	4.16	41.60
10× RT buffer	1.50	15.00
dNTP mix w/dTTP (100 M)	0.15	1.50
RNase inhibitor (20 U/μl)	0.19	1.90
Custom multiplex primer mix 5× (<i>see</i> Subheading 2.4; step 3)	3.00 ^a	30.00 ^b
MultiScribe RT enzyme (50 U/μL)	1.00	10.00
Total	10.00	100.00

^aVolume of each individual RT primer = 3 μl/ number of microRNAs tested

^bFor ten reactions the volume of each individual RT primer = (3 μl/number of miRNAs tested) × 10

the probe mix so that each individual probe in the multiplex is at a concentration of 0.2x. E.g. 10 μl per probe (20x) in a 12 microRNA multiplex is mixed to give 120 μl total probe volume. This is then diluted with 880 μl of Tris-EDTA to give a total custom multiplex probe mix volume of 1000 μl, with each probe at 0.2x.

3. 0.2 ml RNase-free PCR wells (use most appropriate form, i.e., tube, strip, plate).
4. cDNA sample.
5. PCR seal sheets.
6. Thermal cycler capable of the program detailed in Table 4.
7. Nuclease-free water.

**2.6 RT-qPCR
for Individual
MicroRNAs**

1. Life Technologies TaqMan Universal Master-Mix II, no UNG.
2. Life Technologies TaqMan MicroRNA Assay probe 20×. Note that Taqman MicroRNA Assay probes 20× should be protected from the light.
3. Half-skirted white 96-well PCR plate.
4. Optical qPCR adhesive film.
5. Real-time qPCR thermal cycler capable of the program detailed in Table 5.

3 Methods

Use the flowchart (Fig. 1) to direct the protocol workflow through QC and multiplex steps, including verification. RNase-free working practices should be implemented, e.g., cleaning with RNase

Table 4
Thermal cycler program for multiplex pre-amplification

Temperature	Time
95 °C	10 min
Repeat 14 cycles of	
95 °C	15 s
60 °C	4 min
Hold	
4 °C	∞

Table 5
qRT-PCR program for quantification step

Temperature	Time
95 °C	10 min
Repeat 45 cycles of	
95 °C	15 s
60 °C	1 min

neutralizing agents, use of filter pipette tips, etc. Ensure all solutions are prepared prior to beginning each section. Serum isolation is performed at room temperature, all other steps are to be performed on ice (4 °C) unless otherwise stated.

3.1 Serum Isolation from Whole Blood

1. Collect whole blood by venepuncture into an appropriate volume tube with clotting activator and gel for serum separation, ensuring that the tube is filled.
2. Store filled tube in an upright position for 30 min.
3. Centrifuge the tube in a swinging bucket rotor at $2500 \times g$ for 10 min at room temperature.
4. Transfer 1 ml aliquots of the top serum layer into 2 ml RNase-free cryovials. The isolated serum may be used immediately or stored at $-80\text{ }^{\circ}\text{C}$.

3.2 RNA Extraction

Prior to starting, ensure that all materials for the RNA extraction are prepared (*see* Subheading 2.2).

1. Defrost the serum samples on ice, then centrifuge at $400 \times g$ for 5 min at 4 °C.
2. Transfer 200 μl of the sample supernatant to a 2.0 ml RNase-free tube, ensuring the pellet is not disturbed (*see* Note 4).

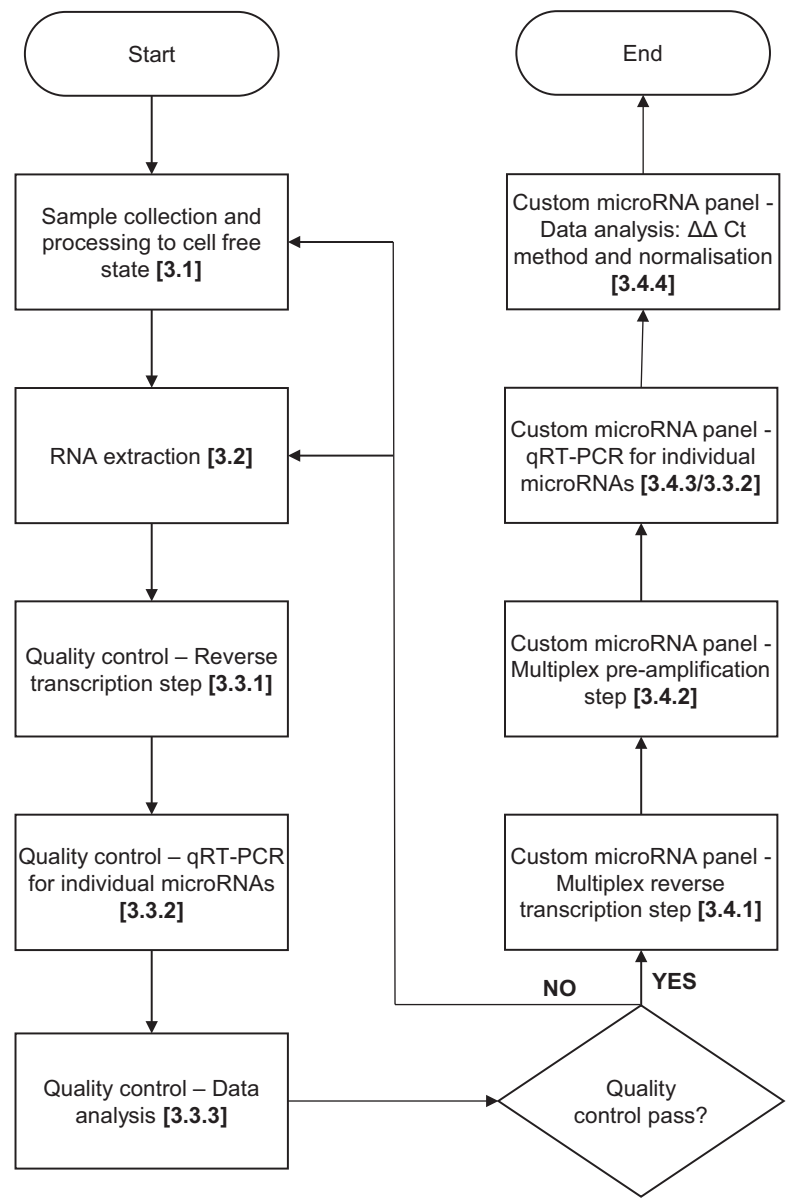


Fig. 1 Workflow plan

3. Add 1000 μl of QIAzol-MS2-cel-miR-39-3p mix (prepared according to Subheading 2.2; items 2–4) to each sample and vortex for 10 s. Incubate at room temperature for 5 min.
 4. Add 200 μl of chloroform to each sample and vortex for 15 s. Incubate at room temperature for 3 min.
 5. Centrifuge the samples at $12,000 \times g$ for 30 min at 4 °C.
- Steps 6–14 are performed at room temperature.

6. Transfer of the top aqueous phase layer into a 2.0 ml RNase-free tube, ensuring that the middle protein layer is not disturbed (*see Note 5*). Discard the bottom and middle layers.
7. Measure the volume of the aqueous phase, and add 1.5 times this volume of 100% ethanol. Mix well by pipetting.
8. Ensure that the RNeasy MinElute spin columns are inserted into 2.0 ml collection tubes. Transfer 750 μl of the sample to the column and centrifuge at $>8000 \times g$ for 30 s at room temperature. Discard the flow-through. Repeat until the entire volume of sample has been passed through the column.
9. Add 700 μl of buffer RWT to the column and centrifuge at $>8000 \times g$ for 30 s. Discard the flow-through.
10. Add 500 μl of buffer RPE to the column and centrifuge at $>8000 \times g$ for 30 s. Discard the flow-through.
11. Add 500 μl of 80% ethanol to the column and centrifuge at $>8000 \times g$ for 2 min. Discard both the collection tube and the flow-through.
12. Transfer the column to a fresh 2.0 ml collection tube and centrifuge at $>8000 \times g$ for 1 min.
13. Transfer the column to a 2 ml RNase-free collection tube and add 50 μl of nuclease-free water to the membrane. Incubate for 2 min at room temperature, then centrifuge at $>8000 \times g$ for 1 min.
14. Add a further 50 μl nuclease-free water to the membrane and incubate for 2 min at room temperature. Centrifuge at $>8000 \times g$ for 1 min. The eluate contains the extracted total RNA (*see Note 6*) which may be used immediately or stored at -80°C for subsequent use.

3.3 Quality Control

3.3.1 Quality Control Reverse Transcription Step

For serum QC, the levels of cel-miR-39-3p, hsa-miR-30b-5p, hsa-miR-23a-3p, and hsa-miR-451a are quantified using the method described below. Details of the assays used are found in Table 1.

1. Create a separate reverse transcriptase mastermix for each quality control miRNA. Prepare the mastermixes in 2 ml RNase-free tubes according to the volumes stated in Table 6. The volume prepared should be sufficient to test the samples and a no-template control (NTC) sample, plus 10% to allow potential pipetting error. Add the reagents in the order listed in the table and mix by gently pipetting.
2. Transfer 10 μl aliquots of each miRNA specific RT-mastermix into the appropriate number of 0.2 ml PCR wells. There should be one well per sample, including the NTC, for each miRNA tested.
3. Gently mix the RNA eluate. Add 5 μl of RNA sample to its corresponding well containing RT-mastermix. For the no template

Table 6
Reaction volumes for microRNA specific (initial QC) reverse transcription mastermix

Reagent	Volume for 1 reaction (μl)	e.g., volume for 12 reactions (μl)
Nuclease-free water	4.16	49.92
10× RT buffer	1.50	18.00
dNTP mix w/dTTP (100 M)	0.15	1.80
RNase inhibitor (20 U/μl)	0.19	2.28
TaqMan microRNA Assay primer 5×	3.00	36.00
MultiScribe RT enzyme (50 U/μL)	1.00	12.00
Total	10.00	120.00

Table 7
Reaction volumes for microRNA specific quantitative real-time PCR mastermix

Reagent	Volume required for 1 reaction (μl)	e.g., Volume required for 36 reactions (μl)
TaqMan Universal Master-Mix II, no UNG	10.00	360.00
Specific TaqMan microRNA Assay probe 20×	1.00	36.00
Total	11.00	396.00

- control well, add 5 μl nuclease-free water. On addition, mix the well contents by gently pipetting. Seal the wells appropriately.
4. Briefly centrifuge at 1000 × *g* before incubating on ice for 5 min.
 5. Place in a thermal cycler and run the program detailed in Table 2.
 6. Dilute the resulting cDNA product by adding 86.5 μl of nuclease-free water to the 15 μl of reaction product. Mix gently by pipetting (*see* **Note 7**).
 7. The resulting solution contains diluted cDNA which may be used immediately or stored at −20 °C for subsequent use.
 8. The dilution performed here means that the diluted cDNA is ‘PCR-ready’ and no further nuclease-free water needs to be added at the quantitative real-time PCR stage (*see* Subheading 3.3.2). The immediate dilution step of the cDNA product aims to minimise subsequent technical variance, as larger volumes will be pipetted.

