UNIVERSITY OF CALIFORNIA,
IRVINE

Integrating microRNA and mRNA dynamics during development and differentiation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology

by

Sorena Rahmanian

Dissertation Committee:
Professor Ali Mortazavi, Chair
Professor Ken W. Cho
Professor Kyoko Yokomori
Associate Professor Robert Spitale
Assistant Professor Zeba Wunderlich

2020

DEDICATION

To

My parents

For being unconditionally loving and supportive

&

To

The Baha'is of Iran

Who stand together in the face of all trials and tribulations

"We cannot segregate the human heart from the environment outside us and say that once one of these is reformed everything will be improved. Man is organic with the world. His inner life molds the environment and is itself also deeply affected by it. The one acts upon the other and every abiding change in the life of man is the result of these mutual reactions."

Shoghi Effendi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my parents for always having faith in me and for respecting and supporting me in all the decisions I have made in my life and to patiently wait for me to make through my journeys.

I would like to thank my advisor professor Ali Mortazavi guided me through difficult times in my research and who always helped me move forward sometimes by pushing me and sometimes by carrying me.

I would like to thank all my committee members, specially professor Robert Spitale for always providing helpful guidance and for believing in my capacities.

Last but not least I am most grateful for all my lab mates and my friends for always encouraging me, cheering me when it was gloomy and helping me when I needed it most.

VITA

Sorena Rahmanian

2008-09	Data Analyst Intern, Teradata, San Diego
2009	CalIT2 summer research, University of California San Diego
2010	B.S. Bioengineering and Biotechnology, University of California San Diego
2011	Summer Internship, Illumina Diagnostics, San Diego
2011	M.S. Bioengineering and Systems Biology, University of California San Diego
2012-2013	Research Associate, AnaptysBio, San Diego
2014-2016	Research Associate, Illumina, San Diego
2017	Scholarship for Biostatistics Summer Initiative, University of Washington
2018	Teaching Assistant, Genetics, University of California Irvine
2020	Ph.D. Mathematical, Computational and Systems Biology, University of California Irvine

FIELD OF STUDY

Mathematical, Computational and Systems Biology

PUBLICATIONS

Rahmanian, S., Rebboah, E., Carvalho, K., McGill, C. J., Spitale, R. C., & Mortazavi, A. *Investigating transcriptome dynamics during HL-60 macrophage differentiation using metabolic labeling*. 'in preparation'

Rahmanian, S., Balderrama-Gutierrez, G., Wyman, D. E., Joan, C. M., Nguyen, K., Spitale, R., & Mortazavi, A. *Long-TUC-seq: A robust method for quantification of metabolically labeled full-length isoforms. BioRxiv*, 073296. https://doi.org/10.1101/2020.05.01.073296

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., **Rahmanian, S.**, Zeng, W., … Mortazavi, A. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *BioRxiv*, 672931. https://doi.org/10.1101/672931

Rahmanian, S., Murad, R., Breschi, A., Zeng, W., MacKiewicz, M., Williams, B., … Mortazavi, A. (2019). Dynamics of microRNA expression during mouse prenatal development. *Genome Research*, *29*(11), 1900–1909. https://doi.org/10.1101/gr.248997.119

Thomas, A., **Rahmanian, S.**, Bordbar, A., Palsson, B. Ø., & Jamshidi, N. (2015). Network reconstruction of platelet metabolism identifies metabolic signature for aspirin resistance. *Scientific Reports*, *4*. https://doi.org/10.1038/srep03925

McConnell, A. D., Zhang, X., Macomber, J. L., Chau, B., Sheffer, J. C., **Rahmanian, S.**, … Bowers, P. M. (2014). A general approach to antibody thermostabilization. *MAbs*, *6*(5). https://doi.org/10.4161/mabs.29680

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., …, **Rahmanian, S.**, … Palsson, B. Ø. (2011). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nature Protocols*, *6*(9). https://doi.org/10.1038/nprot.2011.308

ABSTRACT OF THE DISSERTATION

Integrating microRNA and mRNA dynamics during development and differentiation

By

Sorena Rahmanian

Doctor of Philosophy in Mathematical, Computational and Systems Biology

University of California, Irvine, 2020

Professor Ali Mortazavi, Chair

Developmental processes are extremely complex and precisely coordinated sets of orchestrated changes in the transcriptomic landscape within the cells or tissues involved. These changes are the result of concerted efforts across multiple layers of transcriptional and post-transcriptional regulation. MicroRNAs (miRNAs) are a key class of short, non-coding post-transcriptional regulators with a prominent role in early development and differentiation. The main aim of my research has been to study the role of miRNAs in dynamic processes such as embryonic development in conjunction with the transcriptional changes during those processes. To achieve this goal, we integrated the analysis of miRNA and mRNA data from a set of multiple tissues across different stages of embryonic mouse development. In our study, we first cluster miRNAs and mRNAs separately using a regression-based tool. Then, we used analysis of negative partial correlation of these clusters with each other in parallel with enrichment analysis

of the predicted targets for each miRNA cluster across mRNA clusters. Using this approach, we are able to identify clusters of miRNAs that repress, in a tissue specific manner, the undesired developmental processes pertaining to other tissues.

MicroRNAs affect the steady-state expression of their target mRNAs by destabilizing and degrading them. However, mRNA steady-state expression levels are affected by both transcription and degradation rates, and the changes in steady-state expression measured by RNA-seq can be attributed to either process. A higher resolution of miRNA-mRNA analysis requires studying the dynamics of transcription at the level of individual mRNA molecules that are being made or degraded. Furthermore, many miRNA binding sites fall in UTR regions or exonic/intronic regions of the gene that can vary between isoforms. Identifying exactly which isoforms are expressed can be extremely helpful in distinguishing the degradation rates between different isoform species. Hence, we developed long-TUC-seq, a long-read sequencing protocol that utilizes TUC-seq chemistry (4-thiouridine labeling and its conversion to cytidine using osmium tetroxide) in order to identify RNA molecules that are recently made (or degraded in the case of chase experiments) at transcript isoform resolution.

Finally, in order to consider the dynamics of miRNA biogenesis and degradation itself, we developed micro-TUC-seq, which is a novel method relying on TUC-seq chemistry to identify mature miRNAs generated in a given labeling time window. We apply this method together with regular TUC-seq to decipher the role of miRNAs during HL-60 macrophage differentiation.

CHAPTER 1

Review of mRNA and microRNA dynamics

Dynamics of mRNA expression is coordinated by a plethora of regulatory elements

An intricate network of regulatory elements orchestrates differential gene expression patterns that govern most biological changes and processes across all cell types and a variety of tissues. For example, 90% of genes expressed in the human brain are spatiotemporally regulated (Kang and Blake). More than 2400 genes are differentially expressed across three developmental stages of mouse embryonic brain (Goggolidou et al.). Efforts to understand the drivers of such complex changes in expression profiles have led to the discovery of many cis- and trans-regulatory elements(Gasperini et al.). The expression of genes was mainly known to be driven by cis-regulatory elements such as promoters and enhancers, but recent discoveries have helped us to expand our understanding of these enhancer elements and the complexity involved in their effectiveness. The effect of enhancers on gene expression is controlled by many factors such as the openness of the region of chromatin containing the element, the presence of transcription factors and co-activators that mediate their effects, and the spatial proximity of the enhancer and the promoter of the gene in the nucleus. These are all but some regulatory variables to be considered in predicting and understanding the expression levels of genes in a specific cell or tissue at the time of interest (Fig. 1.1).

The ENCODE (ENCyclopedia Of DNA Elements) Consortium seeks to create a comprehensive catalog of regulatory elements in human and mouse

Many functional genomic assays have been developed over the past couple decades in order to decipher the different layers of gene regulatory networks. Following the completion of the human genome project in 2003 (Collins et al.), the ENCODE consortium was established (www.encodeproject.org) in order to identify and characterize the functional elements of human genome beyond the genes themselves. Since its inception, different labs involved with ENCODE have used a variety of assays to characterize these regulatory elements in different tissues and cell types. The chromosomal architecture is analyzed via Hi-C and ChIA-PET (Fullwood et al.; van Berkum et al.). Open chromatin regions of the genome are assessed by DNase I-seq and ATAC-seq (Boyle et al.; Buenrostro et al.; Thurman et al.) and genome-wide methylation state of genes and promoters are characterized using WGBS and RRBS (Meissner et al.; Lister et al.). Furthermore, the binding of many transcription factors as well as histone markers has been assayed using ChIP-seq (Johnson et al.; Mikkelsen et al.; Pepke et al.). Finally, the actual expression levels of different genes are measured using RNA-seq (Mortazavi et al.; Wilhelm et al.). Together all these assays can be used to build a comprehensive picture of which regulatory elements are playing a role across the genome (Fig. 1.2). The variety of samples and the completeness of the assays have provided the research community with tremendous opportunities for conducting many integrative analyses and studies. One example of such integrative analysis is chromHMM, which builds Hidden Markov Models on the mappings of multiple chromatin markers in different cells to define a chromatin state that can facilitate genome-wide association studies (Chronis et al.; Ernst and Kellis).

Post-transcriptional regulations play just as an important role as transcriptional regulatory elements in defining the expression levels of mRNAs

While the initial ENCODE pilot and majority of phase II of ENCODE (ENCODE2) looked into cis-regulatory elements, more assays were introduced through the later phases that probe potential post-transcriptional elements such as RNA-binding proteins, RNA secondary structures and microRNAs (miRNAs). In ENCODE2, RIP-ChIP (Keene et al.) and RIP-seq

(Cloonan et al.) were employed to study the RNA binding proteins by pulling down the protein of interest and sequencing the enriched RNA. Starting in ENCODE3, other techniques such as eCLIP (Van Nostrand et al.) and RNA Bind-n-Seq (Lambert et al.) were deployed to serve this purpose. Furthermore, recently as part of the current phase of ENCODE (ENCODE4), new assays such as icSHAPE (Flynn et al.; Chan et al.) and LASER (Zinshteyn et al.) are used to characterize the secondary structures in RNA. Finally, another set of assays are aimed at cataloging small non-coding regulatory RNAs such as siRNAs and miRNAs. Small RNA-seq (Fejes-toth et al.), Nanostring and microRNA-seq (Alon et al.) are among these techniques. Together, these techniques allow the researchers to study the structure of RNAs that are expressed and to further understand the relationship between transcription and functionality.

microRNAs are key post-transcriptional regulatory factors that fine tune the expression of their targets

MicroRNAs (miRNAs) are short pieces of non-coding RNA (20-24 nt) that post-transcriptionally regulate the expression levels of their target mRNAs. In the early 1990s the first miRNA was found by chance in *C. elegans* during a developmental study, which revealed that a short RNA made by the lin-4 gene appears to target the expression of lin14 mRNA (Wightman et al.; Lee et al.) It took almost a decade for this class of RNAs to become established and their roles to be well recognized (Lau et al.; Lagos-Quintana et al.; Ambros). MicroRNAs are known to play an important role in developmental processes (Alberti and Cochella; Ivey and Srivastava; Sayed and Abdellatif) and the knockdown of some of their key processing steps had led to lethality in early embryonic stages (Bernstein et al.; Chong et al.). Finally, aberrant miRNA expression is observed in many diseases such as cancer (Zhou et al.; Lee and Dutta; Espinosa and Slack), neurological disorders (Harraz et al.; Wang et al.; Hébert and Strooper) and cardiovascular (Zhao et al.; Romaine et al.; Rooij et al.). Hence, miRNAs have been widely studied as a potential therapeutic approach for many diseases, including going through clinical studies (Rupaimoole and Slack).

Since their discovery in *C. elegans*, miRNAs have been studied in many different organisms, including mammalians such as mouse and human, and a growing number of miRNAs have been identified since then. While the detection of miRNAs relied originally on low throughput methods such as cloning, PCR and microarrays, in recent years scientists have taken advantage of RNA-seq and developed new techniques for high-throughput assess of miRNAs (Tarasov et al.; Motameny et al.; Alon et al.). One of the databases that keeps track of published miRNA sequences and annotations is miRBase (Kozomara and Griffiths-Jones), which annotates 38,589 pre-miRNA and 48,860 mature miRNAs from 271 organisms in v22 (Kozomara et al.).

MicroRNAs are mainly produced through a multi-step canonical pathway, although there are non-canonical pathways that produce miRNAs as well (Havens et al.; Altuvia et al.). MicroRNAs can originate from their host gene as a single miRNA or as a cluster of miRNAs which are under the same transcriptional regulation (Carthew and Sontheimer; Altuvia et al.). In the canonical pathway, the host gene is transcribed by RNA polymerase II to form the primary miRNA (pri-miRNA), which is long-noncoding RNA that is poly-adenylated and capped (Lee et al.). Each pri-miRNA contains one or more hairpin loops which are the sources of miRNAs; the 3’ and 5’ branches of the stem of the loop give rise to the 3’ and 5’ mature miRNA. The pri-miRNA is then cleaved inside the nucleus by the microprocessor complex, composed of Pasha (DGCR8) and a RNase III enzyme called Drosha (Lee et al.). Drosha recognizes the basal single-strand to double strand junction and cuts the pri-miRNA 11 base pairs from there to form a single hairpin loop called precursor miRNA (pre-miRNA) (Nguyen et al.). The pre-miRNA is about 70 nt long and is then transported via exportin5 into the cytoplasm (Lund et al.). The pre-miRNA is further processed by DICER, which is another RNase III enzyme, to form the double-stranded mature miRNA (Ketting et al.). Finally, the double-stranded mature miRNA is loaded into the Argonaute protein (AGO) and one of the strands is selected for targeting, while the other strand is released (Kobayashi and Tomari). Besides the canonical pathway of miRNA biogenesis, there are also non-canonical pathways that produce miRNAs such as mirtrons that are generated via splicing (Ruby et al.).

The miRNA-loaded AGO forms the miRNA induced silencing complex (miRISC) which is guided by the miRNA sequence to its target and represses it (Fabian and Sonenberg). miRISC initially acts by repression of translation, then via mRNA destabilization and degradation (Djuranovic et al.). Repression of translation is mainly facilitated by releasing the initiation factors (eIF4A-I and eIF4A-II) during the initiation phase of translation (Regulation et al.; Fukaya et al.). On the other hand, the degradation of the target mRNA involves initially deadenylation of the transcript with the help of GW182, PABPC, PAN2-PAN3 and CCR4-NOT proteins (J. Liu et al.; Braun, Truffault, et al.; Chen et al.; Behm-Ansmant et al.), which is followed by decapping of the transcript by DCP1-DCP2 (Rehwinkel et al.). Finally, the transcript is degraded via 5’ to 3’ exonuclease activity (Braun, Huntzinger, et al.).

MicroRNAs regulate the transcriptome in a targeted fashion, however the mechanism of their targeting can be elaborate and convoluted

Just like transcription factors and RNA binding proteins, miRNAs regulate expression in a targeted manner. The seed sequence of the miRNA, which is composed of nucleotides 2-8 from the 5’ end of mature miRNA, plays an important role in its target recognition. The majority of the target mRNAs have a target site in their 3’ untranslated region (UTR) that forms Watson-Crick pairing with the seed region (Bartel). Most of these sites also have an “A” across the first position of the miRNA and an additional pairing with position 8 of miRNA(Lewis, Burge, et al.). Other studies have shown that target sites can also be found in the coding regions and 5’ UTR as well (Schnall-Levin et al.). The complementarity of the 3’ region of the miRNA has also been proven to improve the efficacy of targeting, especially in positions 13 to 16 from the 5’ end of the miRNA (Grimson et al.; Moore et al.). Finally, besides aforementioned canonical sites, recent experiments have indicated many non-canonical functional targets with imperfect pairing of the miRNA seed with the target site (Helwak et al.; Seok et al.).

Many experimental methods have been deployed in order to investigate and validate the targets of specific miRNAs. The very first miRNA targeting interaction was discovered by forward genetics in *C. elegans* in 1993 (Lee et al.). Although genetic testing can be laborious and not possible for all organisms, researchers mostly focus on loss-of-function or gain-of-function types of assays, where upon overexpression or knock-down of specific miRNAs, they would study the differential expression of mRNAs using microarrays or RNA-seq (Yu et al.). Finally, a new set of assays has been developed in order to look at the transcriptome-wide targets of all miRNAs. These set of experiments usually entails crosslinking of the RNA-bound proteins to the RNA, then pulling down AGO, and after some further processing, sequencing the RNA to discover the miRNAs and their target RNAs (Helwak et al.; Hafner et al.). However, the last category of the experiments which are more and more common requires more elaborate computational analysis to connect individual miRNAs with their corresponding targets.

Due to the lack of a gold standard for genome-wide discovery of miRNA targets, many groups from early days took on the task of computational prediction of miRNA targets (Lall et al.; Lewis, Burge, et al.; A. Stark et al.). A large group of these prediction tools and algorithms rely on screening and filtering of the transcriptome based on multiple sequence-based features. A couple of the tools that are more established among the community are TargetScan (Lewis, Shi, et al.) and miRanda (Enright et al.). TargetScan originally strictly based its predictions on canonical target sites, their conservation scores and the free energy of miRNA-mRNA duplex. However, over time it evolved to be more flexible with lower conservation regions and imperfect seed pairing and more recently, it considers miRNA 3’ pairing, multiple target sites, as well as targets within ORF regions (Agarwal et al.). miRanda is less stringent than TargetScan; in its newer version, it considers the AU composition of the sequence around the target site and its position in the 3’ UTR as well as predictions of secondary structures in the region containing the target site (Betel et al.).

As more and more experimental validation of the targets became available, a new set of prediction tools have emerged that rely on the data as the training set for different machine learning or similar algorithms. mirSVR is one of these tools that utilizes features from the miRanda package and trains a regression model on expression data from overexpression of specific miRNAs (Betel et al.). In another study, MirTarget uses the miRNA overexpression data to train a Support Vector Machine (SVM) and to identify significant features in predicting the down-regulated targets. Furthermore, these features were integrated with features obtained from training with publicly available CLIP binding data to develop the final model (H. Liu et al.). Finally, in the case of miRwalk2, an ensemble approach has been taken where 13 different predictive tools have been integrated with experimentally validated data (Dweep and Gretz).

Study of nascent and recently made RNAs can help better understand the dynamics of RNA synthesis and decay

Transcriptional and post-transcriptional regulatory elements can help us understand why steady state expression of a gene has changed from one level to another. However, it will not inform us about the kinetics of such a transition and how it has happened. On one hand, the steady state expression level of each gene is a result of how fast it is being transcribed and processed, and on the other hand it is a result of how fast it is being degraded. This is especially important when we are studying the kinetics of the response to an internal or external stimulus. If the stimulus has not affected the factors that regulate transcription rate or degradation of a gene, the steady state expression level of the gene would eventually reach back to its initial level. However, the time it takes to reach its initial level would vary from gene to gene depending on its kinetic parameters; a simple RNA-seq experiment can indicate up-regulation or down-regulation of the gene if it is taken before the time required for it to reach steady state (Fig. 1.3). In order to study such dynamics, many methods and techniques have been developed over the past few years (Yamada and Akimitsu; Rodriguez et al.; R. Stark et al.). In general, these techniques can be divided into those that are studying the nascent transcripts, which capture RNA molecules as they are being transcribed or processed, and those methods that use metabolic labeling, which measure the RNA produced and degraded over a longer window of time.

In addition to all the transcriptional and post-transcriptional regulatory steps, the process of transcription itself can be complex. The work of transcription machinery consists of three main phases: initiation, elongation and termination; however, each step can be very involved with multiple steps that finally determine the overall transcription rate (Core et al.; Revyakin et al.; Sims et al.). Furthermore, the nascent RNA goes through multiple processes as it is being transcribed such as capping, splicing and poly-adenylation (Proudfoot et al.; Black). The rates for all these steps are important in determining the overall transcription rate and many assays have been developed to measure the rate at each of these steps. Nuclear run-on methods isolate the nuclei after pausing transcription, then replace the endogenous nucleotides with exogenous analogs. GRO-seq uses 5-bromouridine 5′-triphosphate (BrU) and PRO-seq uses biotin-modified nucleotides to label the actively transcribing RNAs after resuming transcription (Core et al.; Kwak et al.). Finally, labeled RNA is pulled down and sequenced to discover the active sites of transcription. GRO-seq was used to study the role of enhancer RNAs and divergent or bidirectional transcript initiation (Core et al.; Nagari et al.). Although these methods are useful for detection of transcripts with high turnover, unphysiological restarting of RNA Pol II can introduce some biases (Weber et al.; Mayer et al.). On the other hand, native elongating transcript sequencing (NET-seq) pulls down RNA Pol II-associated RNA using anti-FLAG or antibodies against variety of modifications on C terminal domain (CTD) of Pol II (Churchman and Weissman; Nojima et al.). One limitation of this method is noise that comes from RNA processing intermediates (Mayer and Churchman).

The other set of methods focuses on the RNA molecules that are transcribed over a time window. These methods rely on different nucleotide analogs such as bromouridine (BrU), thiouridine (4SU) and etyniluridine (5EU) to label recently transcribed RNA molecules. One group of these methods uses different techniques to enrich for the labeled RNA, while the other group uses chemical conversion of labeled nucleotides into analogs of different nucleotides. Bru-seq, BruChase-seq and BRIC-seq were all developed around the same time, and they all utilize BrU for the labeling of the RNA and an anti-BrdU antibody to pull down the labeled RNA

(Paulsen et al.; Maekawa et al.). Unlike these methods, 4SU-seq avoids using an antibody that can introduce non-specific binding by using 4SU and thiol-specific reversible biotinylation and streptavidin enrichment. Although the enrichment is easier, this method also suffers from weak biotinylation (Dölken et al.). Similar to 4SU-seq, Transient transcriptome sequencing (TT-seq) utilizes a short labeling with 4SU, however it avoids 5' mapping biases of 4SU-seq by introducing an extra fragmentation step right before 4SU enrichment (Schwalb et al.). Although these methods perform well in detecting high turnover RNA, they require a large amount of input sample and cannot measure the half-life of more stable RNA accurately.