3.3.2 Quantitative
Real-Time PCR
for Individual MicroRNAs

Taqman MicroRNA Assay probes 20× should be protected from the light.

1. Create a separate RT-qPCR mastermix for each target miRNA. Prepare the RT-qPCR mastermixes in 2 ml RNase-free tubes according to the volumes stated in Table 7. The volume prepared should be sufficient to test the samples and no-template control in triplicate, plus 10% volume to allow for potential pipetting error. Add the reagents in the order listed in the table and mix by gently pipetting.
2. Transfer 11 μ l aliquots of each miRNA specific RT-qPCR mastermix into the appropriate number of wells of a white 96-well qPCR plate, i.e., enough to test each sample and the nuclease-free water in triplicate.
3. Add 9 μ l of the diluted cDNA sample [from Subheading 3.3.1 (quality control) or Subheading 3.4.3 (multiplex)] to each corresponding miRNA specific RT-qPCR mastermix well. On addition, mix the well contents by gently pipetting.
4. Seal the plate with an optical qPCR seal.
5. Centrifuge briefly at $1000 \times g$ (*see Note 8*).
6. Place the plate in a RT-qPCR thermal cycler. Ensure the cycler is set to FAM fluorophore detection (520 nm wavelength) and background correction for the plate has been applied, then run the RT-qPCR programme detailed in Table 5.

3.3.3 Quality Control Data Analysis

1. For each sample, calculate the mean raw C_q value for each miRNA from the triplicate raw C_q values.
2. Calculate the hemolysis score (ΔC_q^{Haem}) for each sample using the equation below:

$$\Delta C_q^{\text{Haem}} = \text{mean C}_q \text{ hsa-miR-23a-3p} - \text{mean C}_q \text{ hsa-miR-451a}$$
3. Use the following criteria to evaluate sample quality:
 - If the ΔC_q^{Haem} is >8 , it is considered significantly hemolyzed. If possible, obtain a new serum sample for analysis.
 - If the cel-miR-39-3p mean raw C_q is >25 , the RNA extraction step should be repeated, if possible.
 - If the hsa-miR-30b-5p mean raw C_q is >30 , there is a potential problem with sample integrity (*see Note 9*). Caution should be applied when interpreting results from these samples. If possible, repeat the RNA extraction or obtain a fresh serum sample.

3.4 Custom MicroRNA Panel Multiplex

Prior to starting, ensure that all materials for the multiplex RT step are prepared (*see* Subheading 2.4). Reagents and samples should be defrosted and handled on ice.

3.4.1 Custom MicroRNA Panel Multiplex Reverse Transcription Step

1. Prepare one multiplex RT-mastermix in a 2 ml nuclease-free tube according to the volumes stated in Table 3. The volume prepared should be sufficient to process all samples and a no-template control sample, plus 10% volume to allow for

Table 8
Reaction volumes for multiplex pre-amplification mastermix

Reagent	Volume for 1 reaction (μl)	e.g., Volume for 10 reactions (μl)
Taqman Pre-amp Master Mix 2×	25.0	250.0
Custom multiplex probe mix (Subheading 2.5; step 2)	12.5	125.0
Total	37.5	375.0

potential pipetting error. Add the reagents in the order listed in Table 3 and mix by gently pipetting.

2. Transfer 10 μl aliquots of the multiplex RT-mastermix into 0.2 ml PCR wells. There should be one well per sample, including the no-template control.
3. Gently mix the RNA eluate. Transfer 5 μl of each RNA sample to an individual multiplex RT mastermix well. For the no-template control well, transfer 5 μl nuclease-free water. Seal the wells appropriately.
4. Briefly centrifuge the wells at 1000 × *g* before incubating on ice for 5 min.
5. Place in a thermal cycler and run the programme detailed in Table 2. The resulting solution contains multiplex cDNA.

**3.4.2 Multiplex
Pre-amplification Step**

Prior to starting, ensure that all materials for the multiplex pre-amp step are prepared (*see* Subheading 2.5).

1. Prepare the pre-amp mastermix in a nuclease-free tube according to the volumes stated in Table 8. The volume prepared should be sufficient to process all samples and a no-template control sample, plus 10% volume to allow for pipetting error. Mix the pre-amp mastermix by gently pipetting.
2. Transfer 37.5 μl aliquots of pre-amp mastermix into 0.2 ml PCR wells.
3. Add 12.5 μl of multiplex cDNA from each sample to an individual pre-amp mastermix well and mix the contents by gently pipetting. Seal the wells appropriately.
4. Briefly centrifuge the wells at 1000 × *g* before placing in a thermal cycler and running the program detailed in Table 4. The resulting solution contains amplified cDNA derived from the copies of sequence of the miRNAs of interest.
5. In separate 2 ml RNase-free tubes, dilute the entire 50 μl volume of each individual cDNA sample with 200 μl of nuclease-free water, then pipette gently to mix. Diluted cDNA may be used immediately or stored at −20 °C for subsequent use.

3.4.3 RT-qPCR for Individual MicroRNAs

Using the RT-qPCR protocol described in Subheading 3.3.2, separately quantify the levels of each single miRNA using the diluted multiplex cDNA from the pre-amplification step (Subheading 3.4.2). Note that despite the preceding (reverse transcription and pre-amp) steps being performed in multiplex, this step must be performed in singleplex.

3.4.4 Data Analysis— $\Delta\Delta C_t$ Method and Normalization

1. Calculate the RNA extraction correction-factor for each sample. To do this, first calculate the raw mean cel-miR-39-3p Cq value for each sample. Next, identify the sample with the lowest raw mean cel-miR-39-3p Cq value. Subtract this value from the raw mean cel-miR-39-3p Cq value of each sample to obtain a correction factor for each sample.

$$c = a - b.$$

Where:

a = Sample raw mean cel-miR-39-3p Cq.

b = Lowest raw mean cel-miR-39-3p Cq.

c = RNA extraction correction-factor

2. Correct the levels of housekeeping miRNAs for RNA extraction efficiency. To do this, first calculate the raw mean housekeeping Cq values. Next, subtract the RNA extraction correction-factor (c , above) from the raw mean Cq value of each housekeeping miRNA for each sample.

$$d = e - c.$$

Where:

c = RNA extraction correction-factor (above).

d = Corrected housekeeping miRNA Cq.

e = Raw mean housekeeping miRNA Cq.

3. To calculate biomarker expression relative to an endogenous housekeeping miRNA, first calculate the raw mean biomarker Cq values, then apply the following equation (for serum samples, we suggest using hsa-miR-30b-5p, hsa-miR-30c-5p, and hsa-miR-191-5p as housekeeping miRNAs):

$$f = g - d.$$

Where:

d = Corrected housekeeping miRNA Cq (above).

f = Housekeeping normalized biomarker miRNA ΔCq .

g = Raw mean biomarker miRNA Cq

4. To compare levels of biomarker miRNA in a sample with those from a control patient group, calculate the biomarker miRNA $\Delta\Delta Cq$. First, the biomarker levels in each group must be adjusted for levels of cel-miR-39-3p and the housekeeping miRNA(s), as in steps 2 and 3. Next, calculate the mean housekeeping normalized biomarker miRNA ΔCq of the control group. Then, for each sample, subtract the housekeeping normalized biomarker miRNA ΔCq from the mean control group value.

$$h = f_{\text{control}} - f_{\text{sample}}$$

Where:

h = $\Delta\Delta Cq$ of control and sample biomarker miRNA.

f_{control} = mean housekeeping normalized biomarker miRNA ΔCq of the control group.

f_{sample} = housekeeping normalized biomarker miRNA ΔCq value of the sample.

Then, to find sample relative expression:

sample relative expression = $\log_2 b$.

4 Notes

1. cel-miR-39-3p stock solution (2×10^{10} copies) may be aliquoted and stored at -80°C ; however, the dilute cel-miR-39-3p working solution should not be stored for further use due to risk of degradation at low concentrations. Instead, a fresh solution should be prepared for each RNA extraction batch.
2. We recommend including cel-miR-39-3p and endogenous miRNA housekeeping gene(s), e.g., for serum we recommend hsa-miR-30b-5p.
3. We recommend preparing enough multiplex primer (Subheading 2.4) and probe (Subheading 2.5) mix for the entire sample cohort to minimize batch variation. Unused custom multiplex primer and probe mixes may be aliquoted and stored at -20°C for future use.
4. The protocol can be performed with serum volumes of less than 200 μl . For these samples, the total volume should be made up to 200 μl with sterile $1\times$ PBS before proceeding with the RNA extraction protocol. Record any samples treated in this way and the volume of $1\times$ PBS added in case of any subsequent issues, e.g., with QC.
5. During the RNA extraction step, if the middle protein layer is substantial, recovery of the top RNA containing aqueous phase layer may be impeded. Remove as much of the aqueous phase as possible without disturbing the middle protein layer, then recentrifuge the sample at $12,000 \times g$ for a further 5 min at 4°C . This step can be repeated until at least 500 μl of the top aqueous phase layer is collected in total.
6. It is not possible to accurately quantify the concentration of circulating RNA eluted due to the use of MS2 bacteriophage RNA as a carrier in the RNA extraction process. As such, the amount of RNA entering the next steps is best standardized by using the same volume of eluate, followed by the robust normalization described in this protocol.
7. If the cDNA is to be stored for an extended period of time, dilution of the cDNA product may be performed at the point of use.
8. The qPCR plate may be stored at -20°C until later use. On running, the contents should be defrosted on ice, protected from light. The plate should be centrifuged again to $1000 \times g$ before running the RT-qPCR.