The latest group of these methods utilizes metabolic labeling combined with a chemical conversion step, during which they introduce a substitution in the place of incorporated analog. All these methods use 4SU for the labeling of recently made RNA, however they differ in their secondary conversion step. Thiol (SH)□linked alkylation for the metabolic sequencing of RNA (SLAM□seq) alkylates 4SU via a substitution reaction with iodoacetamide, which consequently is converted to a C during reverse transcription (Muhar et al.). TimeLapse-seq uses oxidative□nucleophilic□aromatic substitution to convert 4SU to an analog of C that has the same hydrogen bond properties as the actual cytidine (Schofield et al.). Finally, thiouridine□to□cytidine□sequencing (TUC□seq) utilizes Osmium tetroxide to convert the 4SU into a perfect C. (Riml et al.; Lusser et al.). All these methods avoid the biases that are introduced via different enrichment techniques; however, they require longer labeling times which could lead to some toxicity, and due to weak incorporation of 4SU in RNA, they do a better job in detecting more genes with higher expression.

Long-read RNA sequencing technologies allow for the study of RNA dynamics at a higher resolution of single isoforms

One of the major sources of biodiversity and RNA complexity in eukaryotes comes from alternative splicing (Black; Pan et al.). Almost all multi-exonic genes (~ 98%) undergo alternative splicing (Wang and Burge); it is a key regulatory step in tissue development and its mis-regulation is associated with many diseases (Baralle and Giudice; Scotti and Swanson). While the Illumina short-read platform is widely used for RNA-sequencing, it has many challenges when it comes to identifying alternative splicing (Broseus and Ritchie). Long-read RNA sequencing technologies are becoming the gold standard for *de novo* transcriptome assembly as well as for isoform-level quantification (Hardwick et al.; Bayega et al.).

The main long-read sequencing platforms are currently from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Both these technologies initially had lower throughput and high error rates, which have substantially improved in recent years. The circular consensus technology with PacBio's new system Sequel II has improved the error rates down to 1% with 8 million reads per SMRT cell (Wenger et al.) ONT has also improved its error rates and throughputs even in its direct RNA sequencing, which sequences RNA directly without any reverse transcription or PCR amplification (Soneson et al.). These advances in third generation sequencing have opened the door for reliable discovery of thousands of new transcripts (Sharon et al.; Wyman et al.).

Studies of recently synthesized RNA benefit from the higher resolution of long-read sequencing

Although long-read sequencing has improved tremendously and proved itself irreplaceable in detecting and quantifying transcript levels, there have been few studies of nascent and labeled transcription using these techniques. Over the last two years, a couple of groups have explored this possibility; however, they both used ONT. Nano-COP uses short 4SU labeling followed by fractionation (similar to NET-seq) and sequencing on ONT (Heather L. Drexler, Karine Choquet). The authors applied this method to study the co-transcriptional dynamics of splicing machinery and discover cooperativity of splicing events. In the other study, the authors use metabolic labeling of the RNA with a 5EU analog for an hour, then perform direct RNA-sequencing without any enrichment. They then train a neural net for base calling of 5EU in addition to the natural nucleotides and use this technique to study the response of GM12878 cells to heat shock (Maier et al.).

Labeled RNA sequencing can shed light into the dynamics of miRNAs and their stabilities

Although miRNAs are a key element in post-transcriptional regulation and decay of mRNAs, they are themselves under coordinated regulation and their stability varies between miRNA species as well as between tissues (Bronevetsky and Ansel; Bail et al.). In general, miRNAs are stable with an average half-life of 119 hours, which is almost 5 days and more than 10-fold longer than for average mRNA (Gantier et al.); however some miRNAs such as miR222-5p and miR125b-1-3p can be very unstable (Guo et al.). In one study, shRNA□induced knockdown of GW182 in HEK293 has led to higher degradation of miRNAs (Yao et al.). Another study claims that the stability of miRNAs significantly depends on their AU content, similar to mRNAs (Sethi and Lukiw). In order to understand and study the stability and transcription dynamics of miRNA, Duffy et al. used 4SU labeling with a longer incubation time to accommodate the high stability of miRNAs followed by methane thiosulfonate (MTS) chemistry (Duffy et al.). They show that the miRNA turnover rates can be estimated by the amount of 4SU labeled miRNA over its steady-state expression level.

Integrative analysis of miRNA and mRNA expression dynamics and their regulatory interactions during mouse embryonic development

As part of ENCODE3, a coordinated set of samples were collected from multiple tissues at different embryonic stages of C57BL/6 mice. These samples were then assayed using different functional genomics assays by the consortium. Most of the samples were assayed for 8 different histone modifications using ChIP-seq, assayed for open chromatin regions using ATAC-seq, characterized for methylation using WGBS and their transcriptome was profiled using RNA-seq and microRNA-seq. In a collaborative effort, our lab received the miRNA data for analysis. A former Ph.D. student, Rabi Murad, initially catalogued the miRNA profiles across the tissues and time points and looked at their cross-species conservation by comparing it with some available data from human tissues.

In order to gain insight into the functionality of these miRNAs and the role that they play during mouse embryonic development, I decided to integrate the miRNA data with the mRNA data that has been collected for the same samples. Since miRNAs work in a very redundant manner and they tend to target a pathway collectively, I clustered the mRNAs across all the samples. We then defined the tissue specificity of each miRNA cluster based on the dynamics of cluster expression in the tissues and calculated tissue-specific correlation between miRNA-mRNA cluster expression. Finally, we coupled this approach with a target enrichment of the miRNA clusters across different mRNA clusters to find the pairs of miRNA-mRNA clusters with the most significant regulatory interactions.

Long-TUC-seq combines labeled RNA with long-read sequencing for enhanced resolution and sensitivity

In a quest to better understand the dynamics of RNA transcription and degradation, we decided to combine RNA metabolic labeling with long-read sequencing. We chose the TUC-seq labeling method for this purpose, because of its lack of enrichment biases compared to the first group of labeled RNA-seq methods and because of its clean conversion to a perfect C without dependency on reverse transcription that could introduce mutations.

We initially tested the TUC-seq chemistry in GM12878 cells, by testing for 4SU incorporation, cell viability post-labeling and RNA integrity post osmium treatment. Once we determined an acceptable range of 4SU concentration and incubation time, we tested whether we could identify the T→ C substitution signature by initially sequencing the sample on the Illumina platform. We observed higher T→C rates in recently synthesized RNA. After establishing a basic pipeline for calling the labeled read, we then switched to PacBio long-read sequencing. We performed long-read and short-read RNA sequencing on the same RNA from 8-hour 4SU treated samples and controls in duplicate. Finally, we conducted a thorough comparison of the two techniques and used the long-TUC-seq results to further analyze the data at isoform levels and to infer isoform-level dynamics.

Micro-TUC-seq can be used to study the stability and transcription dynamics of miRNAs

MicroRNAs also display very dynamic expression and are regulated at many different layers. We decided to further investigate these dynamics by looking at the transcription rate and degradation of miRNAs using a modified version of TUC-seq that we call micro-TUC-seq. Once we established this protocol in GM12878, we tested it in HL60 cell line. We then performed a 5-days differentiation of HL60 cells into macrophages and collected samples across the differentiation time-course. Finally, we performed both TUC-seq and micro-TUC-seq on samples from each time point and analyzed the data.

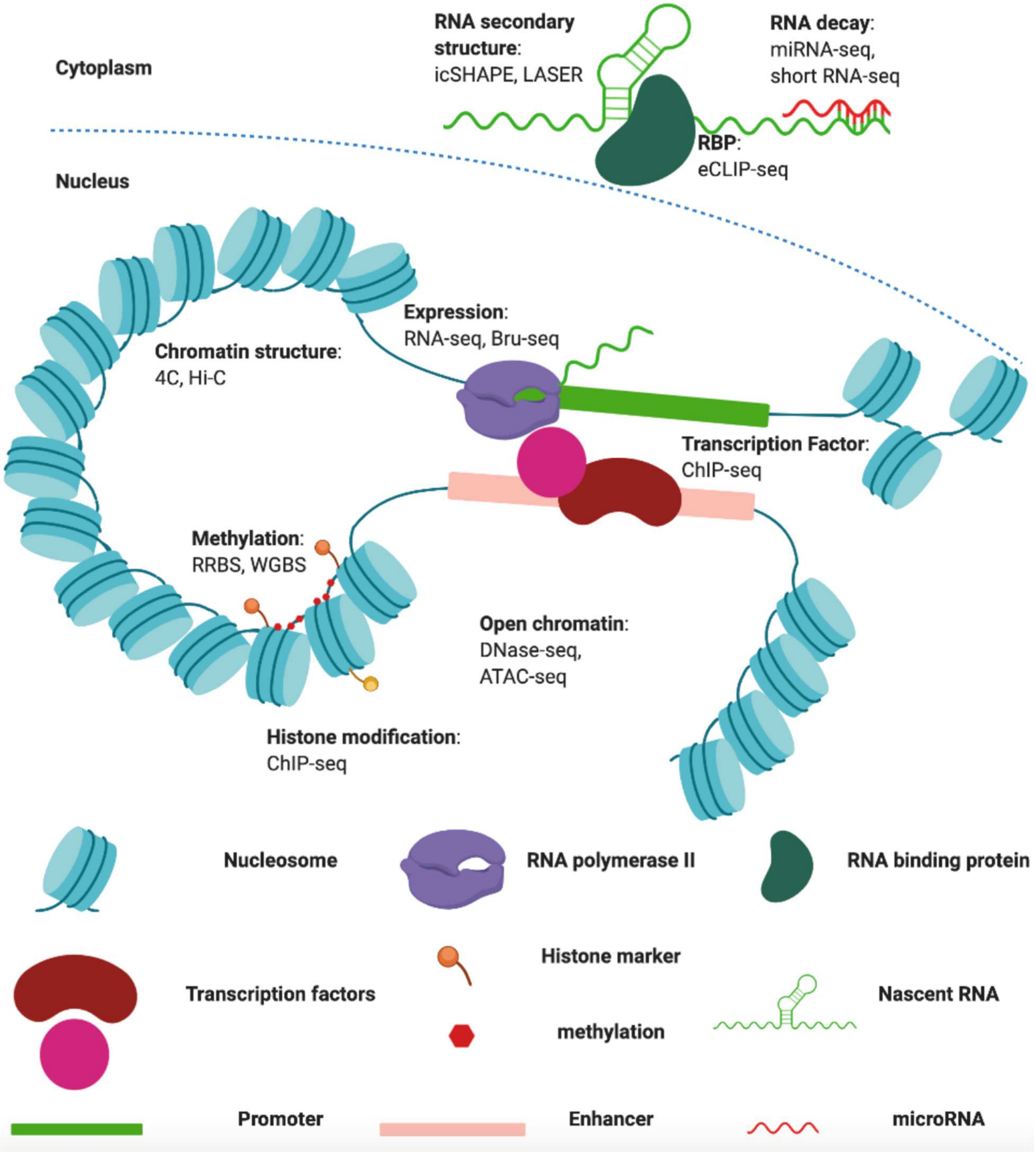

Figure 1.1. Summary of transcriptional and post-transcriptional regulatory elements. A schematic of transcriptional complexity and an array of tools developed to interrogate this complexity. Within the nucleus, the chromosomal architecture and DNA accessibility can be studied by 4C, Hi-C, ATAC-seq, DNase-seq and ChIP-seq. Epigenetic markers such as methylation is assayed by WGBS and RRBS. Transcription factors are studied by ChIP-seq. Outside the nucleus, the post-transcriptional regulatory elements affecting RNA stability can be assayed by icSHAPE, LASER, eCLIP-seq and miRNA-seq.

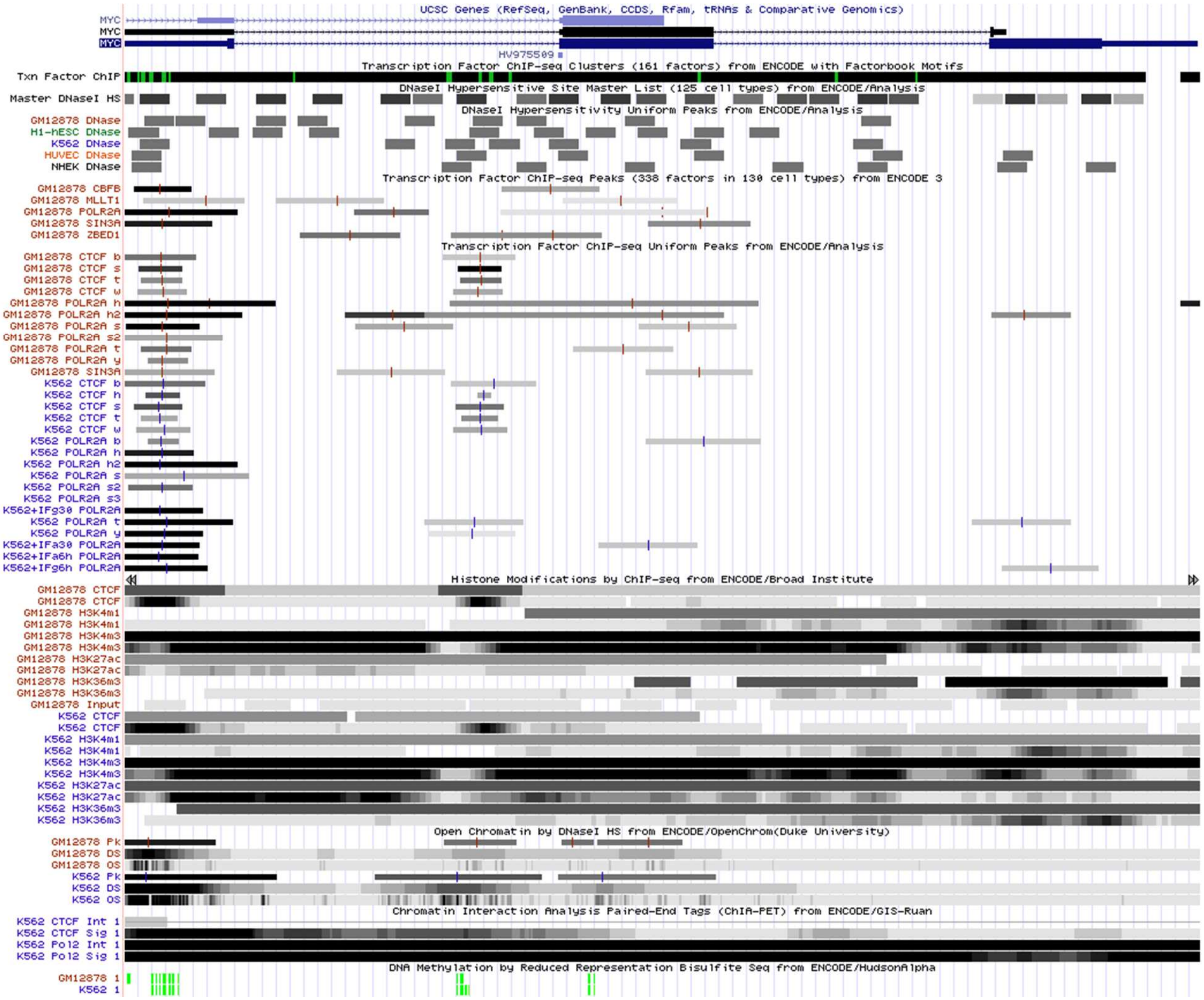

Figure 1.2. ENCODE provides the community with a comprehensive picture of which candidate regulatory elements are found in the genome. A UCSC genome browser shot of representative tracks from the ENCODE project shows a multitude of regulatory elements assayed by different teams. The browser window is showing elements that are mapped to the MYC region of the genome. The elements are selected from assays performed on two of ENCODE tier 1 cell lines: GM12878 and K562. There are many more data tracks available for these cell lines, however for simplicity a few of them have been selected.

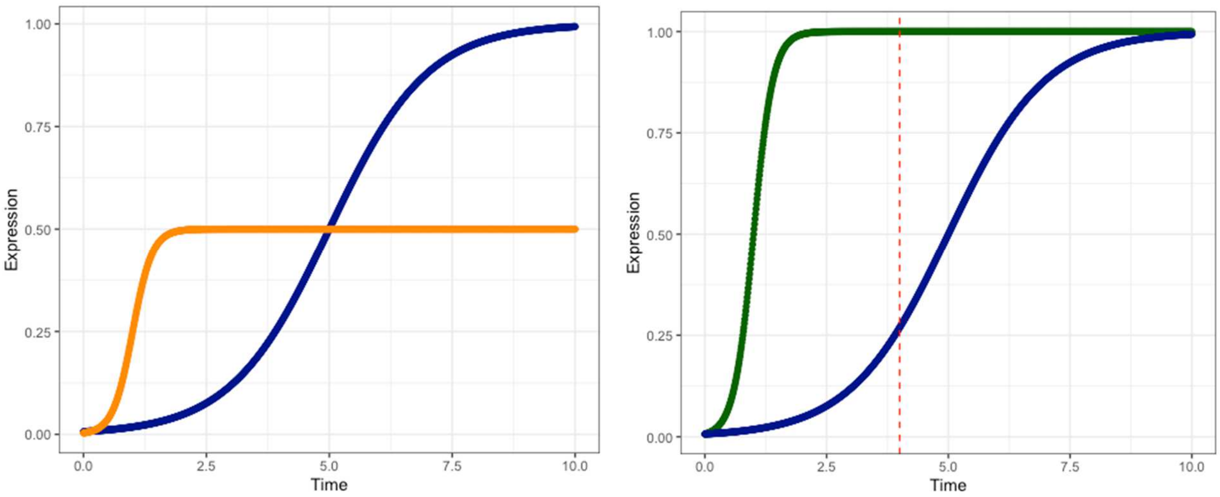

Figure 1.3. Single time point measurements of expression levels do not capture gene dynamics during development. Schematic representation of transcriptional changes upon introduction of a stimulus at time zero. **a**) Showing two different responses to the stimuli, one gene being up-regulated (blue) while the other is down-regulated (yellow). Even though the yellow gene is upregulated, its response is much faster than the blue gene, which can indicate that it has a more primal role in the response to the stimulus compared to the blue gene. **b**) Both blue and green genes are upregulated to the same level in response to the stimulus, however, the green gene's response is much faster. If RNA-seq experiments investigating this differential expression are taken at time 4, before the completion of response by the blue gene, the results can underestimate the blue gene's upregulation.

REFERENCE

Agarwal, Vikram, et al. "Predicting Effective MicroRNA Target Sites in Mammalian MRNAs." *ELife*, vol. 4, no. AUGUST2015, 2015, pp. 1–38, doi:10.7554/eLife.05005.

Alberti, Chiara, and Luisa Cochella. "A Framework for Understanding the Roles of MiRNAs in Animal Development." *Development (Cambridge)*, vol. 144, no. 14, 2017, pp. 2548–59, doi:10.1242/dev.146613.

Alon, S., et al. "Barcoding Bias in High-Throughput Multiplex Sequencing of {miRNA}." *Genome Research*, vol. 21, no. 9, Cold Spring Harbor Laboratory, July 2011, pp. 1506–11, doi:10.1101/gr.121715.111.

Altuvia, Yael, et al. "Clustering and Conservation Patterns of Human MicroRNAs." *Nucleic Acids Research*, vol. 33, no. 8, 2005, pp. 2697–706, doi:10.1093/nar/gki567.

Ambros, Victor. "MicroRNAs: Tiny Regulators with Great Potential." *Cell Press*, vol. 107, 2001, pp. 823–26, doi:10.1007/978-3-642-00150-5_33.

Bail, Sophie, et al. "Differential Regulation of MicroRNA Stability." *Rna*, vol. 16, no. 5, 2010, pp. 1032–39, doi:10.1261/rna.1851510.

Baralle, Francisco E., and Jimena Giudice. "Alternative Splicing as a Regulator of Development and Tissue Identity." *Nature Reviews Molecular Cell Biology*, vol. 18, 2017, pp. 437–51.

Bartel, David P. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell*, vol. 136, no. 2, 2009, pp. 215–33, doi:10.1016/j.cell.2009.01.002.

Bayega, Anthony, et al. "Transcript Profiling Using Long-Read Sequencing Technologies."

Gene Expression Analysis, 2018, pp. 121–47.

Behm-Ansmant, Isabelle, et al. “MRNA Degradation by MiRNAs and GW182 Requires Both CCR4:NOT Deadenylase and DCP1:DCP2 Decapping Complexes.” *Genes and Development*, vol. 20, no. 14, 2006, pp. 1885–98, doi:10.1101/gad.1424106.

Bernstein, Emily, et al. “Dicer Is Essential for Mouse Development.” *Nature Genetics*, vol. 35, no. 3, Springer Nature, Oct. 2003, pp. 215–17, doi:10.1038/ng1253.

Betel, Doron, et al. “Comprehensive Modeling of MicroRNA Targets Predicts Functional Non-Conserved and Non-Canonical Sites.” *Genome Biology*, vol. 11, no. 8, 2010, doi:10.1186/gb-2010-11-8-r90.