9. The raw Cq hsa-miR-30b-5p cut-off of 30 is not applicable to samples with a starting volume of less than 200 μl . A correction cut-off Cq value based on starting volume can be obtained using the following equation:

$$x = \log_2 \frac{200 \mu\text{l}}{\text{starting volume } (\mu\text{l})}$$

Where x is the number of Cq to be added to the existing cut-off of 30, e.g., for a starting volume of 50 μl , x is calculated as 2, and the corrected Cq cut-off is 32.

That notwithstanding, caution should be applied when analyzing data from samples with a corrected hsa-miR-30b-5p Cq of >30 , as use of more dilute serum may increase the likelihood of false-negative results when testing for biomarker miRNA levels.

References

- Witwer KW (2015) Circulating microRNA biomarker studies: pitfalls and potential solutions. *Clin Chem* 61:56–63
- Murray MJ et al (2016) A pipeline to quantify serum and cerebrospinal fluid microRNAs for diagnosis and detection of relapse in paediatric malignant germ-cell tumours. *Br J Cancer* 114:151–162
- Blondal T et al (2013) Assessing sample and miRNA profile quality in serum and plasma or other biofluids. *Methods* 59:S1–S6
- Pritchard CC et al (2012) Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. *Cancer Prev Res* 5:492–497
- Kirschner MB et al (2011) Haemolysis during sample preparation alters microRNA content of plasma. *PLoS One* 6:e24145
- McDonald JS et al (2011) Analysis of circulating microRNA: preanalytical and analytical challenges. *Clin Chem* 57:833–840
- Kroh EM et al (2010) Analysis of circulating microRNA biomarkers in plasma and serum using quantitative reverse transcription-PCR (qRT-PCR). *Methods* 50:298–301
- Murray MJ et al (2014) Serum levels of mature microRNAs in DICER1-mutated pleuropulmonary blastoma. *Oncogene* 3:e87
- Murray MJ et al (2015) Solid tumors of childhood display specific serum microRNA profiles. *Cancer Epidemiol Biomarkers Prev* 24:350–360
- Mitchell PS et al (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A* 105:10513–10518
- Chen C et al (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* 33:e179
- Tang F et al (2006) 220-plex microRNA expression profile of a single cell. *Nat Protoc* 1:1154–1159
- Lao K et al (2006) Multiplexing RT-PCR for the detection of multiple miRNA species in small samples. *Biochem Biophys Res Commun* 343:85–89
- Murray MJ et al (2011) Identification of MicroRNAs from the miR-371–373 and miR-302 clusters as potential serum biomarkers of malignant germ cell tumors. *Am J Clin Pathol* 135:119–125
- Dieckmann KP et al (2016) MicroRNA miR-371a-3p – A novel serum biomarker of testicular germ cell tumors: evidence for specificity from measurements in testicular vein blood and in neoplastic hydrocele fluid. *Urol Int* 97:76–83
- Wulfken LM et al (2011) MicroRNAs in renal cell carcinoma: diagnostic implications of serum miR-1233 levels. *PLoS One* 6:e25787
- Meng X et al (2015) Diagnostic and prognostic potential of serum miR-7, miR-16, miR-25, miR-93, miR-182, miR-376a and miR-429 in ovarian cancer patients. *Br J Cancer* 113:1358–1366
- Marabita F et al (2016) Normalization of circulating microRNA expression data obtained by quantitative real-time RT-PCR. *Brief Bioinform* 17:204–212
- Schwarzenbach H et al (2015) Data normalization strategies for MicroRNA quantification. *Clin Chem* 61:1333–1342

Chapter 19

MicroRNAs, Regulatory Networks, and Comorbidities: Decoding Complex Systems

Francesco Russo, Kirstine Belling, Anders Boeck Jensen,
Flavia Scoyni, Søren Brunak, and Marco Pellegrini

Abstract

MicroRNAs (miRNAs) are small noncoding RNAs involved in the posttranscriptional regulation of messenger RNAs (mRNAs). Each miRNA targets a specific set of mRNAs. Upon binding the miRNA inhibits mRNA translation or facilitate mRNA degradation. miRNAs are frequently deregulated in several pathologies including cancer and cardiovascular diseases. Since miRNAs have a crucial role in fine-tuning the expression of their targets, they have been proposed as biomarkers of disease progression and prognostication.

In this chapter we discuss different approaches for computational predictions of miRNA targets based on sequence complementarity and integration of expression data. In the last section of the chapter we discuss new opportunities in the study of miRNA regulatory networks in the context of temporal disease progression and comorbidities.

Key words microRNA, Regulatory network, Target prediction, Comorbidities, Systems biology, Data integration

1 Introduction

Gene expression regulation has increased in complexity since the recent discovery of noncoding RNAs (ncRNAs) [1, 2]. microRNAs (miRNAs) were identified as new, small ncRNAs with conserved sequences and functions involved in posttranscriptional regulation. miRNAs are widely distributed in animals, plants, and viruses and to date 2588 mature human miRNA transcripts are listed in miRBase v21 (<http://www.mirbase.org/>) [3]. By targeting specific mRNAs (usually in the 3' untranslated region (UTR)), miRNAs lead to translational repression or promote the degradation of their target mRNAs [4, 5]. miRNA genes are under strong selective pressure to preserve the 5' end of mature miRNAs as positions 2 to 7 are crucial for target recognition, termed the “seed” region. The further downstream nucleotides

(particularly nucleotide 8 and 13–16) also contribute to base pairing of miRNAs with their targets [5]. Further evidence suggests that miRNAs also can affect gene expression at translational levels by binding 5' UTR and coding regions of target mRNAs [6, 7] and under specific conditions induce translational upregulation of targets [8]. More than 60% of human protein-coding genes contain at least one conserved miRNA-binding site. A huge number of non-conserved sites also exist [9].

miRNAs have a strong impact on cellular functions (e.g., cell death and proliferation). Therefore, their biogenesis is carefully regulated at transcriptional and posttranscriptional levels [10] and their dysregulation is often associated with human diseases, including cancer [11].

It has also been proposed that miRNAs play a role in development of comorbidities (the presence of two diseases in a patient). miRNA–mRNA interactions can be the underlying causal effects for cases where one disease induces the development of another disease. Examples of such disease pairs are neurological and cardiovascular disorders [12], obesity and metabolic syndrome [13], and complex regional pain syndrome and headache [14]. Comorbidities have been studied in the context of disease associations for many years in epidemiologically research and recently more sophisticated data-driven approaches have linked molecular etiology to disease associations calculated from electronic patient records [15, 16]. The hypothesis of these studies is that strongly associated diseases have a high likelihood of sharing genetic etiology.

Evaluation of miRNA and mRNA expression profiles in blood and tissue samples from patients with comorbidities is of great importance for future studies of comorbidities. Data integration approaches with target prediction and miRNA–mRNA expression correlation analyses will lead to increased knowledge of how miRNA influence comorbidity development. Understanding the function of miRNAs in a single disease is challenging and the complexity is even higher when focus is expanded to comorbidities. Proper experimental design including computational and systems biology approaches can help to decode this complex scenario.

The identification of miRNA targets is a key step to understand the function of miRNAs themselves and for their potential role in diseases. In Fig. 1 we report main variables to take into consideration for a correct identification and interpretation of miRNA targets. The majority of miRNA targets are predicted computationally and several algorithms exist for carrying out the predictions [17–23]. They are based on different criteria using the information of sequence complementarity of miRNA to target. However, only few interactions have in fact been experimentally validated. Other approaches integrate *in silico* target prediction and use miRNA and mRNA expression data to reconstruct posttranscriptional regulatory networks [24, 30]. Such approaches increase the possibility of

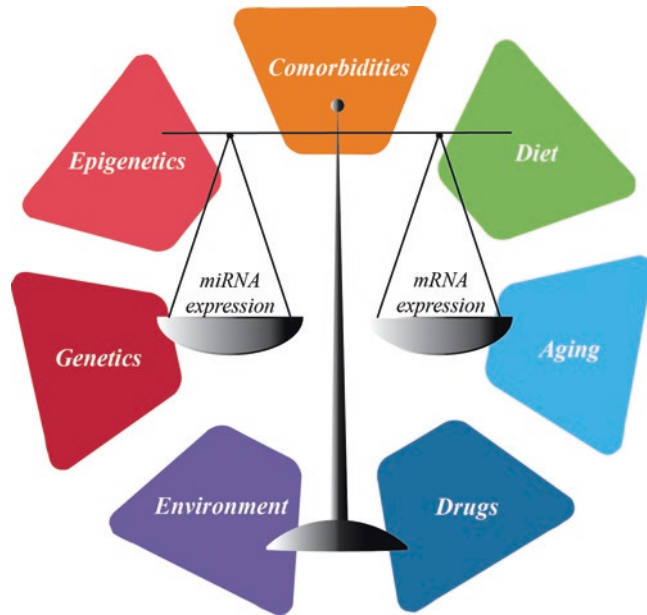


Fig. 1 Main variables to take into consideration for correct identification and interpretation of miRNA targets. Each factor can influence miRNA and mRNA expression and abundance

discovering true positive miRNA–mRNA interactions reducing the number of false positives, which otherwise is a huge issue when solely using sequence complementarity.

In this chapter we discuss the *in silico* prediction of miRNA–mRNA interactions using sequence complementarity methods and afterwards focus on the more robust data integration approaches. In the last part of the chapter we discuss how to experimentally design and analyze and integrate miRNA and mRNA data in the complex scenario of disease comorbidities.