Black, Douglas L. “Mechanisms of Alternative Pre-Messenger RNA Splicing.” *Annual Review of Biochemistry*, vol. 72, no. 1, 2003, pp. 291–336, doi:10.1146/annurev.biochem.72.121801.161720.

Boyle, Alan P., et al. “High-Resolution Mapping and Characterization of Open Chromatin across the Genome.” *Cell*, vol. 132, no. 2, 2008, pp. 311–22, doi:10.1016/j.cell.2007.12.014.

Braun, Joerg E., Vincent Truffault, et al. “A Direct Interaction between DCP1 and XRN1 Couples MRNA Decapping to 5′ Exonucleolytic Degradation.” *Nature Structural and Molecular Biology*, vol. 19, no. 12, Nature Publishing Group, 2012, pp. 1324–31, doi:10.1038/nsmb.2413.

Braun, Joerg E., Eric Huntzinger, et al. “A Molecular Link between MiRISCs and Deadenylases Provides New Insight into the Mechanism of Gene Silencing by MicroRNAs.” *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 12, 2012, pp. 1–16, doi:10.1101/cshperspect.a012328.

Bronevetsky, Yelena, and K. Mark Ansel. “Regulation of MiRNA Biogenesis and Turnover in the Immune System.” *Immunological Reviews*, vol. 253, no. 1, 2013, pp. 304–16, doi:10.1111/imr.12059.

Broseus, Lucile, and William Ritchie. “Challenges in Detecting and Quantifying Intron Retention from next Generation Sequencing Data.” *Computational and Structural Biotechnology Journal*, vol. 18, The Authors, 2020, pp. 501–08, doi:10.1016/j.csbj.2020.02.010.

Buenrostro, Jason D., et al. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology*, vol. 2015, no. January, 2015, pp. 21.29.1-21.29.9, doi:10.1002/0471142727.mb2129s109.

Carthew, Richard W., and Erik J. Sontheimer. “Origins and Mechanisms of MiRNAs and SiRNAs.” *Cell*, vol. 136, no. 4, Elsevier Inc., 2009, pp. 642–55, doi:10.1016/j.cell.2009.01.035.

Chan, Dalen, et al. “Measuring RNA Structure Transcriptome-Wide with IcSHAPE.” *Methods*, vol. 120, Elsevier Inc., 2017, pp. 85–90, doi:10.1016/j.ymeth.2017.02.010.

Chen, Jian Fu, et al. “The Role of MicroRNA-1 and MicroRNA-133 in Skeletal Muscle Proliferation and Differentiation.” *Nature Genetics*, vol. 38, no. 2, Springer Nature, Dec. 2006, pp. 228–33, doi:10.1038/ng1725.

Chong, Mark M. W., et al. “Canonical and Alternate Functions of the MicroRNA Biogenesis Machinery (Genes & Development (2010) 24, (1951-1960)).” *Genes and Development*, vol. 24, no. 19, 2010, p. 2228, doi:10.1101/gad.1953310.and.

Chronis, Constantinos, et al. “Cooperative Binding of Transcription Factors Orchestrates Reprogramming.” *Cell*, vol. 168, no. 3, Elsevier, 2017, pp. 442-459.e20, doi:10.1016/j.cell.2016.12.016.

Churchman, L. Stirling, and Jonathan S. Weissman. “Native Elongating Transcript Sequencing (NET-Seq).” *Current Protocols in Molecular Biology*, vol. 1, no. SUPPL.98, Wiley, Apr. 2012, pp. 14.4.1--14.4.17, doi:10.1002/0471142727.mb0414s98.

Cloonan, Nicole, et al. “Stem Cell Transcriptome Profiling via Massive-Scale MRNA Sequencing.” *Nature Methods*, vol. 5, no. 7, 2008, pp. 613–19, doi:10.1038/nmeth.1223.

Collins, Francis S., et al. “The Human Genome Project: Lessons from Large-Scale Biology.” *Science*, vol. 300, no. 5617, 2003, pp. 286–90, doi:10.1126/science.1084564.

Core, Leighton J., et al. “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.” *Science*, vol. 322, no. 5909, American Association for the Advancement of Science ({AAAS}), Dec. 2008, pp. 1845–48, doi:10.1126/science.1162228.

Djuranovic, Sergej, et al. “MiRNA-Mediated Gene Silencing by Translational Repression Followed by MRNA Deadenylation and Decay.” *Science*, vol. 336, no. April, 2012.

Dölken, Lars, et al. “High-Resolution Gene Expression Profiling for Simultaneous Kinetic Parameter Analysis of RNA Synthesis and Decay.” *Rna*, vol. 14, no. 9, 2008, pp. 1959–72, doi:10.1261/rna.1136108.

Duffy, Erin E., et al. “Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry.” *Molecular Cell*, vol. 59, no. 5, Elsevier Inc., 2015, pp. 858–66, doi:10.1016/j.molcel.2015.07.023.

Dweep, Harsh, and Norbert Gretz. “MiRWalk2.0: A Comprehensive Atlas of MicroRNA-Target Interactions.” *Nature Methods*, vol. 12, no. 8, 2015, p. 697, doi:10.1038/nmeth.3485.

Enright, Anton J., et al. “MicroRNA Targets in Drosophila.” *Genome Biology*, 2003, doi:10.1186/gb-2003-5-1-r1.

Ernst, Jason, and Manolis Kellis. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” *Nature Methods*, vol. 9, no. 3, Nature Publishing Group, 2012, pp. 215–16, doi:10.1038/nmeth.1906.

Espinosa, Carlos E. Stahlhut, and Frank J. Slack. “The Role of MicroRNAs in Cancer.” *Targeted Therapy in Translational Cancer Research*, vol. 79, 2006, pp. 131–40, doi:10.1002/9781118468678.ch8.

Fabian, Marc R., and Nahum Sonenberg. “The Mechanics of MiRNA-Mediated Gene Silencing: A Look under the Hood of MiRISC.” *Nature Structural and Molecular Biology*, vol. 19, no. 6, Nature Publishing Group, 2012, pp. 586–93, doi:10.1038/nsmb.2296.

Fejes-toth, Katalin, et al. “Post-Transcriptional Processing Generates a Diversity of 5′- Modified Long and Short RNAs.” *Nature*, vol. 457, no. 7232, 2009, pp. 1–12, doi:10.1038/nature07759.Post-transcriptional.

Flynn, Ryan A., et al. “Transcriptome-Wide Interrogation of RNA Secondary Structure in Living Cells with IcSHAPE.” *Nature Protocols*, vol. 11, no. 2, Nature Publishing Group, 2016, pp. 273–90, doi:10.1038/nprot.2016.011.

Fukaya, Takashi, et al. “MicroRNAs Block Assembly of EIF4F Translation Initiation Complex in Drosophila.” *Molecular Cell*, vol. 56, no. 1, Elsevier Inc., 2014, pp. 67–78, doi:10.1016/j.molcel.2014.09.004.

Fullwood, Melissa J., et al. “Next-Generation DNA Sequencing of Paired-End Tags (PET) for Transcriptome and Genome Analyses.” *Genome Research*, vol. 19, no. 4, 2009, pp. 521–32, doi:10.1101/gr.074906.107.

Gantier, Michael P., et al. “Analysis of MicroRNA Turnover in Mammalian Cells Following Dicer1 Ablation.” *Nucleic Acids Research*, vol. 39, no. 13, 2011, pp. 5692–703, doi:10.1093/nar/gkr148.

Gasperini, Molly, et al. "Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers." *Nature Reviews Genetics*, 2020.

Goggolidou, P., et al. "A Chronological Expression Profile of Gene Activity during Embryonic Mouse Brain Development." *Mammalian Genome*, vol. 24, no. 11–12, 2013, pp. 459–72, doi:10.1007/s00335-013-9486-7.

Grimson, Andrew, et al. "MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing." *Molecular Cell*, vol. 27, no. 1, 2007, pp. 91–105, doi:10.1016/j.molcel.2007.06.017.

Guo, Yanwen, et al. "Characterization of the Mammalian MiRNA Turnover Landscape." *Nucleic Acids Research*, vol. 43, no. 4, 2015, pp. 2326–41, doi:10.1093/nar/gkv057.

Hafner, Markus, et al. "Genome-Wide Identification of MiRNA Targets by PAR-CLIP." *Methods*, vol. 58, no. 2, 2012, pp. 94–105, doi:10.1016/j.ymeth.2012.08.006.

Hardwick, Simon A., et al. "Getting the Entire Message: Progress in Isoform Sequencing." *Frontiers in Genetics*, vol. 10, no. JUL, 2019, pp. 1–10, doi:10.3389/fgene.2019.00709.

Harraz, Maged M., et al. "MicroRNAs in Parkinson's Disease." *Journal of Chemical Neuroanatomy*, vol. 42, no. 2, Elsevier B.V., 2011, pp. 127–30, doi:10.1016/j.jchemneu.2011.01.005.

Havens, Mallory A., et al. "Biogenesis of Mammalian MicroRNAs by a Non-Canonical Processing Pathway." *Nucleic Acids Research*, vol. 40, no. 10, 2012, pp. 4626–40, doi:10.1093/nar/gks026.

Heather L. Drexler, Karine Choquet, L. Stirling Churchman. "Human Co-Transcriptional Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing." *BioRxiv*, 2019.

Hébert, Sébastien S., and Bart De Strooper. "MiRNAs in Neurodegeneration." *Selfless Insight*, vol. 317, 2007, pp. 1179–81, doi:10.7551/mitpress/8053.003.0075.

Helwak, Aleksandra, et al. "Mapping the Human MiRNA Interactome by CLASH Reveals Frequent Noncanonical Binding." *Cell*, vol. 153, no. 3, Elsevier Inc., 2013, pp. 654–65, doi:10.1016/j.cell.2013.03.043.

Ivey, Kathryn N., and Deepak Srivastava. "MicroRNAs as Developmental Regulators." *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. 7, 2015, pp. 1–9, doi:10.1101/cshperspect.a008144.

Johnson, David S., et al. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science*, vol. 316, no. June, 2007, pp. 1497–503.

Kang, Min Suk, and Randolph Blake. "An Integrated Framework of Spatiotemporal Dynamics of Binocular Rivalry." *Frontiers in Human Neuroscience*, vol. 5, no. AUGUST, 2011, pp. 1–9, doi:10.3389/fnhum.2011.00088.

Keene, Jack D., et al. "RIP-Chip: The Isolation and Identification of MRNAs, MicroRNAs and Protein Components of Ribonucleoprotein Complexes from Cell Extracts." *Nature Protocols*, vol. 1, no. 1, 2006, pp. 302–07, doi:10.1038/nprot.2006.47.

Ketting, R. F., et al. "Dicer Functions in RNA Interference and in Synthesis of Small RNA Involved in Developmental Timing in C. Elegans." *Genes and Development*, vol. 15, no. 20, 2001, pp. 2654–59, doi:10.1101/gad.927801.

Kobayashi, Hotaka, and Yukihide Tomari. "RISC Assembly: Coordination between Small RNAs and Argonaute Proteins." *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1859, no. 1, Elsevier B.V., 2016, pp. 71–81, doi:10.1016/j.bbagrm.2015.08.007.

Kozomara, Ana, et al. “MiRBase: From MicroRNA Sequences to Function.” *Nucleic Acids Research*, vol. 47, no. D1, Oxford University Press, 2019, pp. D155–62, doi:10.1093/nar/gky1141.

Kozomara, Ana, and Sam Griffiths-Jones. “MiRBase: Integrating MicroRNA Annotation and Deep-Sequencing Data.” *Nucleic Acids Research*, vol. 39, no. SUPPL. 1, Oxford University Press ({OUP}), Oct. 2011, pp. D152--D157, doi:10.1093/nar/gkq1027.

Kwak, Hojoong, et al. “Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing.” *Science*, vol. 339, 2013, p. 950, doi:10.1126/science.1229386.

Lagos-Quintana, M., et al. “Identification of Novel Genes Coding for Small Expressed RNAs.” *Science*, vol. 294, no. 5543, 2001, pp. 853–58, doi:10.1126/science.1064921.

Lall, Sabbi, et al. “A Genome-Wide Map of Conserved MicroRNA Targets in C. Elegans.” *Current Biology*, vol. 16, no. 5, Elsevier {BV}, Mar. 2006, pp. 460–71, doi:10.1016/j.cub.2006.01.050.

Lambert, Nicole J., et al. “RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA-Binding Proteins.” *Methods in Enzymology*, 1st ed., vol. 558, no. 1, Elsevier Inc., 2015, doi:10.1016/bs.mie.2015.02.007.

Lau, N. C., et al. “An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis Elegans.” *Science*, vol. 294, no. 5543, 2001, pp. 858–62, doi:10.1126/science.1065062.

Lee, Rosalind C., et al. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14.” *Cell*, vol. 75, no. 5, Elsevier {BV}, Dec. 1993, pp. 843–54, doi:10.1016/0092-8674(93)90529-Y.

Lee, Yong Sun, and Anindya Dutta. “MicroRNAs in Cancer.” *Annual Review of Pathology: Mechanisms of Disease*, vol. 4, no. 1, 2009, pp. 199–227, doi:10.1146/annurev.pathol.4.110807.092222.

Lee, Yoontae, et al. “The Nuclear RNase III Drosha Initiates MicroRNA Processing.” *Nature*, vol. 425, no. September, 2003, pp. 415–19.

Lee, Young Sik, et al. “Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the SiRNA/MiRNA Silencing Pathways.” *Cell*, vol. 117, no. 1, 2004, pp. 69–81, doi:10.1016/S0092-8674(04)00261-2.

Lewis, Benjamin P., Christopher B. Burge, et al. “Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets.” *Cell*, vol. 120, no. 1, 2005, pp. 15–20, doi:10.1016/j.cell.2004.12.035.

Lewis, Benjamin P., I-hung Shi, et al. “Prediction of Mammalian MicroRNA Targets.” *Cell Press*, vol. 115, 2003, pp. 787–98, doi:10.1017/S0140525X0131395X.

Lister, Ryan, et al. “Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences.” *Nature*, vol. 462, no. 7271, Nature Publishing Group, 2009, pp. 315–22, doi:10.1038/nature08514.

Liu, Huanle, et al. “Accurate Detection of M6A RNA Modifications in Native RNA Sequences.” *Nature Communications*, vol. 10, no. 1, 2019, pp. 1–9, doi:10.1038/s41467-019-11713-9.

Liu, Jidong, et al. “A Role for the P-Body Component GW182 in MicroRNA Function.” *Nature Cell Biology*, vol. 7, no. 12, 2005, pp. 1161–66, doi:10.1038/ncb1333.

Lund, Elsebet, et al. “Nuclear Export of MicroRNA Precursors.” *Science*, vol. 303, no. 5654, 2004, pp. 95–98, doi:10.1126/science.1090599.

Lusser, Alexandra, et al. “Thiouridine-to-Cytidine Conversion Sequencing (TUC-Seq) to Measure MRNA Transcription and Degradation Rates.” *The Eukaryotic RNA Exosome*, 2020, pp. 191–211.

Maekawa, Sho, et al. "Analysis of RNA Decay Factor Mediated RNA Stability Contributions on RNA Abundance." *BMC Genomics*, vol. 16, no. 1, 2015, pp. 1–19, doi:10.1186/s12864-015-1358-y.

Maier, Kerstin C., et al. "Native Molecule Sequencing by Nano-ID Reveals Synthesis and Stability of RNA Isoforms." *BioRxiv*, Cold Spring Harbor Laboratory, Apr. 2019, p. 601856, doi:10.1101/601856.

Mayer, Andreas, et al. "Pause & Go: From the Discovery of RNA Polymerase Pausing to Its Functional Implications Andreas." *Current Opinion in Cell Biology*, vol. 46, 2017, pp. 72–80, doi:10.1016/j.ceb.2017.03.002.Pause.

Mayer, Andreas, and L. Stirling Churchman. "Genome-Wide Profiling of RNA Polymerase Transcription at Nucleotide Resolution in Human Cells with Native Elongating Transcript Sequencing." *Nature Protocols*, vol. 11, no. 4, Nature Publishing Group, 2016, pp. 813–33, doi:10.1038/nprot.2016.047.

Meissner, Alexander, et al. "Reduced Representation Bisulfite Sequencing for Comparative High-Resolution DNA Methylation Analysis." *Nucleic Acids Research*, vol. 33, no. 18, 2005, pp. 5868–77, doi:10.1093/nar/gki901.

Mikkelsen, Tarjei S., et al. "Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells." *Nature*, vol. 448, no. 7153, 2007, pp. 553–60, doi:10.1038/nature06008.

Moore, Michael J., et al. "MiRNA-Target Chimeras Reveal MiRNA 3′-End Pairing as a Major Determinant of Argonaute Target Specificity." *Nature Communications*, vol. 6, no. May, Nature Publishing Group, 2015, doi:10.1038/ncomms9864.

Mortazavi, Ali, et al. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods*, vol. 5, no. 7, 2008, pp. 621–28, doi:10.1038/nmeth.1226.

Motameny, Susanne, et al. "Next Generation Sequencing of MiRNAs - Strategies, Resources and Methods." *Genes*, vol. 1, no. 1, 2010, pp. 70–84, doi:10.3390/genes1010070.

Muhar, Matthias, et al. "SLAM-Seq Defines Direct Gene-Regulatory Functions of the BRD4-MYC Axis." *Science*, vol. 360, no. 6390, 2018, pp. 800–05, doi:10.1126/science.aao2793.

Nagari, Anusha, et al. "Computational Approaches for Mining GRO-Seq Data to Identify and Characterize Active Enhancers." *Methods in Molecular Biology*, vol. 1468, 2017, pp. 121–38, doi:10.1007/978-1-4939-4035-6.

Nguyen, Tuan Anh, et al. "Functional Anatomy of the Human Microprocessor." *Cell*, vol. 161, no. 6, Elsevier Inc., 2015, pp. 1374–87, doi:10.1016/j.cell.2015.05.010.

Nojima, Takayuki, et al. "Mammalian NET-Seq Reveals Genome-Wide Nascent Transcription Coupled to RNA Processing." *Cell*, vol. 161, no. 3, 2015, pp. 526–40, doi:10.1016/j.cell.2015.03.027.

Pan, Qun, et al. "Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing." *Nature Genetics*, vol. 40, no. 12, 2008, pp. 1413–15, doi:10.1038/ng.259.

Paulsen, Michelle T., et al. "Use of Bru-Seq and BruChase-Seq for Genome-Wide Assessment of the Synthesis and Stability of RNA." *Methods*, 2014, doi:10.1016/j.ymeth.2013.08.015.

Pepke, Shirley, et al. "Computation for Chip-Seq and Rna-Seq Studies." *Nature Methods*, vol. 6, no. 11S, 2009, p. S22, doi:10.1038/nmeth.1371.

Proudfoot, Nick J., et al. "Integrating MRNA Processing with Transcription." *Cell*, vol. 108, no.

4, 2002, pp. 501–12, doi:10.1016/S0092-8674(02)00617-7.

Regulation, Microrna-mediated Gene, et al. "Translational Repression and EIF4A2 Activity Are Critical for MicroRNA-Mediated Gene Regulation." *Science*, vol. 340, no. April, 2013, pp. 82–85.

Rehwinkel, Jan, et al. "A Crucial Role for GW182 and the DCP1:DCP2 Decapping Complex in MiRNA-Mediated Gene Silencing." *Rna*, vol. 11, no. 11, 2005, pp. 1640–47, doi:10.1261/rna.2191905.

Revyakin, Andrey, et al. "Transcription Initiation by Human RNA Polymerase II Visualized at Single-Molecule Resolution." *Genes and Development*, vol. 26, no. 15, 2012, pp. 1691–702, doi:10.1101/gad.194936.112.

Riml, Christian, et al. "Osmium-Mediated Transformation of 4-Thiouridine to Cytidine as Key To Study RNA Dynamics by Sequencing." *Angewandte Chemie*, vol. 129, no. 43, 2017, pp. 13664–68, doi:10.1002/ange.201707465.

Rodriguez, Joseph, et al. "Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity." *Cell*, vol. 176, no. 1–2, Elsevier Inc., 2019, pp. 213-226.e18, doi:10.1016/j.cell.2018.11.026.

Romaine, Simon P. R., et al. "MicroRNAs in Cardiovascular Disease: An Introduction for Clinicians." *Heart*, vol. 101, no. 12, BMJ, Mar. 2015, pp. 921–28, doi:10.1136/heartjnl-2013-305402.

Rooij, Eva van, et al. "A Signature Pattern of Stress-Responsive MicroRNAs That Can Evoke Cardiac Hypertrophy and Heart Failure." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 48, 2006, pp. 18255–60, doi:10.1073/pnas.0608791103.

Ruby, J. Graham, et al. "Intronic MicroRNA Precursors That Bypass Drosha Processing." *Nature*, vol. 448, no. 7149, 2007, pp. 83–86, doi:10.1038/nature05983.

Rupaimoole, Rajesha, and Frank J. Slack. "MicroRNA Therapeutics: Towards a New Era for the Management of Cancer and Other Diseases." *Nature Reviews Drug Discovery*, vol. 16, no. 3, Nature Publishing Group, 2017, pp. 203–21, doi:10.1038/nrd.2016.246.

Sayed, Danish, and Maha Abdellatif. "MicroRNAs in Development and Disease." *Physiological Reviews*, vol. 91, no. 3, American Physiological Society, July 2011, pp. 827–87, doi:10.1152/physrev.00006.2010.

Schnall-Levin, Michael, et al. "Conserved MicroRNA Targeting in Drosophila Is as Widespread in Coding Regions as in 3′UTRs." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 36, 2010, pp. 15751–56, doi:10.1073/pnas.1006172107.

Schofield, Jeremy A., et al. "TimeLapse-Seq: Adding a Temporal Dimension to RNA Sequencing through Nucleoside Recoding." *Nature Methods*, vol. 15, no. 3, Nature Publishing Group, 2018, pp. 221–25, doi:10.1038/nmeth.4582.

Schwalb, Björn, et al. "TT-Seq Maps the Human Transient Transcriptome." *Science*, vol. 352, no. 6290, 2016, pp. 1225–28, doi:10.1126/science.aad9841.

Scotti, Marina M., and Maurice S. Swanson. "RNA Mis-Splicing in Disease." *Nature Reviews Genetics*, vol. 17, no. 1, Nature Publishing Group, 2016, pp. 19–32, doi:10.1038/nrg.2015.3.

Seok, Heeyoung, et al. "MicroRNA Target Recognition: Insights from Transcriptome-Wide Non-Canonical Interactions." *Molecules and Cells*, vol. 39, no. 5, 2016, pp. 375–81, doi:10.14348/molcells.2016.0013.

Sethi, Prerna, and Walter J. Lukiw. “Micro-RNA Abundance and Stability in Human Brain: Specific Alterations in Alzheimer’s Disease Temporal Lobe Neocortex.” *Neuroscience Letters*, vol. 459, no. 2, 2009, pp. 100–04, doi:10.1016/j.neulet.2009.04.052.

Sharon, Donald, et al. “A Single-Molecule Long-Read Survey of the Human Transcriptome.” *Nature Biotechnology*, vol. 31, no. 11, 2013, pp. 1009–14, doi:10.1038/nbt.2705.

Sims, Robert J., et al. “Recent Highlights of RNA-Polymerase-II-Mediated Transcription.” *Current Opinion in Cell Biology*, vol. 16, no. 3, 2004, pp. 263–71, doi:10.1016/j.ceb.2004.04.004.

Soneson, Charlotte, et al. “A Comprehensive Examination of Nanopore Native RNA Sequencing for Characterization of Complex Transcriptomes.” *Nature Communications*, vol. 10, no. 1, Springer US, 2019, pp. 1–14, doi:10.1038/s41467-019-11272-z.

Stark, Alexander, et al. “Identification of Drosophila MicroRNA Targets.” *PLoS Biology*, vol. 1, no. 3, 2003, doi:10.1371/journal.pbio.0000060.

Stark, Rory, et al. “RNA Sequencing: The Teenage Years.” *Nature Reviews Genetics*, vol. 20, 2019, pp. 631–56.

Tarasov, Valery, et al. “Differential Regulation of MicroRNAs by P53 Revealed by Massively Parallel Sequencing: MiR-34a Is a P53 Target That Induces Apoptosis and G 1-Arrest.” *Cell Cycle*, vol. 6, no. 13, 2007, pp. 1586–93, doi:10.4161/cc.6.13.4436.

Thurman, Robert E., et al. “The Accessible Chromatin Landscape of the Human Genome.” *Nature*, vol. 489, 2012, p. 75082.

van Berkum, Nynke L., et al. “Hi-C: A Method to Study the Three-Dimensional Architecture of Genomes.” *Journal of Visualized Experiments*, no. 39, 2010, pp. 1–7, doi:10.3791/1869.

Van Nostrand, Eric L., et al. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (ECLIP).” *Nature Methods*, vol. 13, no. 6, 2016, pp. 508–14, doi:10.1038/nmeth.3810.

Wang, Gaofeng, et al. “Variation in the {miRNA}-433 Binding Site of {FGF}20 Confers Risk for Parkinson Disease by Overexpression of $\upalpha$-Synuclein.” *The American Journal of Human Genetics*, vol. 82, no. 2, Elsevier {BV}, Feb. 2008, pp. 283–89, doi:10.1016/j.ajhg.2007.09.021.

Wang, Zefeng, and Christopher B. Burge. “Splicing Regulation: From a Parts List of Regulatory Elements to an Integrated Splicing Code.” *Rna*, vol. 14, no. 617, 2008, pp. 802–13, doi:10.1261/rna.876308.802.

Weber, Christopher M., et al. “Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase.” *Molecular Cell*, vol. 53, no. 5, Elsevier Inc., 2014, pp. 819–30, doi:10.1016/j.molcel.2014.02.014.

Wenger, Aaron M., et al. “Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome Aaron.” *Nature Biotechnology*, vol. 37, no. 10, 2020, pp. 1155–62, doi:10.1038/s41587-019-0217-9.Accurate.

Wightman, Bruce, et al. “Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern Formation in C. Elegans.” *Cell*, vol. 75, no. 5, 1993, pp. 855–62, doi:10.1016/0092-8674(93)90530-4.

Wilhelm, Brian T., et al. “Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution.” *Nature*, vol. 453, no. 7199, 2008, pp. 1239–43, doi:10.1038/nature07002.

Wyman, Dana, et al. “A Technology-Agnostic Long-Read Analysis Pipeline for Transcriptome Discovery and Quantification.” *BioRxiv*, 2019, p. 672931, doi:10.1101/672931.

Yamada, Toshimichi, and Nobuyoshi Akimitsu. “Contributions of Regulated Transcription and MRNA Decay to the Dynamics of Gene Expression.” *Wiley Interdisciplinary Reviews: RNA*, vol. 10, no. 1, Wiley, Oct. 2019, p. e1508, doi:10.1002/wrna.1508.

Yao, Bing, et al. “Defining a New Role of GW182 in Maintaining MiRNA Stability.” *EMBO Reports*, vol. 13, no. 12, Nature Publishing Group, 2012, pp. 1102–08, doi:10.1038/embor.2012.160.

Yu, Yue, et al. “Loss-of-Function Screening to Identify MiRNAs Involved in Senescence: Tumor Suppressor Activity of MiRNA-335 and Its New Target CARF.” *Scientific Reports*, vol. 6, no. July, Nature Publishing Group, 2016, pp. 1–13, doi:10.1038/srep30185.

Zhao, Yong, et al. “Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking MiRNA-1-2.” *Cell*, vol. 129, no. 2, 2007, pp. 303–17, doi:10.1016/j.cell.2007.03.030.

Zhou, Jianqing, et al. “MiRNA 206 and MiRNA 574-5p Are Highly Expression in Coronary Artery Disease.” *Bioscience Reports*, vol. 36, no. 1, 2016, pp. 1–7, doi:10.1042/BSR20150206.

Zinshteyn, Boris, et al. “Assaying RNA Structure with LASER-Seq.” *Nucleic Acids Research*, vol. 47, no. 1, Oxford University Press, 2019, pp. 43–55, doi:10.1093/nar/gky1172.

CHAPTER 2

Dynamics of microRNA expression during mouse prenatal development

Note: this chapter is an abbreviated version of the manuscript published in Genome Research.

This chapter includes the sections that I primarily worked on.

ABSTRACT

MicroRNAs (miRNAs) play a critical role as post-transcriptional regulators of gene expression. The ENCODE Project profiled the expression of miRNAs in an extensive set of organs during a time-course of mouse embryonic development and captured the expression dynamics of 785 miRNAs. We found distinct organ and developmental stage specific miRNA expression clusters, with an overall pattern of increasing organ-specific expression as embryonic development proceeds. An analysis of messenger RNA expression clusters compared with miRNA expression clusters identifies the potential role of specific miRNA expression clusters in suppressing the expression of mRNAs specific to other developmental programs in the organ where these miRNAs are expressed during embryonic development. Our results provide the most comprehensive time-course of miRNA expression as part of an integrated ENCODE reference dataset for mouse embryonic development.

INTRODUCTION

Development is a well-orchestrated process primarily controlled by transcriptional regulators with post-transcriptional regulators such as microRNAs (miRNAs) playing an essential role in fine tuning gene expression dynamics. MicroRNAs are small ~22 nucleotide (nt) endogenous non-coding RNAs that regulate gene expression by mediating the post-transcriptional degradation of messenger RNA (mRNA) or by hindering the translation of proteins (Bartel, “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function”; He and Hannon). MicroRNA biogenesis occurs in several steps, starting with transcription of typically polyadenylated primary miRNA (pri-miRNA) transcripts (>200 nt), sometimes referred to as the “host genes”. These pri-miRNA have a characteristic hairpin structure that is cleaved in the nucleus by the enzyme Drosha into pre-miRNA (~60 nt), which are exported to the cytoplasm before finally being processed into 21-24 nt mature miRNA by the enzyme Dicer (Han et al.). The first miRNA was discovered in the nematode *C. elegans* as perturbing its cell developmental lineage (Lee et al.) and since then thousands of miRNAs have been discovered in diverse plants, metazoans, and some viruses (Kozomara and Griffiths-Jones).

Many studies have shown that the deletion of key players in the biogenesis of miRNA such as Ago2, Dicer1 and Dgcr8 lead to embryonic lethality and arrest (Alisch et al.; Morita et al.; Bernstein et al.; Wang et al.). However loss of single miRNAs does not have as dramatic an effect as knocking out all the miRNAs in the organism (Park et al.). This could be due to the redundancy of miRNA-mRNA interactions as each mRNA could be targeted by multiple miRNAs and thus the lack of one miRNA would be compensated by others. Hence there is a strong rationale for studying the role of miRNAs as a functional group or unit. Studies have shown that most genes are potential targets of miRNAs (Friedman et al.) and that miRNAs are involved in regulating diverse cellular processes during development and homeostasis (Vidigal and Ventura). Dysregulation of miRNA expression is known to underlie numerous diseases and developmental defects such as cancer (Lin and Gregory), cardiovascular diseases (Romaine et al.; Zhao et al.), and neurological diseases (Cao et al.).

MicroRNAs have been profiled in various tissues and primary cells in diverse metazoans and plants (Wienholds et al.; Lagos-Quintana et al.; Ehrenreich and Purugganan). Mineno and colleagues used massively parallel signature sequencing (MPSS) technology to profile miRNAs in mouse whole embryos during three embryonic stages (E9.5, E10.5, and E11.5) and were able to detect 390 distinct miRNAs (Mineno et al.). Chiang and colleagues extended this work by sequencing small RNAs from mouse brain, ovary, testes, embryonic stem cells, embryonic stages of complete embryos from three developmental stages, and whole newborns to profile the expression of 398 annotated and 108 novel miRNAs (Chiang et al.). Landgraf and colleagues cloned and sequenced more than 250 small RNA libraries from 26 different organs and cell types from humans and rodents to profile miRNA expression and describe various other miRNA characteristics (Landgraf et al.). More recently, the FANTOM5 project has created a miRNA expression atlas using deep-sequencing data from 396 human and 47 mouse RNA samples (De Rie et al.); however, many of these mouse samples were simply replicates of a handful of mouse cell lines. Previous efforts by the ENCODE Consortium affiliates focused on a meta-analysis of previously published 501 human and 236 mouse small RNA sequencing data sets from a multitude of sources to characterize splicing-derived miRNAs (mirtrons) in the human and mouse genomes (Ladewig et al.). However, the diversity of the source tissues and the different underlying experimental protocols from the disparate primary sources complicated any sort of systematic quantitative analysis. Last but not least, many individual studies have focused on the expression of particular miRNAs in certain tissues in a handful of (typically 2-3) mouse developmental time points. Therefore, a complete and systematic atlas of miRNA expression during development of the major organ systems and broad number of mouse embryonic stages is still missing. This is not only helpful for understanding mouse development, but also for studying the potential role of miRNAs in human development where access to the same time points is either very difficult or outright impossible.

With the growing evidence of the critical role of miRNAs in homeostasis and disease, multiple techniques have been developed for profiling the expression of mature miRNAs, each with their own strengths (Mestdagh et al.). RNA-seq typically refers to the profiling of expressed transcripts 200 nt or longer including the messenger RNAs (mRNA) and long non-coding RNAs (lncRNA) (Mortazavi et al.), which in this work we will refer to as messenger RNA-seq (mRNA-seq), whereas there are also multiple miRNA-specific sequencing protocols such as microRNA-seq (Roberts et al.; Alon et al.) and short RNA-seq (Fejes-Toth et al.). There are also hybridization-based assays such as microarrays as well as molecule counting such as NanoString, which involves hybridization and counting of color-coded molecular barcodes (Wyman et al.; Geiss et al.). As mature miRNAs are processed from longer host pri-miRNAs and the annotated pri-miRNAs are predominantly protein-coding or lncRNA transcripts (Cai et al.), we expect that mRNA-seq should be able to profile the expression of pri-miRNAs. However, there is a significant number of miRNAs whose host genes have not been characterized yet. Furthermore, an important question is whether the expression of pri-miRNAs can reliably predict the expression of their corresponding mature miRNAs. As previously reported (Zeng et al.; De Rie et al.), this would allow the simultaneous profiling of mature miRNA expression along with mRNAs using mRNA-seq. Availability of matching mRNA-seq and microRNA-seq data sets for the same samples in our study provides a unique opportunity to answer this question using a broader set of organs and developmental time points. Furthermore, the corresponding mRNA data can shed light into the targeting of these miRNAs and their functional role during embryonic development.

Each miRNA targets a set of mRNAs through Watson-Crick pairing between miRNA seed region (positions 2-7 from 5' end) and the binding sites on their targets (Bartel, "MicroRNAs: Target Recognition and Regulatory Functions"). Such complementary base-pairing has been used to computationally predict miRNA targets (Bartel, 2009). Expression of miRNAs and mRNAs in matching samples have been used to identify miRNA-mRNA interactions, for example in cancer (McLendon, 2008). Several methods such as biclustering (Jin and Lee) have been used to infer miRNA-mRNA interactions from gene expression data. However, the expression levels of mRNAs are often affected by multiple factors and comparison of mRNA and miRNA expressions cannot establish a functional relationship by itself. Therefore, an approach that integrates miRNA and mRNA expression data and their predicted interactions should provide better inference of their functional interaction networks.

In this study, we used microRNA-seq to characterize the expression patterns of known miRNAs using a set of 16 different mouse organs (we use the term interchangeably with tissues in this study) at 8 embryonic (E10.5 – P0) stages that were specifically selected by the ENCODE Consortium for a wide variety of functional sequencing assays such as mRNA-seq, ChIP-seq, and DNase-seq. The value of this dataset is that the samples and stages are all matched. We show one example of integrative analysis of the microRNA-seq data with matching ENCODE mRNA-seq to compare the characteristics and dynamics of miRNA expression to characterize the changes in overall tissue specificity of particular miRNAs during mouse development. In particular, we compute enrichment of computationally predicted miRNA targets in specific mouse organs along with the negative partial correlation analysis of miRNA and mRNA expression clusters during mouse development to identify developmental processes targeted by miRNAs. We find that groups of miRNAs expressed in one or more organs target groups of developmentally important mRNAs highly expressed in other organs.

RESULTS

A reference miRNA catalog across mouse development

As part of the ENCODE Project, we used microRNA-seq to profile mature miRNAs during mouse embryonic development. This study encompasses 156 microRNA-seq mouse organs with two biological replicates each (Fig. 2.1A). We subsampled one of our deeply sequenced datasets (heart E11.5 with 60 million reads) to evaluate the impact of sequencing depth on the robust detection of miRNAs. While we detect more miRNAs as we sequence deeper, we detect a constant number using a CPM (Counts per Million) based cut off and we selected one million mapped reads to be the sufficient depth of sequencing required for detecting most of miRNAs expressed in a sample (Fig. 2.2). All of our samples in this study were sequenced to a minimum of two million mapped reads per replicate. All data from this study are available from the ENCODE data portal (www.encodeproject.org).

We used a set of three spike-ins of different sequence lengths (22 bp, 25 bp, and 30 bp) in decreasing concentrations (5000, 500, and 50 pM respectively) in our microRNA-seq samples to assess replicate concordance for different library normalization strategies (Fig. 2.3). While the spike-in counts were highly concordant for biological replicates for each sample, they differ for different stages of mouse embryonic development using counts per million (CPM) normalization only. We found that TMM normalization of miRNA CPMs ameliorates such differences in spike-in expression across developmental stages. Therefore, we normalized our data using TMM normalization for downstream analysis.

We used microRNA-seq reads to quantify miRNA expression levels using miRBase version 22 annotations, which includes 1981 mature miRNAs. We detected 785 of these mature miRNAs expressed in at least one of the samples; About 80% of these mature miRNAs correspond to the pre-miRNAs identified as high confidence by miRBase (Kozomara et al., 2013). This set of miRNAs encompasses 72% of high-confidence miRNAs in miRBase compared to the 65% recovered by FANTOM dataset (De Rie et al., 2017) and 71% of miRNAs annotated in MiRGeneDB (Fromm et al., 2018) (Fig. 2.4). We detect additional miRNAs if we use a more lenient cutoff of one read as opposed to two CPM (miRBase: 72%, miRase high-confidence: 92%, and MiRGeneDB: 77%). There are no significant differences in the number of distinct miRNAs expressed in mouse organs and developmental stages and although stage P0 has the highest number of organs profiled as well as the least number of distinct miRNAs detected (Fig. 2.1B). This result is in contrast to the finding that the absolute numbers of expressed miRNAs increase over the developmental time in other model organisms such as *Drosophila melanogaster* (Ninova et al.). At the organ level, we find that the nervous system samples show the highest number of distinct miRNAs expressed (Fig. 2.1B).

Dynamics of miRNA tissue specificity during development

As previous studies have shown, a few highly expressed miRNAs are responsible for most of the detected expression (Lagos-Quintana et al., 2002; Landgraf et al., 2007), with about 50% of the expression corresponding to the top 10 expressed miRNAs (Fig. 2.5). Only 42 miRNAs fall within the top ten expressed miRNAs across our 72 distinct tissue-stage samples. Six of these miRNAs are in the top 10 expressed list for more than half of the samples with miR-16-5p and miR-26a-5p being one of the top expressed miRNAs in every single experiment. To study the specificity of the miRNAs at each stage, we used the Tissue Specificity Index (TSI) as defined previously (Ludwig et al., 2016); using this metric, we found that 40% of the top expressed miRNAs are tissue specific in at least one of the stages that they are highly expressed in. These miRNAs include: miR-1a-3p, miR-208b-3p and miR-351-5p in heart (the last one is only specific in the earlier stages); miR-9-3p, miR-9-5p, miR-124-3p, miR-125b-5p and miR-92b-3p in brain; miR-122-5p and miR-142a-3p in liver; miR-10a-5p in kidney; miR-194-5p in intestine; miR-196b-5p in limb. (Fig. 6)

While there are few miRNAs that are expressed ubiquitously (TSI < 0.15) at the earlier stages of embryonic development, most miRNAs become more tissue-specific as the embryo develops further and the expression of miRNAs shifts from ubiquitous to being organ-specific (Fig. 2.1C; Fig. 2.7). This shift is partly due to changes in the specificity of the miRNAs throughout development with the following miRNAs showing the most change: miR-128-3p, miR-181a-1-3p, miR-138-5p and miR-3099-3p in brain; miR-101a-3p and miR-496a-3p in liver; and miR-140-5p in limb (Fig. 2.8). All of these miRNAs increase in their specificity from being almost ubiquitous to become organ specific. However, there is a group of more than 20 miRNAs that stay mostly organ-specific throughout the developmental time points captured in our study. This group includes some of the well-studied tissue-specific miRNAs such as: miR-9 and miR-92b-3p in brain; miR-1a-3p, miR-208a-3p and miR-133a-3p in heart; miR-122-5p in liver (Fig. 2.9). Finally, there is a group of miRNAs that are present in almost all the tissues at every stage of development including: miR-421-3p, miR-361-5p and miR-744-5p. (Fig. 2.10). In summary, our high-resolution time course captures the distinct patterns of miRNA expression during mouse embryonic development.

Clustering of miRNAs recovers distinct tissue specific clusters

Global analysis of mouse tissues and developmental stages shows distinct miRNA expression patterns as revealed by principal component analysis (PCA) (Fig. 2.1D). Principal component 1 (PC1) accounts for 24.5 percent of the variation and clearly separates the various tissues with the nervous system and liver tissues at the extremes, whereas PC2 (16.5 % variation) represents the time component of mouse development with a temporal gradient between early development at embryonic day 10 (E10.5) and postnatal samples right after birth (P0) (Fig.

2.1D), PC3 (10.8% variation) separates kidney samples from liver, PC4 (6.1% variation) separates heart samples from other tissues and PC5 (4.5% variation) separates kidney samples from limb and craniofacial samples. Overall the first five principal components explained over 60% of the variation in the dataset with most of that variation corresponding to specific tissues.

We used maSigPro (Nueda et al.) to cluster the 785 expressed miRNAs based on the tissue-specific changes in their expression during the development. maSigPro identified 535 of these miRNAs as being differentially expressed (Table 2.1) during embryonic development into 16 clusters based on regression of their expression levels in each of the tissues (Fig. 2.11). Cluster 11 has the highest number of miRNAs (96 miRNAs) that are highly expressed in brain. Additionally, the expression of these miRNAs increase during embryonic development whereas in cluster 2, another brain-specific cluster, the expression of miRNAs goes up initially and comes down after embryonic day 14. miRNA clusters 4, 12 and 14 are the second largest clusters, with 54 miRNAs each. Clusters 4 and 12 are composed of miRNAs mostly expressed in liver and heart respectively, whereas miRNAs in cluster 14 are highly expressed in all the tissues except in liver and brain. Analysis of tissue specific miRNAs in each cluster reveals that more than half of the miRNA clusters are enriched for specificity to only one organ with the rest of clusters enriched for specificity in 2 or 3 different organs (Fig. 2.12). Thus, miRNAs during development show distinct clustered expression in select tissues.