2 Materials

The majority of the tools described in this chapter are publicly available online or as R packages. R is a free software environment for statistical computing and graphics (<https://www.r-project.org/>).

3 Methods

3.1 Sequence Complementarity Based Methods

In the last two decades many computational methods have been developed to identify miRNA–target interactions. In the early years of this research area, algorithms were based on sequence complementarity using target recognition rules retrieved from published studies reporting experimentally validated interactions.

These approaches are still among the most popular algorithms nowadays and very useful for novel approaches that have been developed more recently. In the next subsections we briefly discuss the most popular algorithms used in this field based on sequence complementarity. We have selected methods that use different approaches to give the readers an overview.

3.1.1 TargetScan

One of the first computational target prediction algorithm was TargetScan, a rule-based method proposed in 2003 [17, 18]. TargetScan predicts biological targets of miRNAs by searching for the presence of canonical sites (8mer, 7mer, and 6mer sites) that match the seed region of each miRNA. “Canonical sites” are regions in the mRNA that contain an exact miRNA seed match. Some canonical sites are more effective at mRNA control than others. One option of the tool is to consider only conserved sites across species.

TargetScan is available for the species *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, and *D. rerio*. TargetScanHuman considers matches to human 3' UTRs and their orthologues, as defined by UCSC whole-genome alignments (<http://genome.ucsc.edu>). Recently, a new version of TargetScan has been published (version 7.0) [18] that includes an improved quantitative model of canonical targeting using a compendium of experimental datasets. The proposed novel model not only considers the miRNA binding site type but also other 14 features to predict the most effectively targeted mRNAs using multiple linear regression models. The resulting models, one for each binding site type, were collectively called the *context++ model*. The considered features included characteristics of miRNAs, features of the sites (including their contexts and positions within the mRNAs), and features of the mRNAs. Furthermore, authors showed that TargetScan performed significantly better than existing models and was as informative as the best high-throughput in vivo cross-linking approaches such as HITS-CLIP [31] and CLASH [32]. TargetScan is available online at http://www.targetscan.org/vert_70/ as precompiled predictions or as perl script.

3.1.2 miRanda-mirSVR

The miRanda algorithm computes optimal sequence complementarity between a set of mature miRNAs and a given mRNA using a *weighted dynamic programming algorithm* [19]. Weights are position-dependent and reflect the relative importance of the 5' and 3' regions of miRNAs. In addition, miRanda uses an estimate of the free energy of miRNA–mRNA duplex as a filter.

Sequence conservation at and near miRNA binding sites is a strong indication of functional constraints in evolution. miRanda filters out less-conserved predicted target sites using the *PhastCons* conservation score. This score measures the evolutionary conservation of sequence blocks across multiple vertebrates using a

phylogenetic *Hidden Markov Model* [33]. To improve the prediction, authors proposed another algorithm called mirSVR [20] for scoring and ranking the efficiency of miRanda-predicted miRNA target sites. This approach uses a *Support Vector Regression (SVR)* to train on a wide range of features, including secondary structure accessibility of the site and conservation. mirSVR is also able to identify a significant number of experimentally determined noncanonical and nonconserved sites. MiRanda is available online at <http://www.microrna.org/microrna/home.do>. It is also possible to use the script of the algorithm that has been written in C.

3.1.3 PITA

The Probability of Interaction by Target Accessibility (PITA) algorithm is based on the evidence that target accessibility has a critical role in miRNA–mRNA interactions for a wide range of target types, and PITA is able to accurately capture these effects. It uses initial seeds for each miRNA in 3' UTRs and then applies a model to each putative site [21]. It computes an energy-based score for miRNA–mRNA interactions, $\Delta\Delta G$, equal to the difference between the free energy gained by the binding of the miRNA to the target, ΔG_{duplex} , and the free energy lost by unpairing the target-site nucleotides, ΔG_{open} . Further, the method combines sites for the same miRNA to obtain a total interaction score for the miRNA and UTR.

PITA has been applied to all 3' UTRs of fly, worm, mouse, and human, resulting in catalogues of target predictions for these organisms that are available online at http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html. Through a web interface the user is able to upload both miRNA and mRNA sequences or to browse pre-calculated interactions. Moreover, it is possible to download the script version of the program for large scale predictions.

3.1.4 RNA22

RNA22 is a pattern-based approach for discovery of miRNA binding sites and it can be applied to any genome without requiring genome-specific model retraining [22]. It does not use a cross-species sequence conservation filter allowing the discovery of nonconserved miRNA binding sites. Instead the method is based on the *Teiresias* algorithm [34] that is used to discover variable-length motifs (“patterns”) in mature miRNA sequences followed by a *second-order Markov chain* to estimate statistical significance of each pattern. After the pattern discovery step follows a step of “target islands” identification. Target islands refer to any regions of the UTR that is reverse complement of the mature miRNA. The Vienna package [35] is used to predict the structure of each miRNA/island-segment duplex and its Gibbs free energy (“folding energy”). RNA22 recognizes a binding site based on the presence of multiple, distinct, statistically significant patterns that have been discovered by processing known mature miRNA sequences.

RNA22 is available online at <https://cm.jefferson.edu/rna22/> and it allows uploading user defined miRNA and mRNA sequences and setting several parameters, e.g., the folding energy cutoff. Moreover, the user is also able to download a program that runs locally allowing to submit batch queries to the RNA22 server. Precomputed predictions are also available for fly, worm, mouse, and human.

3.1.5 RNAhybrid

RNAhybrid predicts multiple potential miRNA binding sites [23]. The program finds the energetically most favorable hybridizations of small RNAs to large RNAs. The tool computes an optimization of miRNA–mRNA duplexes by an accurate statistical analysis of *Minimum Free Energies (MFEs)*. Then, it normalizes MFEs by sequence lengths of miRNA and targets and models the normalized MFEs as distributions. A second program, RNACalibrate [23], evaluates the predictions for every miRNA and assigns p-values to normalized MFEs. The significance of multiple binding sites in a single target is evaluated with Poisson statistics.

The web page (<http://bibiserv.techfak.uni-bielefeld.de/rna-hybrid/>) allows users to upload miRNA and mRNA sequences. It is also possible to download the RNAhybrid binary package for Windows, src package for almost every Unix like system and binary package for OSX.

3.2 Data Integration Approaches

All available algorithms based on sequence complementarity produce a significant number of false positives. This is mainly caused by the current lack of knowledge of miRNA–mRNA interaction and the miRNA action itself. Integration of target predictions with gene expression profiles improves the detection of true functional interactions. Several novel methods integrate gene expression assuming that miRNAs downregulate target mRNAs, thus, looking for anti-correlation either upregulation of miRNA and down-regulation of target or vice versa. In this section we discuss methods that make use of expression correlation, i.e., the strength and direction of a relationship between two random variables such as *Pearson correlation* (parametric) and *Spearman correlation* (non-parametric). Other methods use *Mutual Information*, a measure of mutual dependence of two variables. Mutual information is sensitive to any functional relationship and not just to linear dependencies. Finally, recent methods try to infer miRNA–mRNA *causal regulatory relationships* starting from observational data (e.g., gene expression profiles). The aim of these methods is to discover causal effects of miRNAs on mRNAs. In fact, it is important to highlight that *Correlation* or *Association* are not *Causality* for an observed event [36]. This key concept has been recently underlined by Altman and Krzywinski [36]. Correlation implies association, but not causation. Conversely, causation implies association, but not correlation.

3.2.1 *Magia*

MiRNA and genes integrated analysis (MAGIA) is a web tool for integrative analysis of target predictions, miRNA and mRNA expression data [24, 25]. The tool allows several kinds of analyses: (a) retrieve and browse miRNA target predictions for drosophila, mouse, rat, and human using target prediction algorithms discussed above such as TargetScan and PITA with the possibility of combining them with Boolean operators; (b) integrate gene expression data and use different functional measures: parametric and nonparametric correlation indexes, a variational Bayesian model, mutual information, and a meta-analysis approach based on *P*-value combination. The results of the analyses are bipartite regulatory networks of the best miRNA–mRNA interactions. Moreover, in the second version of the tool [25] authors have introduced integration of transcription factor (TF) binding sites giving the possibility to discover new regulatory circuits of miRNA–mRNA–TF.

MAGIA allows the user to follow several steps in the analysis in a clear way. The aim of the tool is to refine target predictions using gene expression data. First, the user can select the organism of interest and the gene or transcript annotation (EntrezGene, RefSeq or ENSEMBL annotations). Second, the user can select target prediction algorithms, either single or multiple algorithms, to find intersection of predictions. Finally, the user can upload normalized expression profiles for mRNA and miRNA.

A powerful aspect of MAGIA is that it takes into account whether data has been collected from experiments by measuring expression profiles of miRNAs and targets in exactly the same biological samples. The tool allows use of a meta-analysis approach based on a *P*-value combination in the case of different biological samples or Spearman and Pearson correlation, mutual information, and a variational Bayesian model if expression profiles come from the same biological samples.

MAGIA is available online at <http://gencomp.bio.unipd.it/magia2/start/>.

3.2.2 *mirConnX*

mirConnX combines sequence information with gene expression data to create a disease-specific, genome-wide regulatory network [26]. The aim of this tool is to provide a unique environment for transcriptional and posttranscriptional regulation considering miRNA–mRNA–TF regulatory networks integrating mRNA and miRNA expression data measured under the same set of conditions. Expression profiles are preprocessed to remove lowly expressed miRNAs and mRNAs followed by the construction of an association network using statistical approaches including Pearson, Spearman and Kendall correlations. This undirected network is then compared to a prior pre-compiled, species-specific static network, which is derived from TF–gene binding, miRNA–mRNA prediction and literature evidence. The result is a directed, weighted graph. An integration function allows obtaining a final directed network containing robust interactions.

mirConnX web tool allows the visualization and exploration of the network and identifies network motifs. The tool is available at <http://www.benoslab.pitt.edu/mirconnx/>.