Integrative analysis of miRNA and mRNA expression profiles during mouse development identifies significant anti-correlations of developmentally important genes with miRNAs predicted to target them

In order to understand the connection between miRNAs and the expression of their targets, we developed an integrative analysis pipeline to connect miRNAs to their mRNA targets

(Fig. 2.13). As a first step, we quantified the tissue specificity of miRNA clusters by computing a tissue specificity matrix. The tissue specificity of each miRNA cluster was determined based on the expression changes of miRNAs in each tissue during development. The tissues that had the highest standard deviation of a given miRNA cluster's expression in different stages were identified as the tissue specificity of that cluster. The tissue-specificity of the miRNA clusters calculated in this manner are highly concordant with the tissue-specificities obtained by the specificity analysis of the individual miRNAs in each cluster. There is at least one miRNA cluster identified for each tissue and at least one tissue identified as tissue specific for each miRNA cluster (Fig. 12; Table 2.2).

We clustered mouse developmental mRNAs from ENCODE using maSigPro into 30 clusters incorporating 14,827 differentially expressed genes out of the 20,686 genes that were expressed at least once during the development with a replicate average expression of at least two TPM (Fig. 2.14). About one third of these clusters are specific to a single tissue, with the rest being expressed in multiple tissues. The largest three clusters are clusters 9, 3 and 12 with 1,773, 1,214 and 884 genes in them respectively and all three of these clusters correspond to genes expressed in brain. Most of the tissue specific clusters correspond to liver, heart and lung after the brain.

After identifying the clusters of miRNA and mRNA that are dynamically expressed during development, we calculated the partial correlation between each of the miRNA clusters and each of the mRNA clusters. The partial correlation matrix was built using Pearson correlation between each pair of clusters (miRNA-mRNA) within the context of tissues that the miRNA cluster was active in (using only the miRNA tissue specificity as the context) (Fig. 2.15). Using this partial correlation approach, 60% of the miRNA-mRNA cluster interactions are anti-correlated with a mean correlation coefficient value of -0.47. This anti-correlation was used to filter out the positive interactions after target enrichment analysis.

We collected the predicted targets of each of the miRNAs from five different resources and prediction algorithms using miRNAtap (Pajak & Simpson 2018). We used the unique set of all the predicted targets for miRNAs in each of the miRNA clusters to build a contingency table that contains the distribution of each of these unique target sets among the mRNA clusters. We then performed a χ-square test on the contingency table to study the enrichment of targets in different mRNA clusters (Fig. 2.13) and applied a p-value cut off of 10^{-4} (Bonferroni corrected P-value: 0.05/(16*30)) to determine the mRNA clusters that were significantly enriched for miRNA cluster targets. 18 interactions between 11 unique miRNA clusters and 7 unique mRNA cluster were identified as significant, however only 9 of these interactions passed the filter for negative partial correlation (Fig. 2.16; Table 2.3). We also evaluated the conservation of the 3'UTR 8mer target sites for different miRNA seeds and gene cluster interactions (Fig. 2.17). We found that 8 out of our 9 significant interactions fall in the top 30 percentile of conserved 8mer target sites. In particular, miR clusters 11, 14 and 6 have the highest number of conserved targets across the gene clusters. From these 9 significant interactions, we chose to further analyze two pairs. The interaction between miRNA cluster 11 and mRNA cluster 18 (Fig. 2.13) had a P-value of 10^{-5} for the target enrichment and a correlation coefficient of -0.73. The miRNAs involved in this interaction are highly expressed and increase during time in brain whereas the target genes are expressed more highly in other tissues such as limb, cranioface, and heart at the same stages of development. Gene ontology of the targets revealed that this miRNA cluster targets genes involved in the development of skeletal system, cardiac development and vasculature development (p-values < 10^{-15}), by presumably downregulating them in the brain.

Another interaction between miRNA cluster 6 and mRNA cluster 28 has a P-value of $5.2*10^{-5}$and negative correlation coefficient of -0.68. This miRNA cluster increases expression mainly in the heart, lung and kidney (Fig. 2.13) whereas the mRNA targets are highly expressed in brain organ and their expression is very limited in heart (Fig. 2.13). Gene ontology analysis of this interaction enriches for terms involved with head and brain development (p-values < 10^{-5}; Fig. 2.3H). In both of these cases as well as several of the others, the miRNA cluster is enriched for targets that are developmentally important genes for tissues other than the tissue in which the miRNAs are highly expressed.

DISCUSSION

In this study we provide a comprehensive resource of miRNA expression dynamics across mouse developmental stages in multiple organs. Our catalogue of organ and developmental stage specific miRNAs provides a valuable resource for elucidating the role of miRNAs and highlighting certain key properties of miRNAs during mouse development. we detected only 42% of the annotated miRNAs in mouse expressed a minimum of 2 CPM (72% of miRNAs annotated as highly confident) in the 16 different organs that are representative of major organ systems during mouse embryogenesis. This result suggests that only a subset of miRNAs might be involved in regulating gene expression during mouse development with the remaining either expressed in other tissues or more likely expressed later in post-natal development and adult tissues (Ludwig et al.). There is also little variability in the number of miRNAs detected per tissues with the heart and nervous system tissues exhibiting the highest number of detected miRNAs. Interestingly, the miRNA output of most embryonic samples is dominated by the expression of a few highly expressed miRNAs that usually consist of nontissue-specific and ubiquitously but highly expressed miRNAs, which matches reports from human and mouse cell types (De Rie et al.).

Although tissue specificity of miRNAs has been well studied and well reported in multiple model organisms (Gao et al.; Ludwig et al.; Lagos-Quintana et al.), a comprehensive study of the dynamics of such tissue-specific miRNAs across mouse development was lacking. Our analysis fills this knowledge gap. We show that most of the tissue-specific miRNAs are dynamically regulated across development, with different subsets of miRNAs in the brain and heart expressed at different levels during embryonic development.

Finally, the clustering of the miRNAs based on the dynamics of their expression in different tissues allowed us for a unique opportunity to study the functionality and role of these miRNAs in a cooperative way. This approach revealed that some of these tissue specific clusters of miRNAs likely act as suppressors of genes involved in the development of other tissues than those in which the cluster of miRNAs are expressed. While the coregulation of the mRNAs could be simply due to the sharing of cis-regulatory elements, we note that many of the target genes of our miRNAs are transcription factors that are themselves important for mouse development, which strongly suggests that post-transcriptional regulation needs to be incorporated into models of transcriptional regulation being built from ChIP-seq, open chromatin, and mRNA expression data. The availability of miRNA expression levels in matching tissues and time points of the Mouse ENCODE dataset of embryonic development provides a unique opportunity to integrate the analysis of miRNAs with other functional genomic data used to build the Mouse Encyclopedia of DNA Elements.

FIGURES

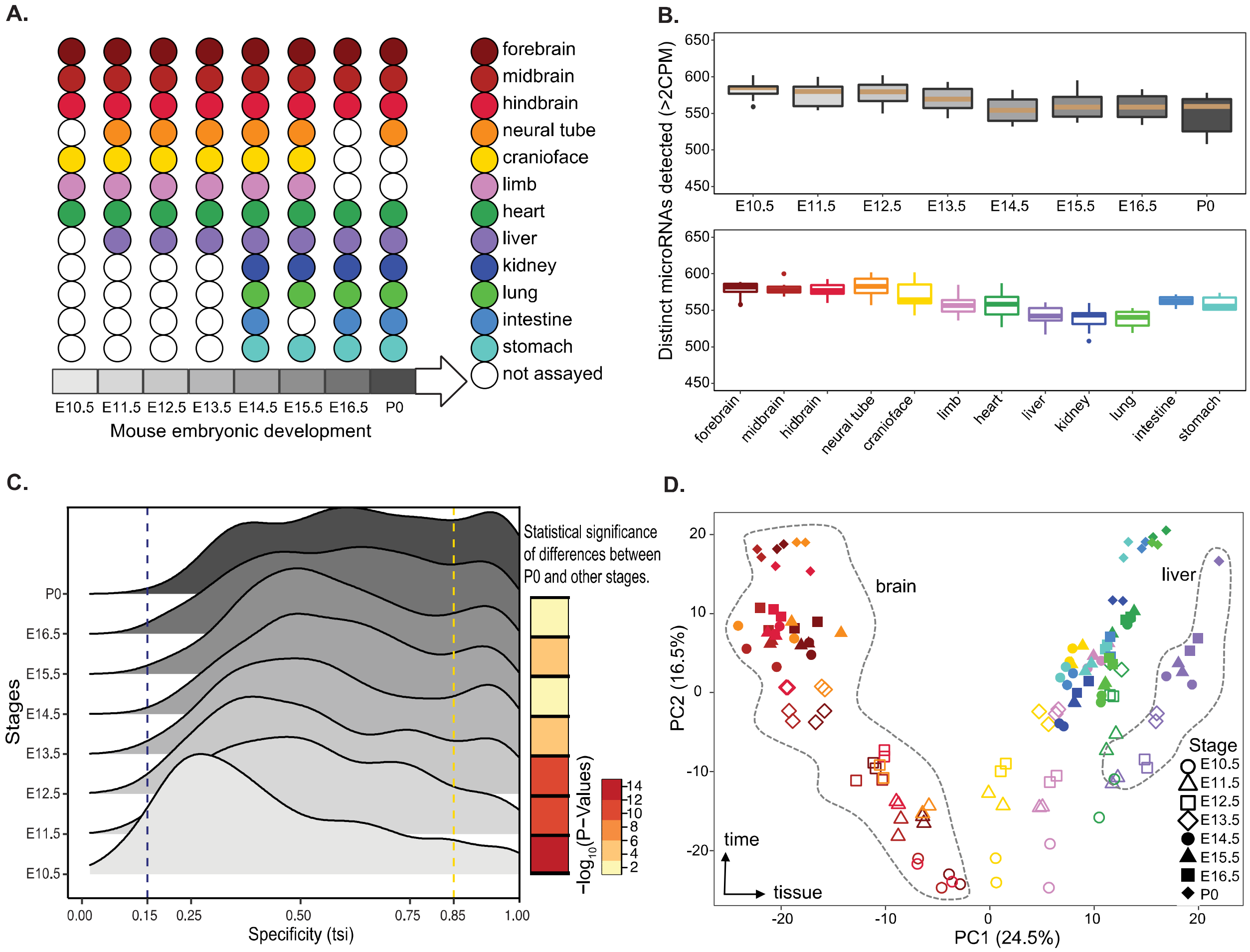

Figure 2.1: Overview of mouse ENCODE miRNA data sets. a) Representative major organ systems were profiled in a time course of mouse embryonic development. **b**) Number of distinct miRNAs detected in different organs and developmental stages (minimum 2 CPM). There are no significant differences between the number of miRNAs detected at different stages or within different organs. Developmental stage and organ colors correspond to Fig. 1A. **c**) The distribution of tissue specificity of miRNAs expressed at each developmental stage measured as tissue specific index (TSI). The miRNAs are significantly more tissue specific at stage of P0 compared to E10.5 (Kolmogorov–Smirnov (KS) test p-value < 2.2e-16). **d**) Principal component

analysis (PCA) of 12 mouse organs across 8 developmental stages. Organs are represented by various colors corresponding to Fig. 1A while shapes denote the different developmental stages.

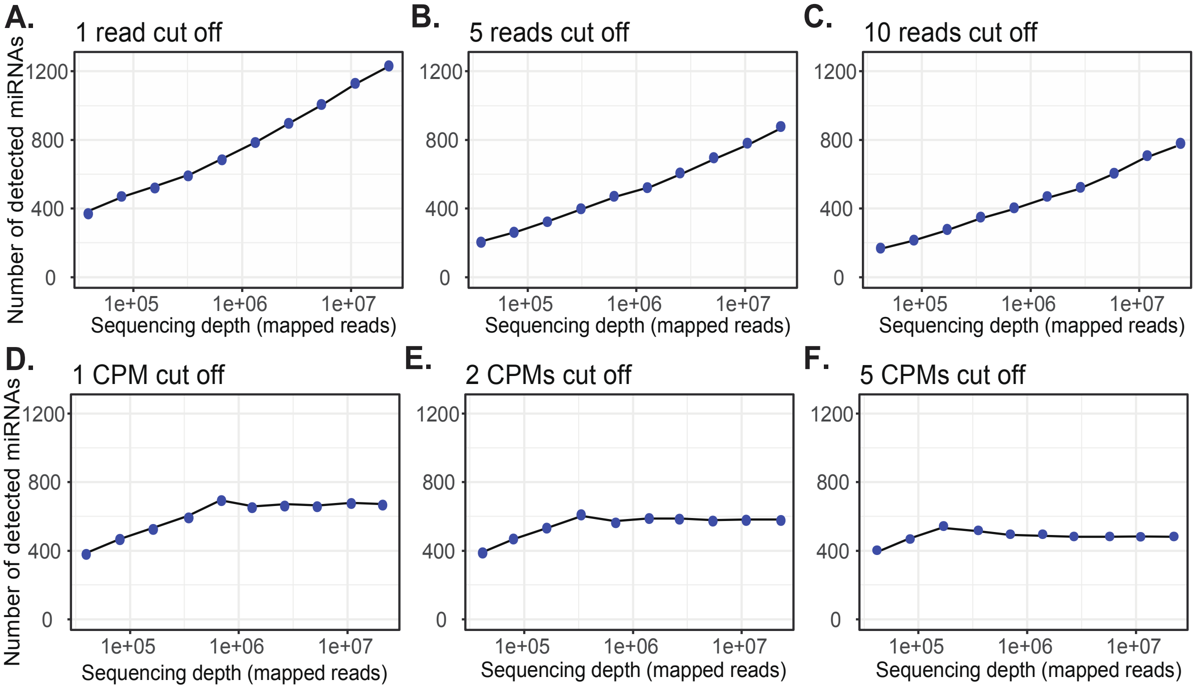

Figure 2.2: Sequencing depth analysis of microRNA-seq data. The deeply sequenced (~ 22 M mapped reads) heart E11.5 replicate 1 sample was subsampled to assess miRNA detection at different sequencing depths (11M, 5.4M, 2.7M, 1.4M, 680K, 340K, 160K, 80K and 40K) using different cutoffs: **a)** 1 read, **b)** 5 reads and **c)** 10 reads, **d)** 1 CPM , **e)** 1M for 2 CPMs and **f)** 5 CPMs.

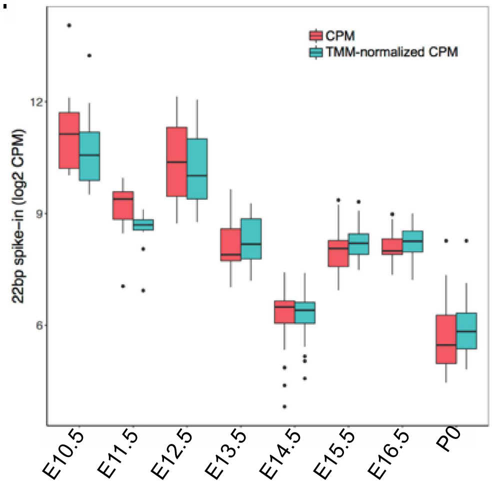

Figure 2.3: TMM normalization of miRNA expressions. Expression levels of the 22 bp spike-in in mouse microRNA-seq samples across different stages of embryonic development for non-normalized counts-per-million (CPM, red boxplots) and TMM-normalized CPMs (cyan boxplots).

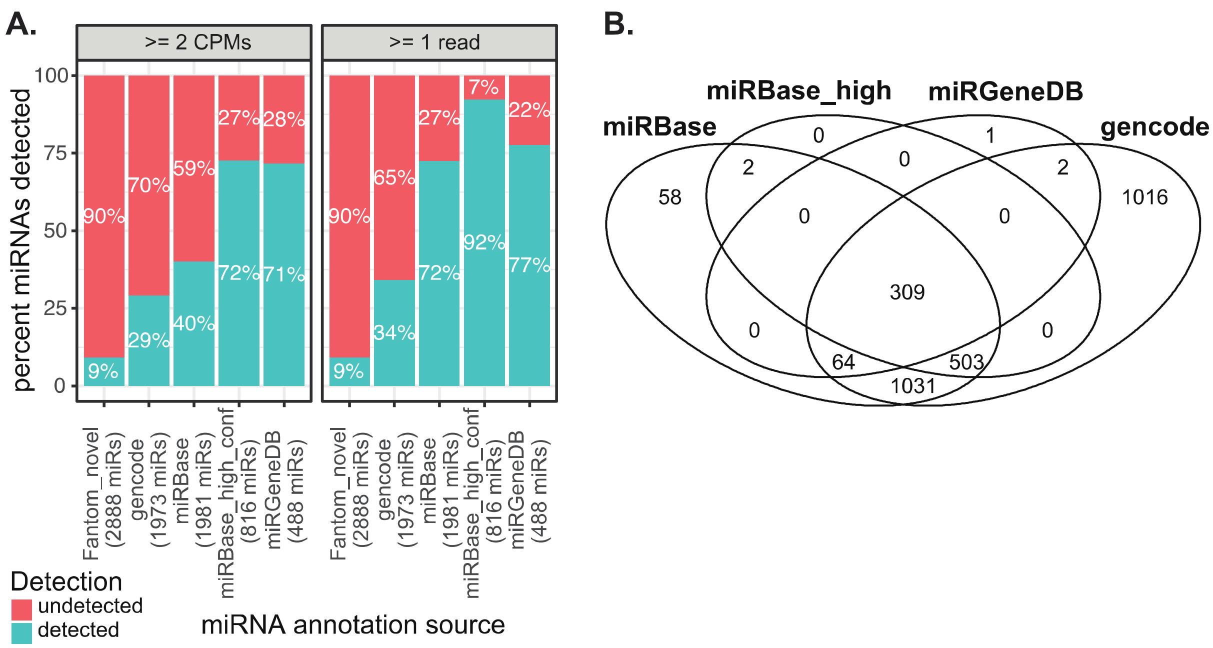

Figure 2.4: Coverage of miRNAs from different databases during mouse embryonic development. a) We mapped our microRNA-seq data to mature miRNAs from different sources such as miRbase, miRbase high confidence set, miRGeneDB and the Fantom 5 novel miRNA set. Our data includes more than 70% of miRNAs annotated in the high confidence set of miRBase and miRGeneDB at a minimum of 2 CPM for at least one of the organs at one of the developmental time points. These numbers were even higher when using a softer cut off of 1 read **b)** Overlap of annotated miRNAs in different databases.

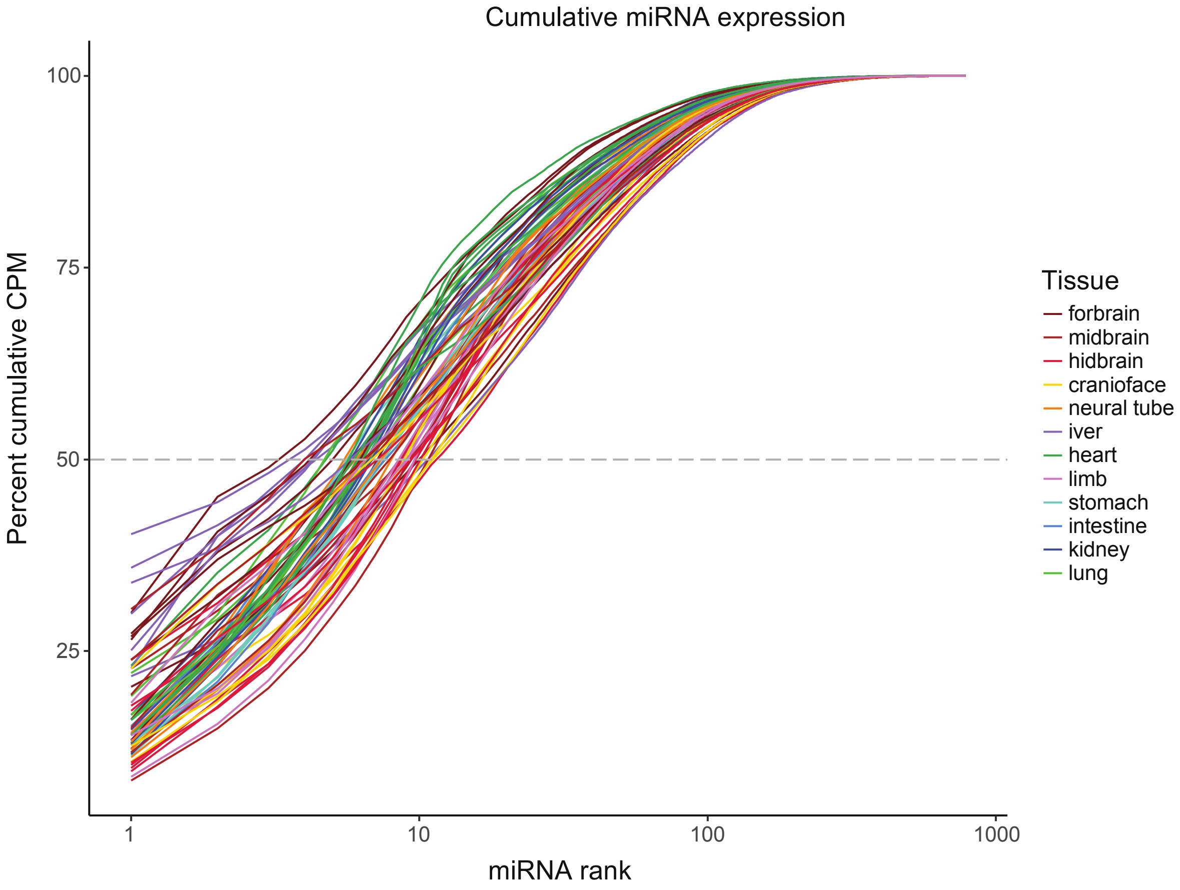

Figure 2.5: A few highly expressed miRNAs dominate miRNA-seq datasets. Cumulative distribution of sequencing reads, using miRNAs ranked by expression levels. The top 10 miRNAs account for more than half of the miRNA sequencing reads.

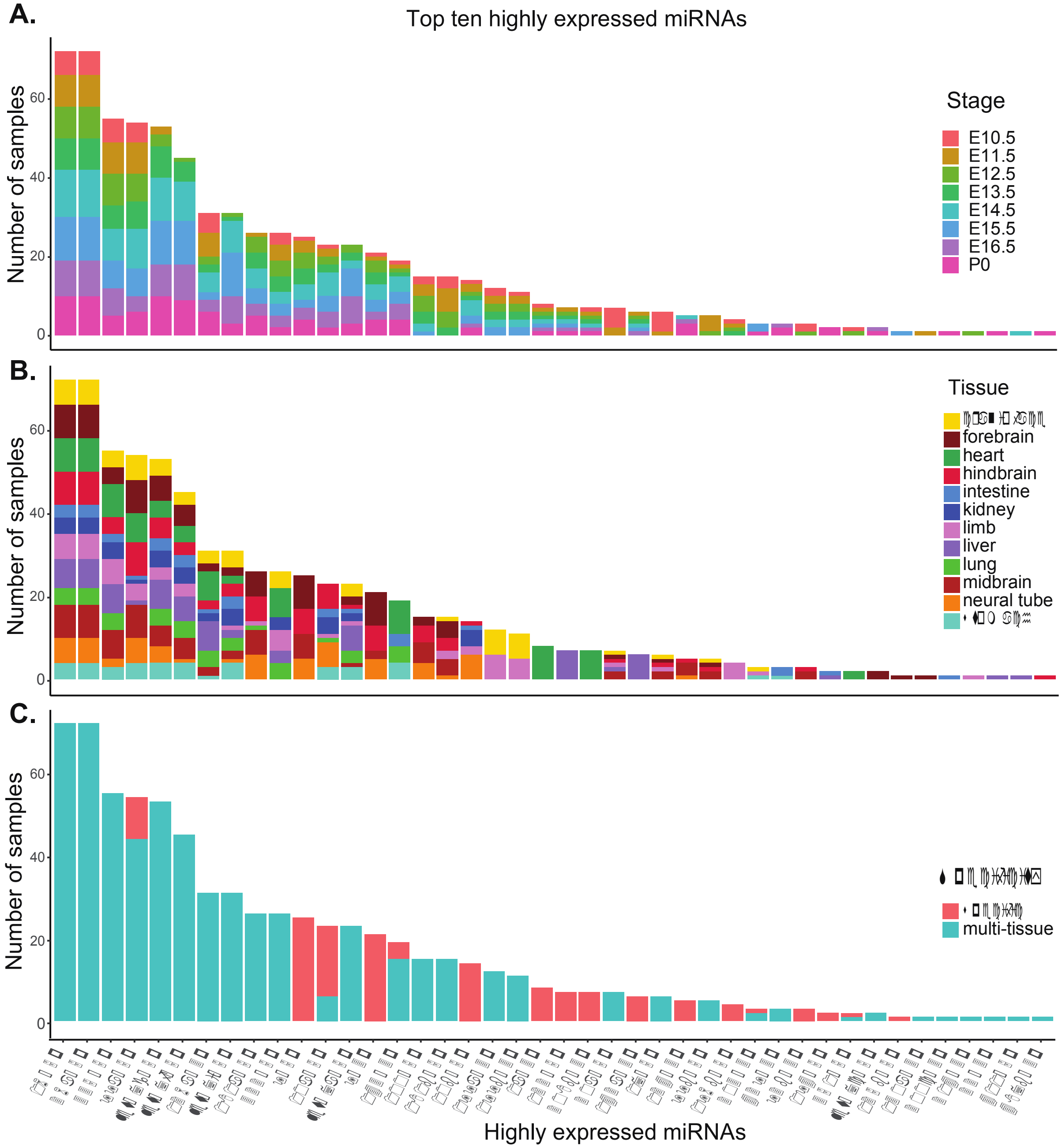

Figure 2.6: Characterization of the specificity of highly expressed miRNAs. Histograms of the 43 miRNAs that form the top ten highly expressed miRNAs in all the samples, ranked by the

number of samples they are highly expressed in; colored by their **a)** stage, **b)** organ, and **c)** specificity.

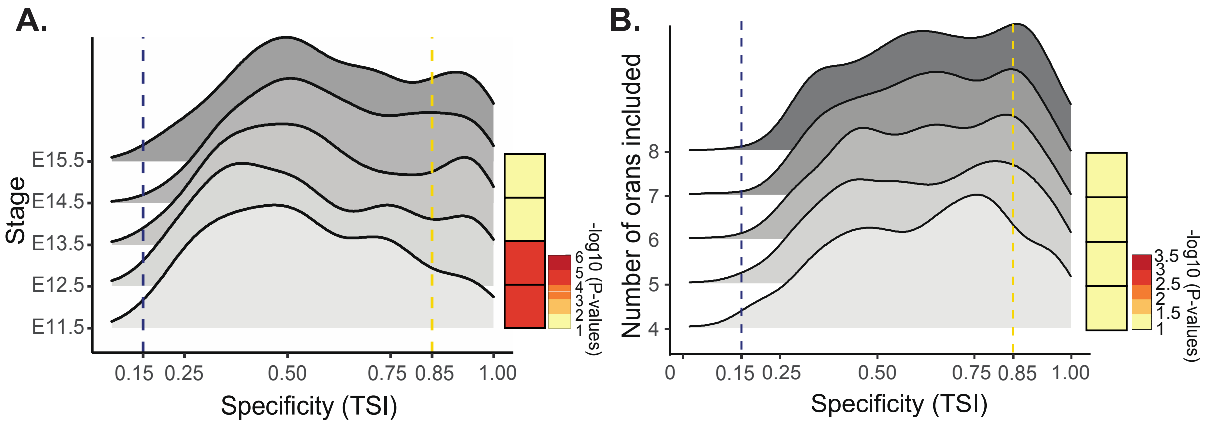

Figure 2.7: Alternative analysis of tissue specificity profiles. a) In order to ensure that the observed differences in the distribution of miRNA tissue specificity is not an artifact of different number of organs used for each stage, we looked at the different distributions of the five intermediate stages, constraining the calculation of TSI to the same organs (forebrain, heart, cranioface, liver and limb) and observed that there is still statistically significant differences between TSI distribution of earlier stages and later stages **b)** Then we looked at the TSI profile of P0 samples and re-calculated the TSI using different number of tissues, which did not result in any significant differences in the re-calculated distributions.

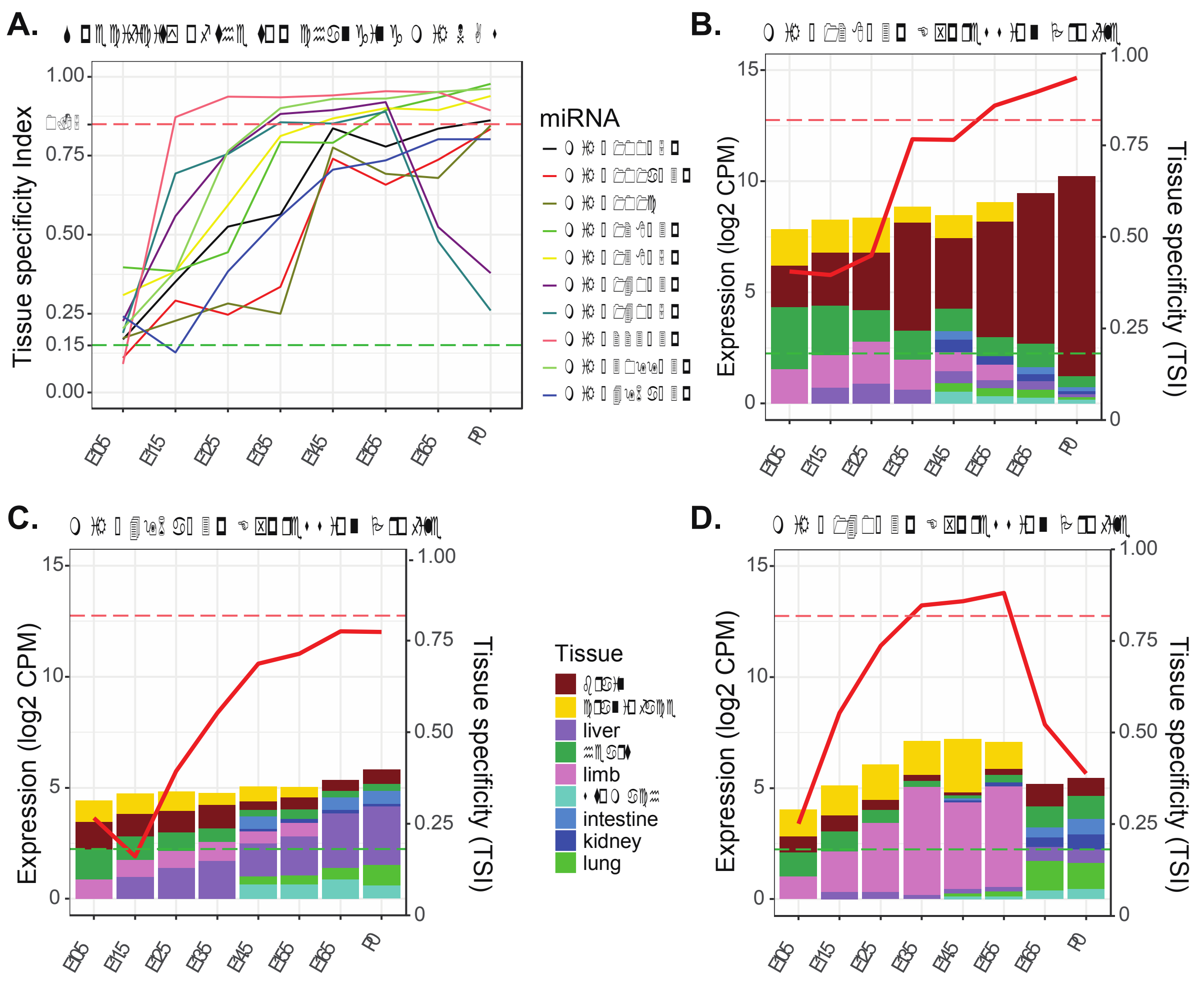

Figure 2.8: Analysis of miRNAs that change their specificity. a) The tissue specificity profile of ten miRNA with the most change in their specificity through the embryonic development. **(b-d)** Expression profile and bar graphs for three of these miRNAs: **b)** miR-128-3p, **c)** miR-496a-3p, **d)** miR-140-3p. The red curve traces the tissue specificity (TSI) on the second axis and each bar is colored proportional to the expression of the miRNA in different tissues. The fall in tissue specificity of miR-140-3p at the last two time point may be due to the lack of limb samples for those two time-point.

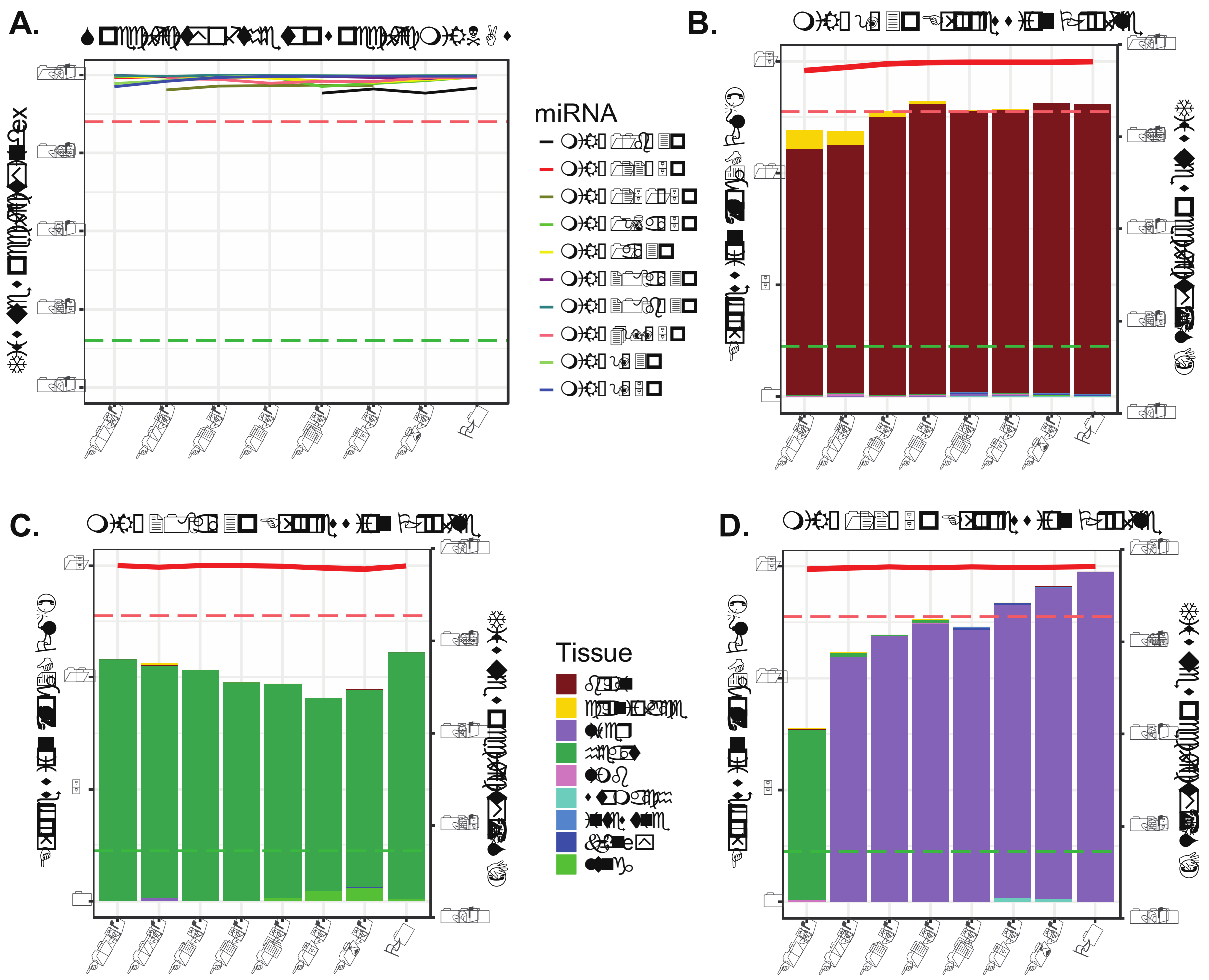

Figure 2.9: Analysis of miRNAs that are highly specific. a) The tissue specificity profile of ten miRNA with the highest specificity through the embryonic development. **(b-d)** The bar graphs show the expression profile of three of these miRNAs: **b)** miR-9-3p, **c)** miR-208a-3p, **d)** miR-122-5p. The red curve traces the tissue specificity (TSI) on the second axis and each bar is colored proportional to the expression of the miRNA in different tissues. miR-122-5p is highly specific at all stages but is shown as specific to heart at the E10.5 and liver specific at every other stage, possibly due to lack of a liver sample for stage E10.5.

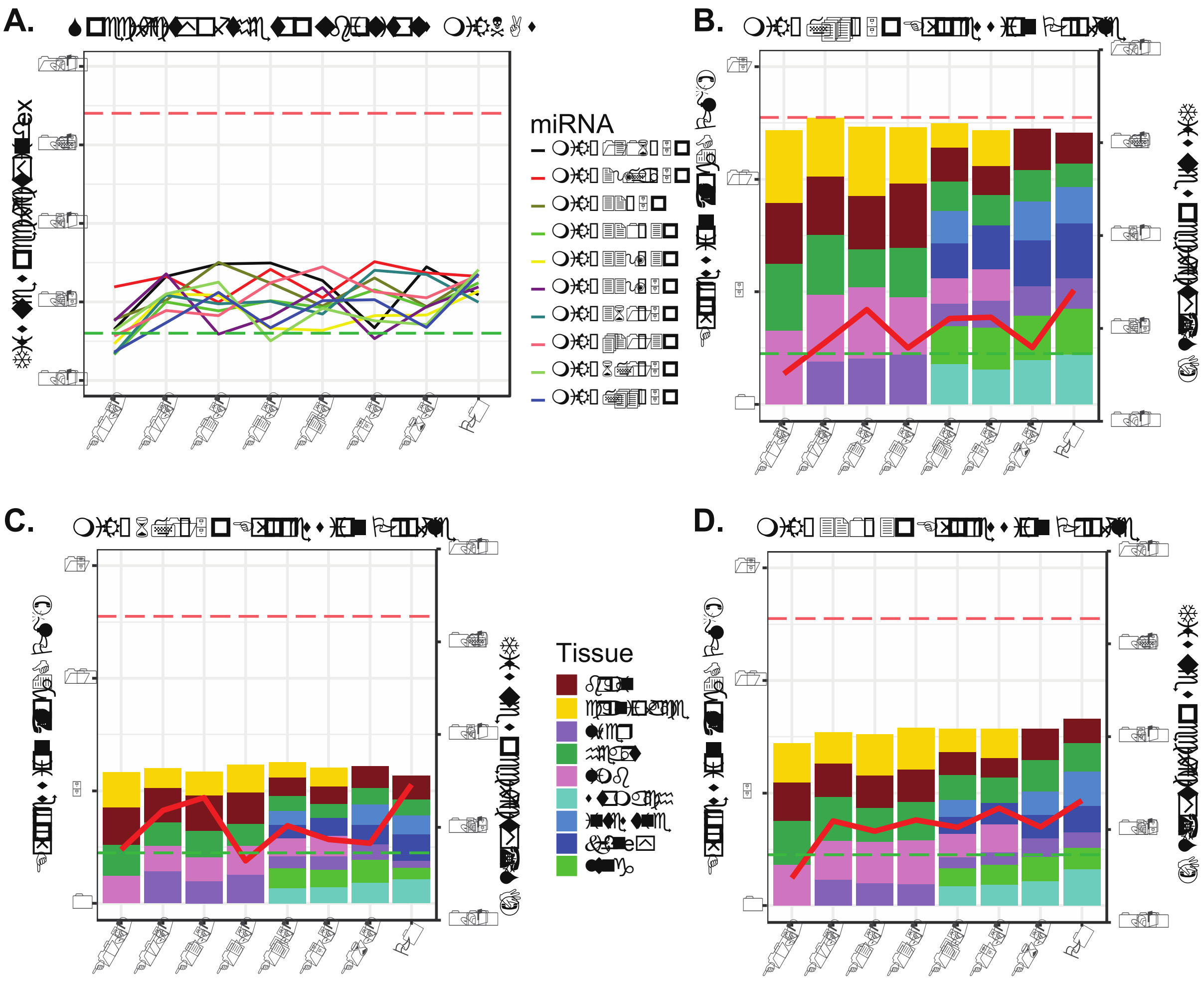

Figure 2.10: Analysis of miRNAs that are mostly ubiquitous. a) The tissue specificity profile of ten miRNA with the lowest specificity through the embryonic development. **(b-d)** The bar graphs show the expression profile of three of these miRNAs: **b)** miR-744-5p, **c)** miR-671-5p, **d)** miR-320-3p. The red curve traces the tissue specificity (TSI) on the second axis and each bar is colored proportional to the expression of the miRNA in different tissues.