3.2.3 GenMir++

The Generative model for miRNA regulation (GenMir++) [27] is a Bayesian model and learning algorithm that includes gene expression data and a set of candidate miRNA targets obtained from target prediction algorithms. The model assumes that downregulation of target mRNA is related to the action of multiple miRNAs. High expression of one or many miRNAs causes a modification in the expression of mRNA that is negatively shifted with respect to an estimated background expression level. For any transcript, regulating miRNAs will be selected using a set of unobserved binary indicator variables. Given the expression data and the predicted interactions, the problem of finding functional miRNA targets consists of inferring which indicator variables are turned on and which are turned off.

GenMir++ is available as MATLAB code at <http://www.psi.toronto.edu/genmir/>.

3.2.4 miRNet

Recently, a new web-based tool called miRNet has been published [28]. It offers statistical, visual, and network-based approaches to study miRNA regulatory networks. miRNet includes a comprehensive knowledge base integrating miRNA–target interaction data from 11 databases: miRTarBase [37], TarBase [38], miRecords [39], SM2miR [40], Pharmaco-miR [41], miR2Disease [42], PhenomiR [43], StarBase [44], EpimiR [45], miRDB [46], and miRanda [19, 20]. Moreover, it allows differential expression analysis of data from microarray, RNA sequencing, quantitative PCR as well as is a powerful network visualization system and provides enrichment analysis. miRNet currently supports eight species: *S. mansoni*, *C. elegans*, *D. rerio*, *D. melanogaster*, *M. musculus*, *R. norvegicus*, *B. taurus*, and *H. sapiens*.

To facilitate the analysis, authors have implemented an interactive flowchart to allow users to choose the appropriate setup based on input data. The input can be a list of miRNA, mRNA, long ncRNA IDs or a data table containing expression values. It is also possible to select miRNA, disease, or gene names from a list of available database entries. miRNet supports differential expression analysis using limma [47], edgeR [48] and HTqPCR methods. For functional enrichment analysis miRNet use standard hypergeometric test and unbiased random sampling.

miRNet is available online at <http://www.mirnet.ca/>.

3.2.5 miRLab

miRLAB is an R package to discover miRNA–mRNA regulatory relationships using several tools including popular target prediction and ensemble methods [29]. miRLAB contains both miRNA and mRNA expression datasets, a pipeline to retrieve datasets from

The Cancer Genome Atlas (TCGA, <http://cancergenome.nih.gov/>), and computes differential expression. Users can also provide their own datasets and compare the prediction results with those predicted by the embedded benchmark methods.

The authors have re-implemented several computational methods including correlation, regression, and causal inference approaches. The correlation methods included are: Pearson, Spearman, Kendall, distance correlation, Hoeffding's D measure, and randomized dependence coefficient. In this context miRNA–mRNA pairs are ranked based on the correlation coefficient where negative correlations are in the top ranking. The authors also included Mutual Information for nonlinear relationship discovery. Other commonly used methods are based on regression models of which miRLAB includes the high-dimensional techniques Lasso [49] and Elastic-net [50], which both can be used to infer associations between variables. Another interesting method implemented in miRLAB is a causal inference algorithm called *Intervention calculus when the Directed Acyclic Graph is Absent (IDA)* [51, 52], which estimates the causal effect that a variable has on others. miRNAs are considered causes while mRNAs are considered effects. IDA follows two main steps consisting of learning a causal structure from observational data, then inferring the causal effects and scoring them. miRLAB is one of the few tools that integrate this novel approach, but it is also computationally time consuming. Future validation is needed in order to understand whether causal based methods perform better than standard correlation in the context of miRNA–mRNA prediction, but it is considered a growing and promising field [53].

miRLAB further provides an option to incorporate the target information from different sequence based prediction algorithms (e.g., TargetScan) obtaining strong miRNA–mRNA relationships based on expression data and physical interactions.

miRLAB is available at <http://bioconductor.org/packages/release/bioc/html/miRLAB.html>.

3.3 miRNAs and Comorbidities

Given the increased evidence of the role of miRNAs in diseases and biological processes, it is likely that miRNAs are involved in driving the co-occurrence of diseases (i.e., comorbidities) [12–14]. miRNAs are expressed across tissues and can affect molecular players in several pathways, and thus, miRNA deregulation might cause co-occurrence of two or more diseases. Therefore, including miRNA regulation in comorbidity analysis can uncover new knowledge about disease associations. In this context many variables such as environmental and genetic factors also play a role. These are important factors to take into consideration in the experimental design of gene expression measurements and later downstream analyses. In fact, each factor could change miRNA and mRNA expressions (Fig. 1) resulting in misinterpretation of biological results.

Recently, several studies have explored disease association in a data-driven manner where the association between all diseases are calculated based on electronic health records (EHR) or registry data [16, 54, 55]. These studies have demonstrated that genetic factors can be derived from EHR data analysis [55] and that significantly associated diseases share genetic etiology than random pairs of diseases [54, 56]. A crucial factor to be considered in such data-driven studies of disease associations is the temporal order that the diagnoses develops in. Recently, a data-driven approach of temporal disease progression has been proposed using EHR data covering the whole population of Denmark, considering 14.9 years of registry data on 6.2 million patients [16]. This approach mapped how patients progress through different diseases. Such maps can be utilized to identify groups of patients with different disease development that might be explained by genetic factors such as miRNAs. These types of approaches open up for new opportunities in the discovery of molecular biomarkers. If coupled with expression data, they can serve as a powerful tool for uncovering novel association between miRNA and disease. miRNAs have been proposed to be novel candidate biomarkers for disease progression for cancer and other pathologies [57, 58]. Moreover, it has been shown that miRNAs circulate in the human bloodstream complexed in vesicles such as exosomes, microvesicles, high-density lipoproteins (HDLs), and Ago2 protein [59–62]. Since circulating miRNAs are stable in body fluids, they can serve as fast and noninvasive novel candidate biomarkers for the early detection and progression of diseases. With these recent discoveries, it is possible to design proper experiments and computational analyses to better integrate miRNA and mRNA expressions in the context of disease comorbidities (Fig. 2). Considering a large cohort of patients with different risk factors for a specific primary disease, an ideal experimental design would consist of follow-up of patients where blood samples are taken at each time point. The blood samples can be used to extract RNAs and then quantify miRNA and mRNA expressions using Next Generation Sequencing technologies or microarrays. Following patients from the time of the first diagnosis of the primary disease through months or years can allow us to monitor the onset of secondary diseases that show strong correlation with the primary disease. At each time point, miRNAs and mRNAs are likely to have fluctuating expression levels due to risk factors, the presence of secondary disease or because the age is an important parameter to take into account (Fig. 1). The data integration methods described above should be considered in a dynamic way where crucial changes in miRNA–mRNA interactions are observed. This approach will elucidate the role of miRNAs and their targets in disease progression and comorbidities, allowing the discovery of novel molecular players.

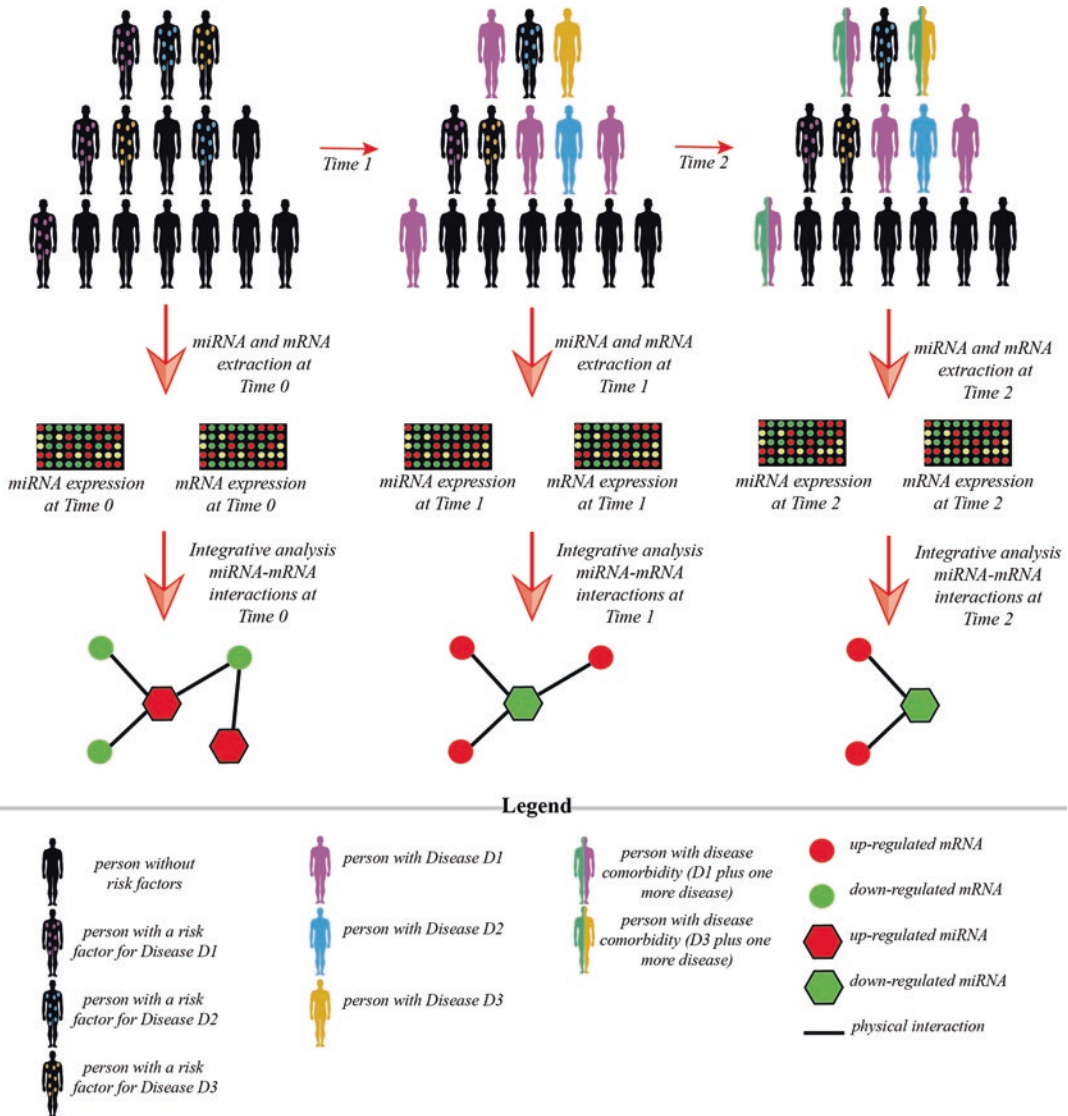


Fig. 2 Experimental and computational design to better integrate miRNA and mRNA expressions in the context of disease comorbidities. In the first step (*Time 0*) a patient cohort is included in the study. At each time point (starting from *Time 0*) blood samples are collected, following RNA extraction and quantification. Then, the downstream analyses are performed including prediction of miRNA–mRNA interactions using methods discussed in this chapter. *Colored dots* in body shapes indicate specific risk factors for diseases. Single colored body shape (e.g., *pink*) indicates one patient with one disease, while multicolored body shapes (e.g., *pink* and *green*) indicate disease comorbidities. *Green* and *red* circles indicate downregulated and upregulated targets, while *green* and *red* hexagons are downregulated and upregulated miRNAs