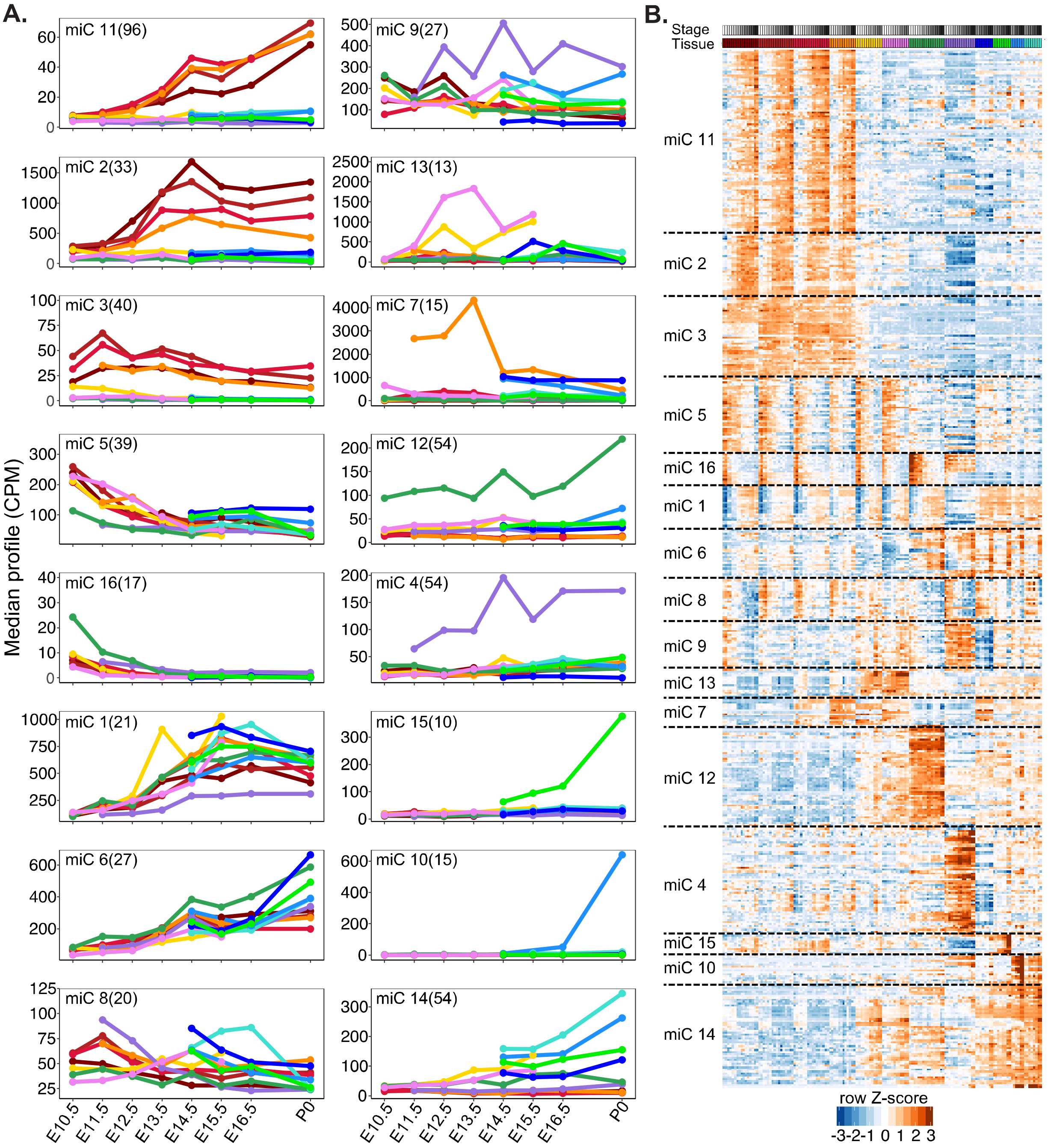

Figure 2.11: Clustering of mouse miRNAs during embryonic development time-course. a)

Clustering of miRNAs using maSigPro into 16 non-redundant groups based on median expression level of the miRNAs in each cluster. Organ colors correspond to Fig. 1A. **b**) Heatmap of the normalized expression levels (z-scores) of miRNAs in each cluster from Fig. 2A. Organ and stage colors correspond to Fig. 1A.

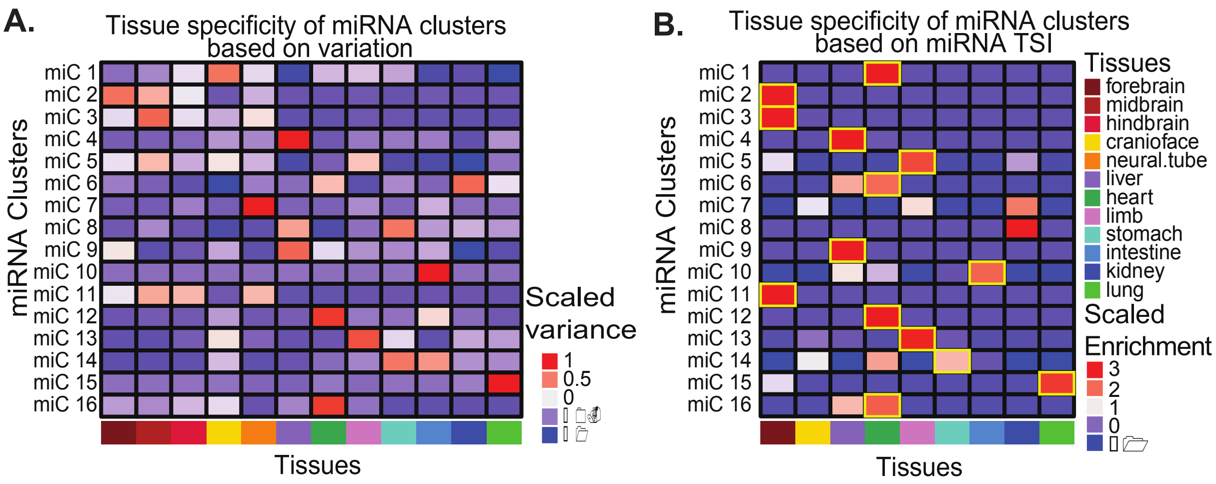

Figure 2.12: MicroRNA cluster tissue-specificity. **a**) Heatmap of the variance of miRNA cluster mean expression during the time course in each organ. These values were scaled for each cluster separately to identify the tissue specificity of that cluster. **b**) We further determined the tissue-specificity of the miRNA clusters by enrichment analysis of the tissue-specific miRNAs in each cluster. Tissue-specificity of the majority of the miRNA clusters correspond with what has been determined using the variance of the average miRNA expression of each cluster within each tissue, with gold boxed clusters indicating where the tissue specificities are concordant with cluster tissue specificity in panel A.

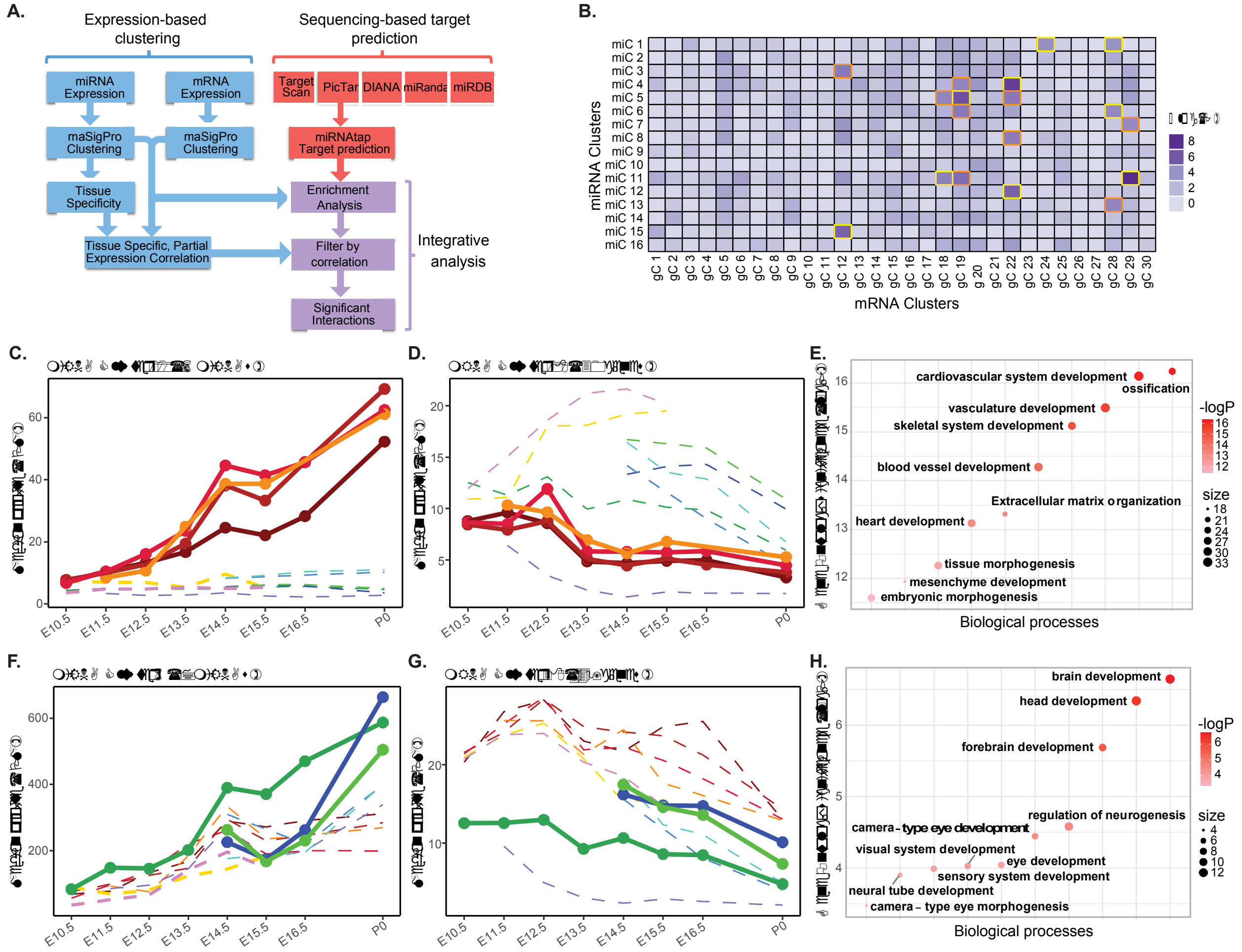

Figure 2.13: Identification of miRNA-mRNA cluster interactions. a) Potential targets of each miRNA cluster were obtained by applying an ensemble approach. Interactions were called as significant if they had a negative tissue-specific partial correlation and were enriched beyond the Bonferroni-corrected P-value of 10^{-4}. **b)** Heatmap of miRNA cluster target enrichment calculated using c^2 statistics. The 18 interactions identified as enriched are boxed in orange and gold. The interactions boxed in gold have negative partial correlation and are identified as significant interactions. **c)** miRNA cluster 11 corresponds to brain-specific miRNAs upregulated during development. **d)** mRNA cluster 18 genes are highly expressed in other organs such as limb, cranioface and heart. **e)** Gene ontology of miRNA cluster 11 targets in mRNA cluster 18 shows

enrichments in developmentally important genes with roles outside the brain. **f)** miRNA cluster 6 increases significantly during heart development. **g)** mRNA cluster 28 genes are over-expressed in brain. **h)** Gene ontology analysis of miRNA cluster 6 targets in mRNA cluster 28 revealed terms such as brain, head, and forebrain development.

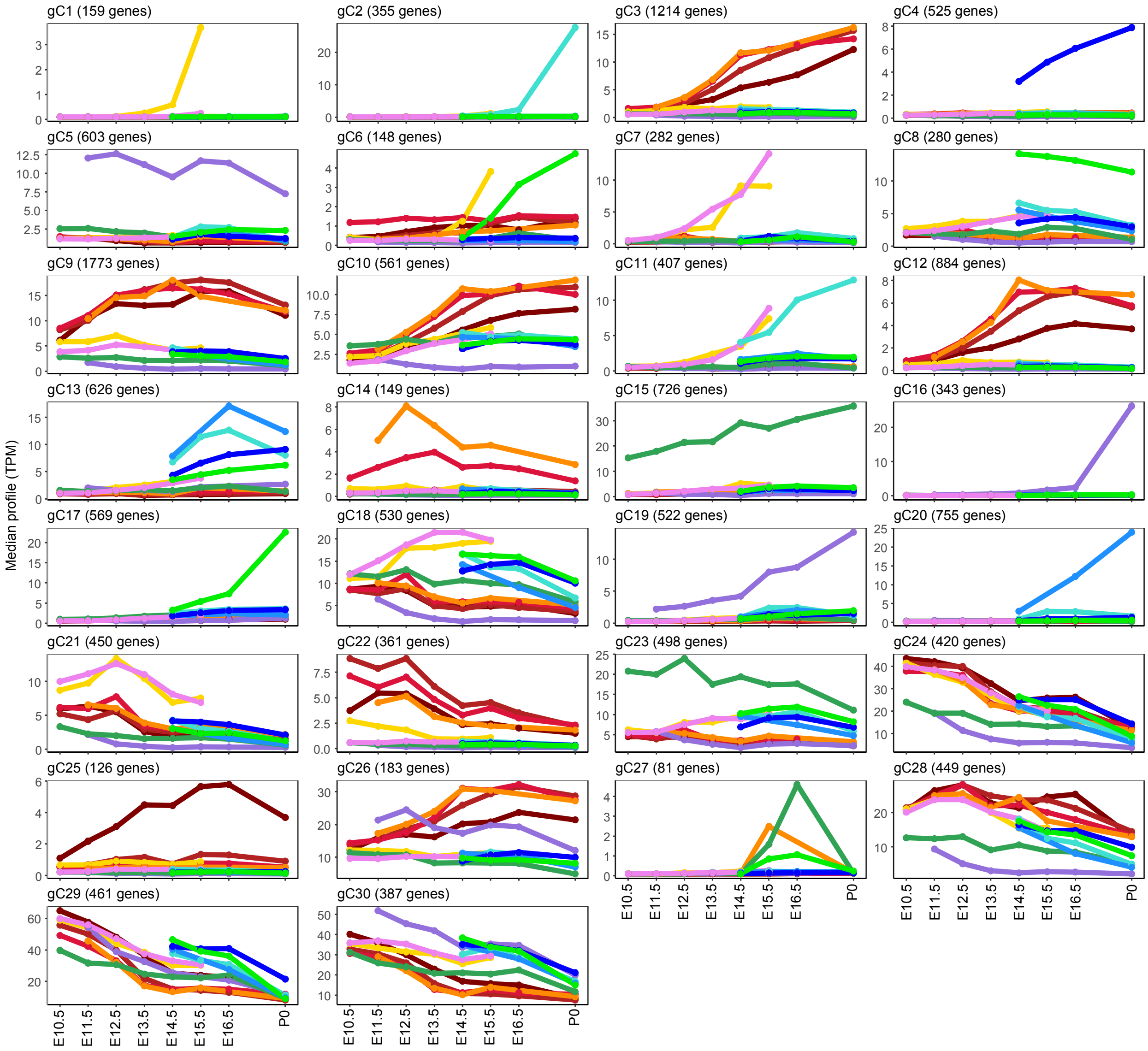

Figure 2.14: mRNA maSigPro cluster expression profiles. The median expression profiles of 30 mRNA clusters obtained by maSigPro.

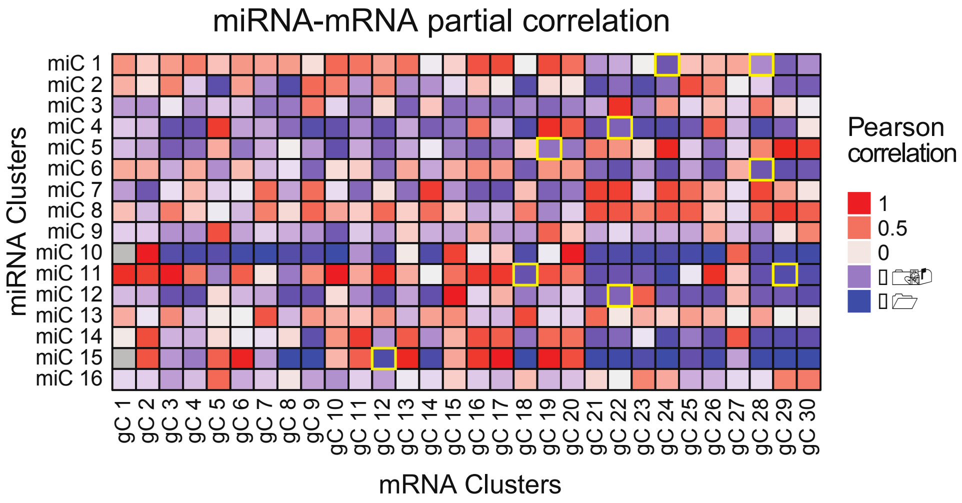

Figure 2.15: Expression anti-correlation analysis: Tissue-wise partial Pearson correlation between miRNAs clusters and each mRNA clusters identifies significant anti-correlations. The highlighted boxes indicate the significant interactions identified by target enrichment analysis and have a negative partial correlation.

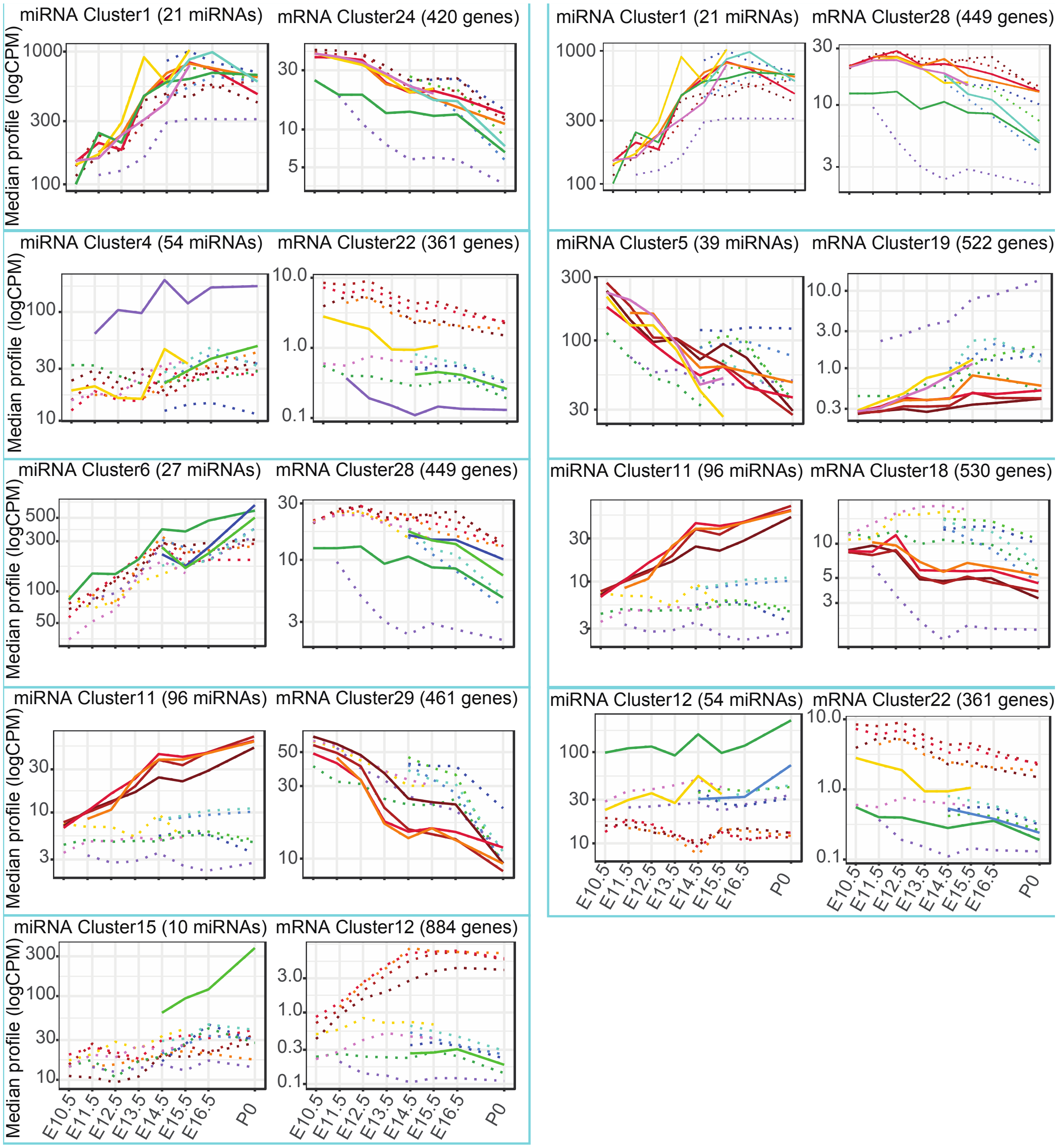

Figure 2.16: Expression profiles of all the significant interactions. The miRNA and mRNA expressions for the miRNA and mRNA clusters in the identified significant interactions.