4 Future Perspectives

In this chapter we discuss computational target prediction approaches for miRNA–mRNA interactions. Recently, it has been shown that miRNAs can bind other molecules such as long

noncoding RNAs, circular RNAs and pseudogenes [63]. By attenuating shared miRNAs the different kinds of RNAs could cross talk and regulate each other. These RNAs are known as competing endogenous RNAs (ceRNAs) [63]. Recently, a new generation of miRNA target prediction analysis has been proposed [64, 65]. These algorithms do not only include complementary sequence comparisons and gene expression, but also consider other novel variables in the context of the ceRNA cross talk. In this complex scenario the miRNA inhibition depends on several aspects: (1) number of miRNA binding sites, (2) miRNA binding affinity, (3) unbound miRNA expression level, and (4) target expression level [64, 65]. Future methods will integrate more data such as proteins, methylation, and copy number variation, increasing our understanding of miRNA regulation.

Acknowledgments

Francesco Russo has been supported by a fellowship sponsored by Progetto Istituto Toscano Tumori-Grant 2012 Prot. A00GRT. Novo Nordisk Foundation (grant agreement NNF14CC0001).

References

1. ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tamma H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316:1484–1488
3. Kozomara A, Griffiths-Jones S (2014) miR-Base: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 42:D68–D73
4. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297
5. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
6. Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics* 28:771–776
7. Lytle JR, Yario TA, Steitz JA (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci U S A* 104:9667–9672
8. Vasudevan S, Tong Y, Steitz JA (2007) Switching from repression to activation: microRNAs can up-regulate translation. *Science* 318:1931–1934
9. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19:92–105
10. Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* 10:126–139
11. Lujambio A, Lowe SW (2012) The microcosmos of cancer. *Nature* 482:347–355
12. Hébert SS (2009) Putative role of MicroRNA-regulated pathways in comorbid neurological and cardiovascular disorders. *Cardiovasc Psychiatr Neurol* 2009:849, 519
13. Hulsmans M, Sinnaeve P, Van der Schueren B, Mathieu C, Janssens S, Holvoet P (2012) Decreased miR-181a expression in monocytes of obese patients is associated with the occurrence of metabolic syndrome and coronary artery disease. *J Clin Endocrinol Metab* 97:E1213–E1218
14. Orlova IA, Alexander GM, Qureshi RA, Sacan A, Graziano A, Barrett JE, Schwartzman RJ, Ajit SK (2011) MicroRNA modulation in

- complex regional pain syndrome. *J Transl Med* 9:195
15. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søbey K, Bredkjær S, Juul A, Werge T, Jensen LJ, Brunak S (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* 7:e1002141
 16. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, Jensen PB, Jensen LJ, Brunak S (2014) Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 5:4022
 17. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. *Cell* 115:787–798
 18. Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005
 19. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res* 36:D149–D153
 20. 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:R90
 21. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39:1278–1284
 22. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* 10:1507–1517
 23. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126:1203–1217
 24. Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C (2010) MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Res* 38:W352–W359
 25. Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C (2012) MAGIA²: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res* 40:W13–W21
 26. Huang GT, Athanassiou C, Benos PV (2011) mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res* 39:W416–W423
 27. Huang JC, Morris QD, Frey BJ (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J Comput Biol* 14:550–563
 28. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia J (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res* 44:W135–W141
 29. Le TD, Zhang J, Liu L, Liu H, Li J (2015) miRLAB: an R based dry lab for exploring miRNA-mRNA regulatory relationships. *PLoS One* 10:e0145386
 30. Baglioni M, Russo F, Geraci F, Rizzo M, Rainaldi G, Pellegrini M (2015) A new method for discovering disease-specific MiRNA-target regulatory networks. *PLoS One* 10:e0122473
 31. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479–486
 32. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent non-canonical binding. *Cell* 153:654–665
 33. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050
 34. Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 14:55–67
 35. Hofacker IL, Fontana W, Stadler P, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f Chemie* 125:167–188
 36. Altman N, Krzywinski M (2015) Points of Significance: Association, correlation and causation. *Nat Methods* 12:899–900
 37. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ, Tsai TR, Ho SY, Jian TY, Wu HY, Chen PR, Lin NC, Huang HT, Yang TL, Pai CY, Tai CS, Chen WL, Huang CY, Liu CC, Weng SL, Liao KW, Hsu WL, Huang HD (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* 44:D239–D247
 38. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, Fevgas A, Dalamagas T, Hatzigeorgiou AG (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res* 43:D153–D159

39. 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:D105–D110
40. 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* 29:409–411
41. Rukov JL, Wilentzik R, Jaffé I, Vinther J, Shomron N (2014) Pharmaco-miR: linking microRNAs and drug effects. *Brief Bioinform* 15:648–659
42. 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:D98–104
43. 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:R6
44. Li JH, Liu S, Zhou H, Qu LH, Yang JH (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 42:D92–D97
45. 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* 2014, bau023.
46. Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 14:1012–1017
47. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47
48. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140
49. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Roy Stat Soc Ser B* 1:267–288
50. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J Roy Stat Soc Ser B* 67:301–320
51. Maathuis MH, Kalisch M, Bühlmann P (2009) Estimating high-dimensional intervention effects from observational data. *Ann Stat* 37: 3133–3164
52. Maathuis MH, Colombo D, Kalisch M, Bühlmann P (2010) Predicting causal effects in large-scale systems from observational data. *Nat Methods* 7:247–248
53. Le TD, Liu L, Tsykin A, Goodall GJ, Liu B, Sun BY, Li J (2013) Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics* 29:765–771
54. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1,257,601
55. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, Melamed R, Rabadan R, Bernstam EV, Brunak S, Jensen LJ, Nicolae D, Shah NH, Grossman RL, Cox NJ, White KP, Rzhetsky A (2013) A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* 155:70–80
56. Hidalgo CA, Blumm N, Barabási AL, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* 5:e1000353
57. Botla SK, Savant S, Jandaghi P, Bauer AS, Mücke O, Moskalev EA, Neoptolemos JP, Costello E, Greenhalf W, Scarpa A, Gaida MM, Buchler MW, Strobel O, Hackert T, Giese NA, Augustin HG, Hoheisel JD (2016) Early epigenetic down-regulation of microRNA-192 expression promotes pancreatic cancer progression. *Cancer Res* 76:4149–4159
58. Mees ST, Mardin WA, Sielker S, Willscher E, Senninger N, Schleicher C, Colombo-Benkmann M, Haier J (2009) Involvement of CD40 targeting miR-224 and miR-486 on the progression of pancreatic ductal adenocarcinomas. *Ann Surg Oncol* 16:2339–2350
59. Russo F, Di Bella S, Nigita G, Macca V, Laganà A, Giugno R, Pulvirenti A, Ferro A (2012) miRandola: extracellular circulating microRNAs database. *PLoS One* 7:e47786
60. Russo F, Di Bella S, Bonnici V, Laganà A, Rainaldi G, Pellegrini M, Pulvirenti A, Giugno R, Ferro A (2014) A knowledge base for the discovery of function, diagnostic potential and drug effects on cellular and extracellular miRNAs. *BMC Genomics* 15:S4
61. Mathivanan S, Fahner CJ, Reid GE, Simpson RJ (2012) ExoCarta 2012: database of exosomal proteins, RNA and lipids. *Nucleic Acids Res* 40:D1241–D1244
62. Kalra H, Simpson RJ, Ji H, Aikawa E, Altevogt P, Askenase P, Bond VC, Borràs FE, Breakefield X, Budnik V, Buzas E, Camussi G, Clayton A, Cocucci E, Falcon-Perez JM, Gabrielsson S, Gho YS, Gupta D, Harsha HC, Hendrix A, Hill AF, Inal JM, Jenster G, Krämer-Albers EM, Lim SK, Llorente A, Lötvall J, Marcilla A, Mincheva-Nilsson L, Nazarenko I, Nieuwland

- R, Nolte-'t Hoen EN, Pandey A, Patel T, Piper MG, Pluchino S, Prasad TS, Rajendran L, Raposo G, Record M, Reid GE, Sánchez-Madrid F, Schiffelers RM, Siljander P, Stensballe A, Stoorvogel W, Taylor D, Thery C, Valadi H, van Balkom BW, Vázquez J, Vidal M, Wauben MH, Yáñez-Mó M, Zoeller M, Mathivanan S (2012) Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. *PLoS Biol* 10:e1001450
63. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146:353–358
 64. Yuan Y, Ren X, Xie Z, Wang X (2016) A quantitative understanding of microRNA-mediated competing endogenous RNA regulation. *Quantitat Biol* 4:47
 65. Yuan Y, Liu B, Xie P, Zhang MQ, Li Y, Xie Z, Wang X (2015) Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. *Proc Natl Acad Sci U S A* 112(10):3158–3163

Label-Free Direct Detection of MiRNAs with Poly-Silicon Nanowire Biosensors

Jing He, Jianjun Zhu, Bin Jiang*, and Yulan Zhao*

Abstract

The diagnostic and prognostic value of microRNAs (miRNAs) in diseases becomes promising. Owing to fast response and high sensitivity, silicon nanowire (SiNW) biosensor has been considered a potential tool for miRNAs detection. Here, we describe a booming method to detect miRNAs with poly-silicon nanowire biosensors. Standard and real miRNA samples are applied in this study. The results show a limitation of 1 fM in the detection of standard miRNA sample with our poly-nanowire devices. Meanwhile, one-base mismatched sequence could be distinguished. Furthermore, these poly-SiNW arrays can detect snRNA U6 in total RNA samples extracted from HepG2 cells with a detection limitation of 0.2 µg/mL.

Key words miRNA, SiNW

1 Introduction

MicroRNAs (miRNAs) are a class of highly conserved noncoding RNAs within 18–25 nucleotides. MiRNAs play a vital role in post-transcriptional level of gene expression through binding to 3' UTR or coding region of mRNA [1, 2]. They are involved in various physiological and pathological processes such as cell differentiation, proliferation, and apoptosis [3, 4]. Dysregulation of miRNAs is related to cancers [5], atherosclerosis [6], and other diseases. The level of miRNAs in body fluid including serum has been revealed to alter in diseases, indicating their diagnostic and prognostic value in clinic [7, 8]. However, due to their characteristics like small size, low level, and sequence similarity among members, the clinical application of miRNAs is limited so far.

Current methods for miRNAs detection are mainly divided into two groups: methods based on amplification or hybridization [9]. In general, methods such as quantitative real-time PCR and

*Joint senior authors.

microarray do not match perfectly with the requirements of clinical detection, such as easy utility, fast response, low cost, high sensitivity, and specificity. Silicon nanowire (SiNW) biosensors have advantages such as label-free detection, high sensitivity, rapid response, and good selectivity. Zhang et al. first reported SiNW biosensors could detect 1 fM miRNA directly with peptide nucleic acid (PNA) probe [10]. Dorvel et al. were able to achieve 100 fM detection level of miR-10b with using SiNW and hafnium oxide dielectrics, and claimed a theoretical limit of 1 fM [11]. Importantly, they took single-stranded DNA (ssDNA) as probe for cost-saving. Notably, miRNAs detected in these studies are standard samples.

Here, we introduce our poly-SiNW biosensors to detect both standard and real miRNA sample with ssDNA as probes. Since polysilicon is the major materials used in commercial manufacture of SiNW, our study could provide experimental evidence for poly-SiNW application for miRNA detection.

2 Materials

1. Milli-Q water.
2. ssDNA and RNA oligonucleotides (Table 1).
3. Poly-silicon nanowire biosensors surrounded with a SiO₂ layer were (provided by Shanghai Integrated Circuit Research & Development Center, China) (*see* Fig. 1).
4. 2% APTES: Add 1 mL APTES to 49 mL 95% ethanol.
5. 1.25% glutaraldehyde: Add 1.25 mL 50% glutaraldehyde to 48.75 mL H₂O.
6. 1× SSC solution: 150 mM NaCl, 15 mM sodium citrate.

Table 1
Sequences of standard sample and probe

Standard sample and probe	Sequence
Let-7b	5'-UGAGGUAGUAGGUUGUGUGGUU-3'
Let-7c	5'-UGAGGUAGUAGGUUGUAUGGUU-3'
Mismatch (MM)	5'-AUGCAUGCAUGCAUGCAUGCAA-3'
Let-7b probe	5'-AACCACACAACCTACTACCTCA-3'
snRNA U6 probe	5'-TGCTAATCTTCTCTGT-3'

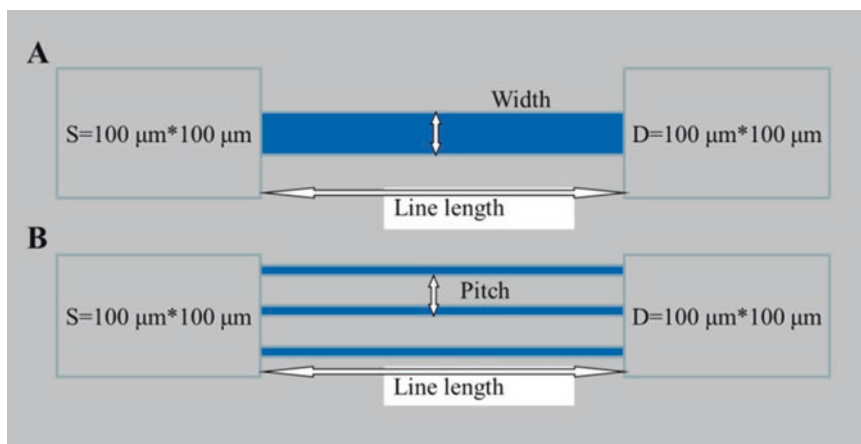


Fig. 1 Structures of SiNW biosensor. S: Source, D: Drain. SiNW length is 100 μm . Pitch is width sum of silicon nanowires and lateral blank area. **(a)** Schematic diagram of RR17, RR19, and RR20; **(b)** Schematic diagram of FF47, FF49, FF50, FF57, FF59, and FF60

3 Methods

3.1 Modification of Poly-SiNW Biosensors

1. Clean the SiNW biosensors with isopropyl alcohol and ethanol (*see Note 1*).
2. Immerse SiNW biosensors into 2% APTES solution for 2 h under 253.7 nm UV light (*see Note 2*).
3. Wash the SiNW device for three times with absolute ethanol and air-dry for 10 min.
4. Place the SiNW device in 1.25% glutaraldehyde for 1 h (*see Note 3*).
5. Wash the device for three times with Milli-Q water and air-dry for 10 min.

3.2 Probe Preparation and Hybridization

1. Incubate SiNW device with 100 nM ssDNA probe in 1 $\times$ SSC at room temperature overnight (*see Note 4*).
2. Wash the device with 1 $\times$ SSC for three times (*see Note 5*).
3. Before hybridization, measure the voltage-current curve of SiNW device and calculate the resistance (R_0) (*see Note 6*).
4. Add enough target sample in 0.01 $\times$ SSC to the device surface and incubate for 1 h (*see Note 7*).
5. Wash the device with 0.01 $\times$ SSC for three times and air-dry for 10 min.
6. Measure the voltage-current curve to calculate the resistance (R).
7. Compare R_0 and R and calculate R/R_0 (*see Note 8*).

4 Notes

1. Clean the SiNW biosensors with isopropyl alcohol and ethanol help remove contaminants.
2. Silanation is carried out by reaction with 2% APTES in 95% ethanol solution for 2 h to convert surface silanol groups to amines.
3. Treatment of 1.25% glutaraldehyde for 1 h helps create the aldehyde-modified SiNW surface so that the ssDNA probe can bind to the device (*see* Fig. 2).
4. Instead of PNA, we use ssDNA as probe which is of lower cost.
5. Wash the device with 1× SSC for three times to remove unreacted probe and air-dry for 10 min.
6. Voltage-current curves of each sample are measured in Cascade probe station (Cascade Microtech) within a certain range (0–5 V).
7. A series of standard samples of let-7b, let-7c, and mismatch miRNA sequences are prepared by adding miRNA powders in 0.01× SSC solution. Total RNA samples were extracted from cultured human liver cancer cell line HepG2 using Trizol by standard protocol.
8. Changes in resistances reflect hybridization efficiency. Data are analyzed through comparing changes in resistances. After preliminary screening, we found the current-voltage curves (I–V

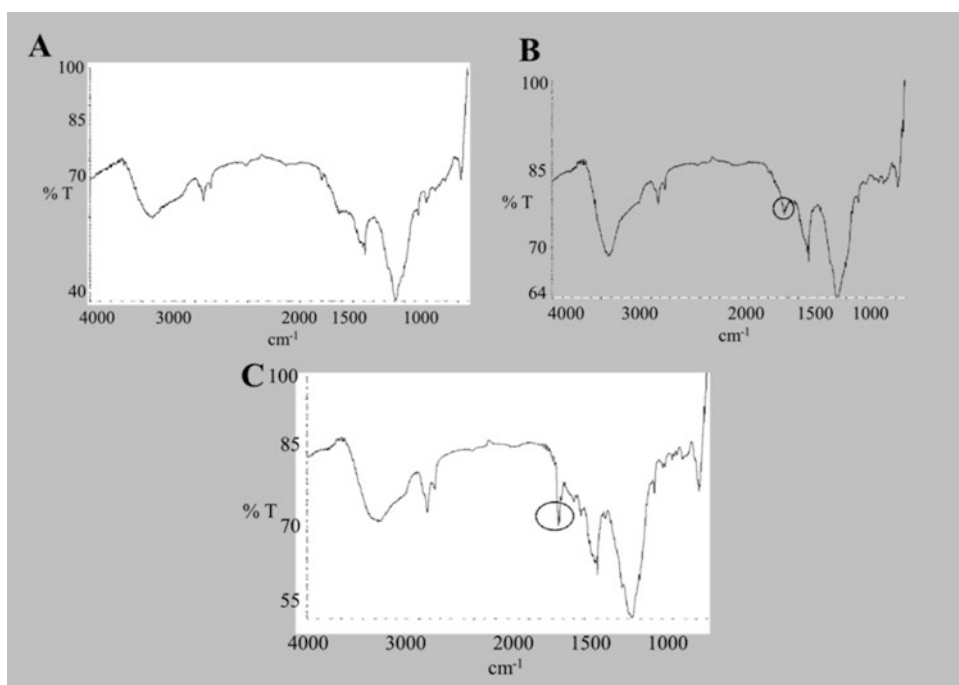


Fig. 2 Chemical modification of SiNW. (a) SiNW before chemical modification. (b) After modification of 2% APTES, bending vibration of N-H is enhanced in 1650 cm^{-1} . (c) After modification of 1.25% glutaraldehyde, stretching vibration of C-O is enhanced in 1730 cm^{-1}

curves) were commonly not straight lines, suggesting resistant of each sample was not a constant. Hence, we analyzed the resistance changed by the following process: (1) to obtain the current-voltage function (cubic function is suitable according to our experience) according to the raw curve, which is defined as $f(I) = aU^3 + bU^2 + cU + d$; (2) to obtain the conductance-voltage function by derivation $f(G) = dI/dU = 3aU^2 + 2bU + c$; (3) to calculate the area under curve (AUC) of the conductance-voltage curve in a certain range (0–5 V in this study) $\int_0^5 f(G)$ to indicate the conductance of each sample; (4) to calculate the resistance change (R/R_0) by the following equation

$$R / R_0 = \frac{\int_0^5 f(G_0)}{\int_0^5 f(G)}$$

Here, U is the voltage; I is the current; G_0 is the conductance of a biosensor before sample loading; G is the conductance of this biosensor after miRNA sample hybridization; and a , b , c , and d are constants (*see* Fig. 3).

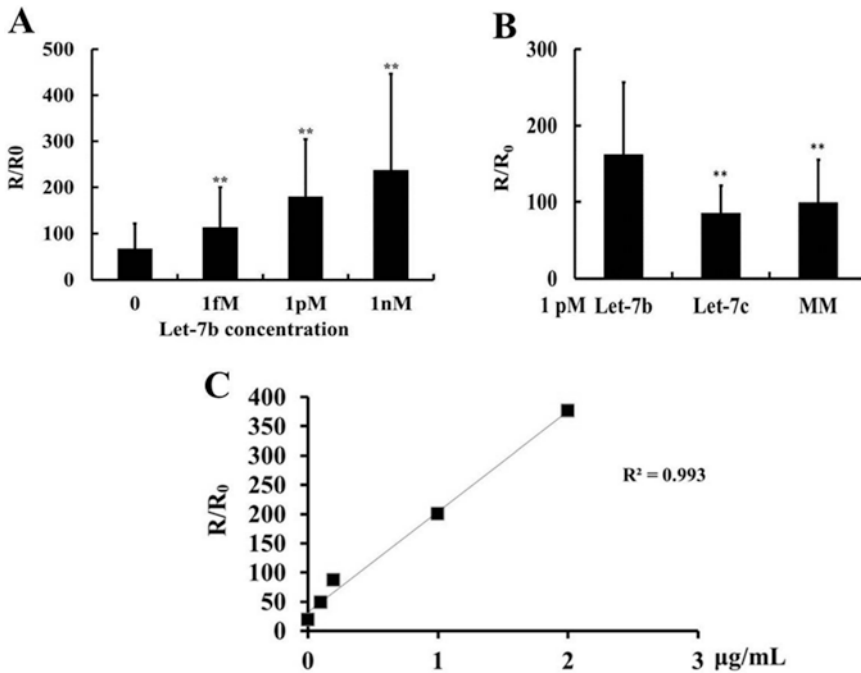


Fig. 3 (a) Detection of microRNA standard sample. Let-7b standard sample: 1 nM, 1 pM, 1 fM, and 0. Probe: 100 nM Let-7b probe. The detection limit for Let-7b standard sample is 1 fM. (b) Detection of one-base mismatched microRNA standard sample: 1 pM Let-7b, 1 pM Let-7c, and 1 pM mismatch (MM). Probe: 100 nM Let-7b probe. Single base difference between Let-7b and Let-7c is significantly identified at 1 pM concentration level. (c) Detection of snRNA U6 in total RNA from HepG2 in five concentrations. There is a well linear relationship among a series of concentrations

Acknowledgment

The analysis was supported by National Natural Science Foundation of China (NSFC 30800401 and 81201604).

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297
2. Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* 455:1124–1128
3. Guo D, Li Q, Lv Q, Wei Q, Cao S et al (2014) MiR-27a targets sFRP1 in hFOB cells to regulate proliferation, apoptosis and differentiation. *PLoS One* 9:e91354
4. Chen X, Wang X, Ruan A, Han W, Zhao Y et al (2014) miR-141 Is a Key Regulator of Renal Cell Carcinoma Proliferation and Metastasis by Controlling EphA2 Expression. *Clin Cancer Res* 20:2617–2630
5. Esquela-Kerscher A, Slack FJ (2006) Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer* 6:259–269
6. Lusis AJ (2000) Atherosclerosis. *Nature* 407:233–241
7. Brase JC, Wuttig D, Kuner R, Sultmann H (2010) Serum microRNAs as non-invasive biomarkers for cancer. *Mol Cancer* 9:306
8. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK et al (2011) MicroRNAs in body fluids--the mix of hormones and biomarkers. *Nat Rev Clin Oncol* 8:467–477
9. de Planell-Saguer M, Rodicio MC (2013) Detection methods for microRNAs in clinic practice. *Clin Biochem* 46:869–878
10. Zhang GJ, Chua JH, Chee RE, Agarwal A, Wong SM (2009) Label-free direct detection of MiRNAs with silicon nanowire biosensors. *Biosens Bioelectron* 24:2504–2508
11. Dorvel BR, Reddy B Jr, Go J, Duarte Guevara C, Salm E et al (2012) Silicon nanowires with high-k hafnium oxide dielectrics for sensitive detection of small nucleic acid oligomers. *ACS Nano* 6:6150–6164

ERRATUM TO

Chapter 8 Cell-Free Urinary MicroRNAs Expression in Small-Scale Experiments

**Ludek Zavesky, Eva Jandakova, Radovan Turyna, Daniela Duskova,
Lucie Langmeierova, Vit Weinberger, Lubos Minar, Ales Horinek,
and Milada Kohoutova**

Tamas Dalmay (ed.), *MicroRNA Detection and Target Identification: Methods and Protocols*, Methods in Molecular Biology, vol. 1580, DOI 10.1007/978-1-4939-6866-4_8, © Springer Science+Business Media LLC 2017

DOI 10.1007/978-1-4939-6866-4_21

The original version of this Chapter unfortunately contained an omission in the Acknowledgements section. The corrected Acknowledgement is given below:

The financial support from the Charles University Prague (projects PRVOUK-P27/LF1/1, Progres Q28/LF1 and PRVOUK-P25/LF1/2) and from the Ministry of Health of the Czech Republic (project FNBr 65269705) is appreciated. We would like to thank Avast Foundation, Feron, a.s., ČEPS, a.s. and CEZ Foundation for their kind support of our research.

The updated original online version for this chapter can be found at
http://dx.doi.org/10.1007/978-1-4939-6866-4_8

Tamas Dalmay (ed.), *MicroRNA Detection and Target Identification: Methods and Protocols*, Methods in Molecular Biology, vol. 1580, DOI 10.1007/978-1-4939-6866-4_21, © Springer Science+Business Media LLC 2017

INDEX

A

Absolute quantification 32, 239, 240, 244
Argonaute.....107–115, 142, 186, 188

B

Biofluids 7, 21–25, 27–29, 36–42
Bioinformatics 137–139, 194, 195
Biosensors..... 60–65, 68–70, 297–302

C

Cancer
 endometrial.....100
 ovarian 100
Cerebrospinal fluid (CSF) 41, 267
Circular strand displacement polymerization 72
Comorbidities..... 282, 283, 289–291
Computational target prediction 181, 185,
 226, 284, 291
CRISPR/Cas9.....79–95
Cross-linking and immunoprecipitation (CLIP)..... 80, 115,
 127, 132, 137–139, 175, 179

D

Data integration 282, 283, 286–289
Degradome analysis.....216–221
Denaturation 1–6, 55, 91
Diagnostics 72, 99–101, 149, 297
Differential expression..... 25, 32, 36–39, 100,
 150, 168–170, 181, 182, 194, 196–198, 200–205, 208,
 216, 288, 289
Droplet microfluidics.....71–78

E

Electrochemical sensor 255, 257
Exosomes..... 36, 60, 62, 68, 70, 100, 243, 290

F

Folding-based sensor 250

G

Gold microelectrode.....250, 253–254

H

High-throughput.....7–19, 46, 84, 127–129,
 138, 139, 154, 176, 185, 194, 217, 250, 284

I

Ion-exchange 59–70, 262
Isothermal amplification.....71–78

L

Label-free sensor 250
Library construction 46, 47, 49–51, 53–56
Liquid biopsy.....21

M

Machine learning (ML) 128, 177, 178, 189, 190
Microfluidics 7, 8, 60–62, 65–68, 71–78
MicroRNA
 profiling 21–23, 25, 27,
 37–39, 209
 target identification 124, 176
miScript..... 8, 9, 13–15, 17

N

Next generation sequencing (NGS) 21–43, 46,
 128, 179, 226, 290
Non-enzymatic amplification.....250
Normalization 9, 16, 25, 33, 43, 60, 105,
 169, 194, 196–198, 201–203, 213, 240
Nucleic acid amplification 71

P

Plasma 17–19, 22, 27, 29, 30, 37, 40, 42, 43,
 61, 66, 74, 75, 77, 240, 242, 243, 245, 265–267
Platform comparison 22, 33
Polyacrylamide gel electrophoresis (PAGE) 2, 50,
 55, 86, 113
Polymerase chain reaction (PCR)
 digital.....239–246
 qPCR.....7–19, 21–43, 81–83, 90, 92,
 93, 137, 240, 265–267, 270, 275, 277, 278
 real-time 15, 22, 25, 83, 86, 91,
 103–105, 240, 274–275, 297

R

Reduction of ligation bias 46
 Regulatory networks.....80, 128, 282, 287, 288

S

Serum 17–19, 21–43, 59, 61, 109,
 240, 257–258, 262, 267, 297
 Silicon nanowire (SiNW).....298–300
 Small RNA
 loci 195, 214
 profiling.....45, 46
 Surface acoustic wave 59–70
 Systems biology.....282

T

TaqMan..... 8, 9, 11–13, 17, 101, 103,
 241–244, 266, 268–270, 274, 276
 Thermodynamic stability..... 118, 120,
 130, 226

U

Urine 38, 100–103,
 105, 240

V

Virtual machine.....149–173