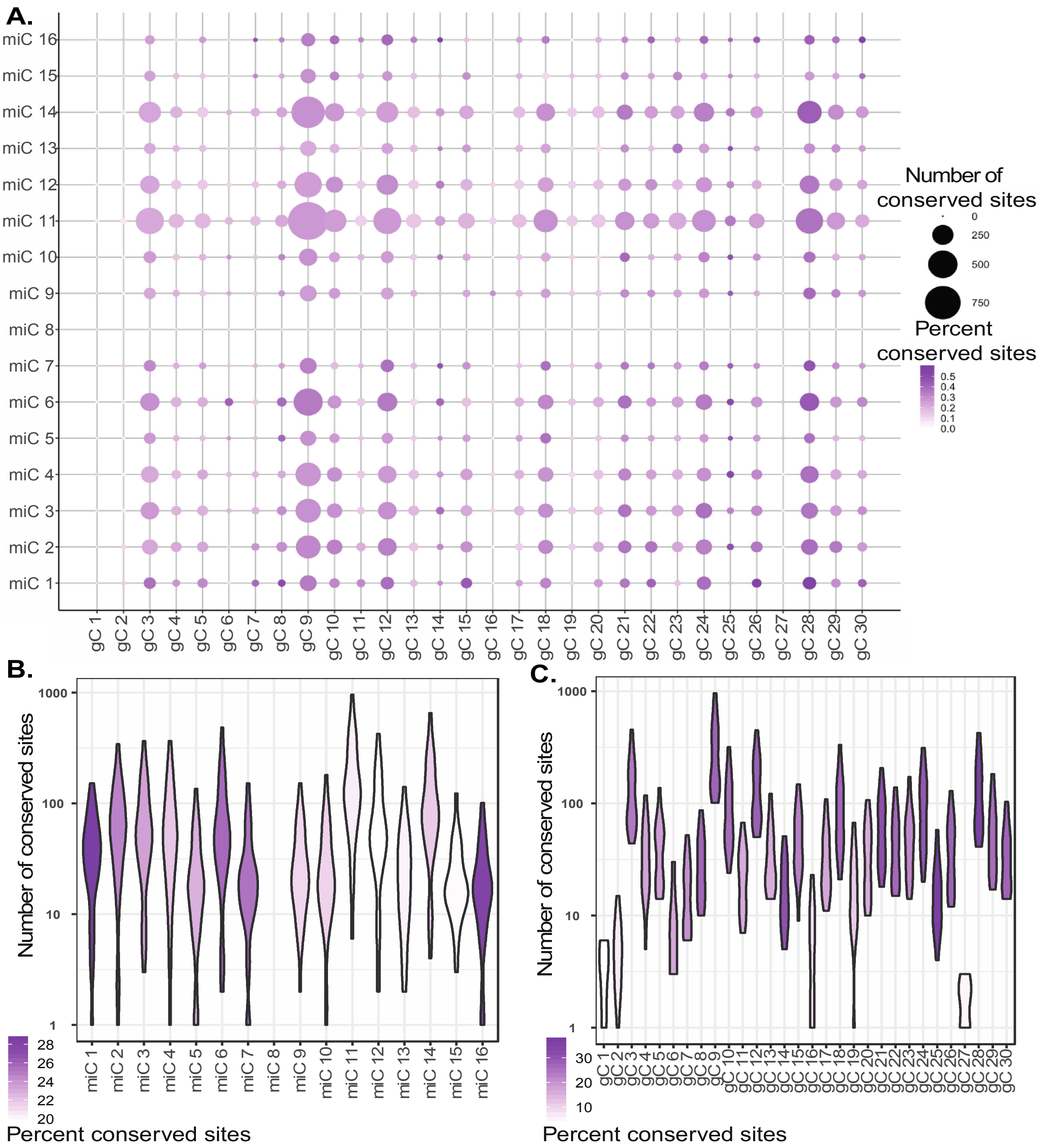

Figure 2.17: Conservation of microRNAs target sites. We studied the number of conserved target sites of conserved miRNAs in each miRNA cluster in the 3'UTRs of each gene cluster. A criterion of PhastCons score of minimum 0.9 for at least 4 nucleotides of the target site was used to call the site conserved. Panel **a)** shows the number of conserved miRNAs as the size of the points and the percentage of target sites conserved as the color (darker shade corresponds to the

higher conserved fraction) **b)** The analysis of conserved targets in gene clusters shows gene cluster 25 and 29 with the highest fraction of conserved targets while clusters 9 and 12 contains the highest number of conserved sites. c) On the other hand, miRNA clusters 1 and 16 has the highest fraction of conserved targets, whereas clusters 6,11 and 14 have the highest number of conserved targets.

TABLES

miRNA	cluster	miRNA	cluster	miRNA	cluster
mmu-let-7a-1-3p	1	mmu-miR-340-5p	2	mmu-miR-3074-1-3p	3
mmu-let-7a-5p	1	mmu-miR-344c-3p	2	mmu-miR-3081-3p	3
mmu-let-7b-3p	1	mmu-miR-488-3p	2	mmu-miR-3093-3p	3
mmu-let-7b-5p	1	mmu-miR-488-5p	2	mmu-miR-3093-5p	3
mmu-let-7c-2-3p	1	mmu-miR-6977-3p	2	mmu-miR-325-3p	3
mmu-let-7d-3p	1	mmu-miR-701-5p	2	mmu-miR-325-5p	3
mmu-let-7d-5p	1	mmu-miR-708-3p	2	mmu-miR-344e-3p	3
mmu-let-7e-5p	1	mmu-miR-708-5p	2	mmu-miR-505-5p	3
mmu-let-7f-1-3p	1	mmu-miR-7a-5p	2	mmu-miR-598-5p	3
mmu-let-7f-5p	1	mmu-miR-7b-5p	2	mmu-miR-670-3p	3
mmu-let-7g-5p	1	mmu-miR-9-3p	2	mmu-miR-670-5p	3
mmu-let-7i-3p	1	mmu-miR-9-5p	2	mmu-miR-760-3p	3
mmu-let-7i-5p	1	mmu-miR-935	2	mmu-miR-872-5p	3
mmu-let-7j	1	mmu-miR-99a-3p	2	mmu-miR-877-5p	3
mmu-miR-151-5p	1	mmu-miR-99a-5p	2	mmu-miR-92b-3p	3
mmu-miR-320-3p	1	mmu-miR-100-3p	3	mmu-miR-92b-5p	3
mmu-miR-467e-5p	1	mmu-miR-100-5p	3	mmu-miR-101a-3p	4
mmu-miR-669c-5p	1	mmu-miR-1198-5p	3	mmu-miR-101b-3p	4
mmu-miR-672-5p	1	mmu-miR-124-5p	3	mmu-miR-101c	4
mmu-miR-676-3p	1	mmu-miR-124b-3p	3	mmu-miR-12191-3p	4
mmu-miR-98-5p	1	mmu-miR-1251-3p	3	mmu-miR-122-3p	4
mmu-let-7c-1-3p	2	mmu-miR-1251-5p	3	mmu-miR-122-5p	4
mmu-let-7c-5p	2	mmu-miR-135a-1-3p	3	mmu-miR-126b-3p	4
mmu-miR-103-3p	2	mmu-miR-135a-2-3p	3	mmu-miR-127-3p	4
mmu-miR-12194-3p	2	mmu-miR-135b-3p	3	mmu-miR-127-5p	4
mmu-miR-125b-1-3p	2	mmu-miR-135b-5p	3	mmu-miR-129b-3p	4
mmu-miR-125b-5p	2	mmu-miR-153-5p	3	mmu-miR-129b-5p	4
mmu-miR-128-2-5p	2	mmu-miR-216a-3p	3	mmu-miR-136-3p	4
mmu-miR-137-5p	2	mmu-miR-216a-5p	3	mmu-miR-136-5p	4
mmu-miR-149-5p	2	mmu-miR-216b-3p	3	mmu-miR-142a-3p	4
mmu-miR-181a-1-3p	2	mmu-miR-216b-5p	3	mmu-miR-142a-5p	4
mmu-miR-181a-5p	2	mmu-miR-217-3p	3	mmu-miR-144-3p	4
mmu-miR-181b-5p	2	mmu-miR-217-5p	3	mmu-miR-144-5p	4
mmu-miR-181c-3p	2	mmu-miR-218-5p	3	mmu-miR-154-3p	4
mmu-miR-181d-5p	2	mmu-miR-219a-1-3p	3	mmu-miR-154-5p	4
mmu-miR-301a-3p	2	mmu-miR-219a-2-3p	3	mmu-miR-16-5p	4
mmu-miR-3078-5p	2	mmu-miR-219a-5p	3	mmu-miR-185-5p	4
mmu-miR-330-3p	2	mmu-miR-301b-3p	3	mmu-miR-1968-5p	4
mmu-miR-340-3p	2	mmu-miR-3067-5p	3	mmu-miR-1981-5p	4

miRNA	cluster	miRNA	cluster	miRNA	cluster
mmu-miR-223-3p	4	mmu-miR-18b-5p	5	mmu-miR-30d-3p	6
mmu-miR-223-5p	4	mmu-miR-1955-5p	5	mmu-miR-30d-5p	6
mmu-miR-292b-3p	4	mmu-miR-196b-5p	5	mmu-miR-30e-3p	6
mmu-miR-29a-3p	4	mmu-miR-1983	5	mmu-miR-30e-5p	6
mmu-miR-29b-3p	4	mmu-miR-19b-1-5p	5	mmu-miR-3105-3p	6
mmu-miR-300-3p	4	mmu-miR-19b-2-5p	5	mmu-miR-3105-5p	6
mmu-miR-3070-3p	4	mmu-miR-20a-5p	5	mmu-miR-425-5p	6
mmu-miR-3098-5p	4	mmu-miR-20b-3p	5	mmu-miR-511-3p	6
mmu-miR-329-3p	4	mmu-miR-20b-5p	5	mmu-miR-547-3p	6
mmu-miR-3544-3p	4	mmu-miR-296-3p	5	mmu-miR-10a-3p	7
mmu-miR-376a-3p	4	mmu-miR-301b-5p	5	mmu-miR-10a-5p	7
mmu-miR-376a-5p	4	mmu-miR-32-3p	5	mmu-miR-10b-3p	7
mmu-miR-376b-3p	4	mmu-miR-345-3p	5	mmu-miR-10b-5p	7
mmu-miR-376b-5p	4	mmu-miR-363-3p	5	mmu-miR-182-5p	7
mmu-miR-381-3p	4	mmu-miR-363-5p	5	mmu-miR-183-3p	7
mmu-miR-409-5p	4	mmu-miR-421-3p	5	mmu-miR-183-5p	7
mmu-miR-410-3p	4	mmu-miR-6399	5	mmu-miR-196a-1-3p	7
mmu-miR-434-3p	4	mmu-miR-674-3p	5	mmu-miR-196a-5p	7
mmu-miR-434-5p	4	mmu-miR-674-5p	5	mmu-miR-211-5p	7
mmu-miR-451a	4	mmu-miR-758-5p	5	mmu-miR-615-3p	7
mmu-miR-486a-3p	4	mmu-miR-7654-3p	5	mmu-miR-615-5p	7
mmu-miR-486a-5p	4	mmu-miR-872-3p	5	mmu-miR-741-3p	7
mmu-miR-486b-3p	4	mmu-miR-92a-1-5p	5	mmu-miR-96-3p	7
mmu-miR-486b-5p	4	mmu-miR-92a-3p	5	mmu-miR-96-5p	7
mmu-miR-494-3p	4	mmu-miR-99b-3p	5	mmu-miR-1187	8
mmu-miR-496a-3p	4	mmu-miR-99b-5p	5	mmu-miR-466a-3p	8
mmu-miR-5104	4	mmu-miR-107-3p	6	mmu-miR-466b-3p	8
mmu-miR-5114	4	mmu-miR-126a-3p	6	mmu-miR-466b-5p	8
mmu-miR-5123	4	mmu-miR-126a-5p	6	mmu-miR-466c-3p	8
mmu-miR-539-5p	4	mmu-miR-139-5p	6	mmu-miR-466e-3p	8
mmu-miR-7670-3p	4	mmu-miR-148b-5p	6	mmu-miR-466e-5p	8
mmu-miR-106a-5p	5	mmu-miR-191-5p	6	mmu-miR-466f	8
mmu-miR-106b-5p	5	mmu-miR-201-5p	6	mmu-miR-466f-5p	8
mmu-miR-1197-5p	5	mmu-miR-21a-3p	6	mmu-miR-466k	8
mmu-miR-125a-3p	5	mmu-miR-21a-5p	6	mmu-miR-466m-5p	8
mmu-miR-130a-3p	5	mmu-miR-22-3p	6	mmu-miR-466o-5p	8
mmu-miR-130b-3p	5	mmu-miR-22-5p	6	mmu-miR-466p-3p	8
mmu-miR-155-5p	5	mmu-miR-221-3p	6	mmu-miR-467d-5p	8
mmu-miR-17-3p	5	mmu-miR-222-3p	6	mmu-miR-6539	8
mmu-miR-17-5p	5	mmu-miR-3068-5p	6	mmu-miR-669a-5p	8
mmu-miR-181b-2-3p	5	mmu-miR-30a-3p	6	mmu-miR-669l-5p	8
mmu-miR-181d-3p	5	mmu-miR-30a-5p	6	mmu-miR-669m-5p	8
mmu-miR-18a-5p	5	mmu-miR-30c-2-3p	6	mmu-miR-669p-5p	8
mmu-miR-18b-3p	5	mmu-miR-30c-5p	6	mmu-miR-7688-5p	8

miRNA	cluster	miRNA	cluster	miRNA	cluster
mmu-miR-1193-3p	9	mmu-miR-125b-2-3p	11	mmu-miR-323-3p	11
mmu-miR-1193-5p	9	mmu-miR-1264-3p	11	mmu-miR-326-3p	11
mmu-miR-134-5p	9	mmu-miR-1264-5p	11	mmu-miR-328-3p	11
mmu-miR-1946b	9	mmu-miR-128-3p	11	mmu-miR-329-5p	11
mmu-miR-299a-3p	9	mmu-miR-129-1-3p	11	mmu-miR-330-5p	11
mmu-miR-299a-5p	9	mmu-miR-129-2-3p	11	mmu-miR-331-3p	11
mmu-miR-3072-3p	9	mmu-miR-129-5p	11	mmu-miR-331-5p	11
mmu-miR-337-5p	9	mmu-miR-1298-3p	11	mmu-miR-338-3p	11
mmu-miR-341-3p	9	mmu-miR-1298-5p	11	mmu-miR-338-5p	11
mmu-miR-341-5p	9	mmu-miR-132-3p	11	mmu-miR-342-3p	11
mmu-miR-369-3p	9	mmu-miR-132-5p	11	mmu-miR-342-5p	11
mmu-miR-376c-3p	9	mmu-miR-135a-5p	11	mmu-miR-344-3p	11
mmu-miR-377-5p	9	mmu-miR-137-3p	11	mmu-miR-344b-3p	11
mmu-miR-379-3p	9	mmu-miR-138-1-3p	11	mmu-miR-344d-3p	11
mmu-miR-379-5p	9	mmu-miR-138-2-3p	11	mmu-miR-344f-3p	11
mmu-miR-380-3p	9	mmu-miR-138-5p	11	mmu-miR-344f-5p	11
mmu-miR-409-3p	9	mmu-miR-139-3p	11	mmu-miR-346-5p	11
mmu-miR-410-5p	9	mmu-miR-146b-3p	11	mmu-miR-3475-3p	11
mmu-miR-411-3p	9	mmu-miR-146b-5p	11	mmu-miR-370-3p	11
mmu-miR-411-5p	9	mmu-miR-149-3p	11	mmu-miR-382-5p	11
mmu-miR-412-5p	9	mmu-miR-150-3p	11	mmu-miR-383-3p	11
mmu-miR-431-5p	9	mmu-miR-150-5p	11	mmu-miR-383-5p	11
mmu-miR-493-3p	9	mmu-miR-153-3p	11	mmu-miR-384-3p	11
mmu-miR-494-5p	9	mmu-miR-181c-5p	11	mmu-miR-384-5p	11
mmu-miR-540-3p	9	mmu-miR-1839-3p	11	mmu-miR-431-3p	11
mmu-miR-673-5p	9	mmu-miR-1839-5p	11	mmu-miR-433-3p	11
mmu-miR-679-3p	9	mmu-miR-1843a-5p	11	mmu-miR-448-3p	11
mmu-miR-12206-5p	10	mmu-miR-1843b-5p	11	mmu-miR-448-5p	11
mmu-miR-192-3p	10	mmu-miR-186-5p	11	mmu-miR-487b-3p	11
mmu-miR-192-5p	10	mmu-miR-187-3p	11	mmu-miR-544-3p	11
mmu-miR-194-1-3p	10	mmu-miR-1912-3p	11	mmu-miR-544-5p	11
mmu-miR-194-2-3p	10	mmu-miR-1969	11	mmu-miR-551b-5p	11
mmu-miR-194-5p	10	mmu-miR-204-3p	11	mmu-miR-592-5p	11
mmu-miR-203b-3p	10	mmu-miR-204-5p	11	mmu-miR-598-3p	11
mmu-miR-215-3p	10	mmu-miR-212-3p	11	mmu-miR-6540-3p	11
mmu-miR-215-5p	10	mmu-miR-212-5p	11	mmu-miR-6540-5p	11
mmu-miR-3073b-5p	10	mmu-miR-29b-2-5p	11	mmu-miR-666-3p	11
mmu-miR-30b-5p	10	mmu-miR-29c-3p	11	mmu-miR-666-5p	11
mmu-miR-31-3p	10	mmu-miR-300-5p	11	mmu-miR-668-3p	11
mmu-miR-31-5p	10	mmu-miR-3059-3p	11	mmu-miR-668-5p	11
mmu-miR-582-5p	10	mmu-miR-3059-5p	11	mmu-miR-672-3p	11
mmu-miR-98-3p	10	mmu-miR-3061-5p	11	mmu-miR-770-3p	11
mmu-miR-1224-5p	11	mmu-miR-3099-3p	11	mmu-miR-770-5p	11
mmu-miR-124-3p	11	mmu-miR-3106-5p	11	mmu-miR-7a-1-3p	11

miRNA	cluster	miRNA	cluster	miRNA	cluster
mmu-miR-7a-2-3p	11	mmu-miR-501-5p	12	mmu-miR-200b-5p	14
mmu-miR-7b-3p	11	mmu-miR-503-3p	12	mmu-miR-200c-3p	14
mmu-miR-873a-3p	11	mmu-miR-503-5p	12	mmu-miR-200c-5p	14
mmu-miR-873a-5p	11	mmu-miR-504-5p	12	mmu-miR-203-3p	14
mmu-miR-874-3p	11	mmu-miR-532-3p	12	mmu-miR-203-5p	14
mmu-miR-879-5p	11	mmu-miR-532-5p	12	mmu-miR-205-3p	14
mmu-miR-1231-5p	12	mmu-miR-542-3p	12	mmu-miR-205-5p	14
mmu-miR-133a-3p	12	mmu-miR-542-5p	12	mmu-miR-210-3p	14
mmu-miR-133a-5p	12	mmu-miR-6353	12	mmu-miR-210-5p	14
mmu-miR-133b-3p	12	mmu-miR-652-3p	12	mmu-miR-224-5p	14
mmu-miR-151-3p	12	mmu-miR-675-3p	12	mmu-miR-23a-3p	14
mmu-miR-188-5p	12	mmu-miR-675-5p	12	mmu-miR-23a-5p	14
mmu-miR-190a-5p	12	mmu-miR-676-5p	12	mmu-miR-23b-3p	14
mmu-miR-1a-1-5p	12	mmu-miR-700-3p	12	mmu-miR-24-1-5p	14
mmu-miR-1a-3p	12	mmu-miR-7083-5p	12	mmu-miR-24-2-5p	14
mmu-miR-208a-3p	12	mmu-miR-7689-3p	12	mmu-miR-24-3p	14
mmu-miR-208a-5p	12	mmu-miR-133b-5p	13	mmu-miR-26a-1-3p	14
mmu-miR-208b-3p	12	mmu-miR-140-3p	13	mmu-miR-26a-2-3p	14
mmu-miR-208b-5p	12	mmu-miR-140-5p	13	mmu-miR-26a-5p	14
mmu-miR-30c-1-3p	12	mmu-miR-199a-3p	13	mmu-miR-26b-3p	14
mmu-miR-3102-5p.2-5p	12	mmu-miR-199a-5p	13	mmu-miR-26b-5p	14
mmu-miR-322-3p	12	mmu-miR-199b-3p	13	mmu-miR-27a-3p	14
mmu-miR-322-5p	12	mmu-miR-199b-5p	13	mmu-miR-27a-5p	14
mmu-miR-335-3p	12	mmu-miR-206-3p	13	mmu-miR-27b-3p	14
mmu-miR-335-5p	12	mmu-miR-214-3p	13	mmu-miR-27b-5p	14
mmu-miR-351-3p	12	mmu-miR-214-5p	13	mmu-miR-28a-3p	14
mmu-miR-351-5p	12	mmu-miR-3095-3p	13	mmu-miR-28a-5p	14
mmu-miR-362-5p	12	mmu-miR-452-3p	13	mmu-miR-34a-5p	14
mmu-miR-378a-3p	12	mmu-miR-652-5p	13	mmu-miR-365-3p	14
mmu-miR-378a-5p	12	mmu-miR-141-3p	14	mmu-miR-375-3p	14
mmu-miR-378b	12	mmu-miR-141-5p	14	mmu-miR-429-3p	14
mmu-miR-378c	12	mmu-miR-143-3p	14	mmu-miR-455-3p	14
mmu-miR-378d	12	mmu-miR-143-5p	14	mmu-miR-455-5p	14
mmu-miR-450a-1-3p	12	mmu-miR-145a-3p	14	mmu-miR-490-3p	14
mmu-miR-450a-2-3p	12	mmu-miR-145a-5p	14	mmu-miR-490-5p	14
mmu-miR-450a-5p	12	mmu-miR-145b	14	mmu-miR-574-3p	14
mmu-miR-450b-3p	12	mmu-miR-147-3p	14	mmu-miR-574-5p	14
mmu-miR-450b-5p	12	mmu-miR-148a-3p	14	mmu-miR-802-3p	14
mmu-miR-483-3p	12	mmu-miR-148a-5p	14	mmu-miR-802-5p	14
mmu-miR-483-5p	12	mmu-miR-152-3p	14	mmu-miR-146a-5p	15
mmu-miR-499-3p	12	mmu-miR-193a-5p	14	mmu-miR-195a-3p	15
mmu-miR-499-5p	12	mmu-miR-200a-3p	14	mmu-miR-195a-5p	15
mmu-miR-500-3p	12	mmu-miR-200a-5p	14	mmu-miR-34b-3p	15
mmu-miR-501-3p	12	mmu-miR-200b-3p	14	mmu-miR-34b-5p	15

miRNA	cluster	miRNA	cluster	miRNA	cluster
mmu-miR-34c-3p	15	mmu-miR-291b-5p	16	mmu-miR-302a-5p	16
mmu-miR-34c-5p	15	mmu-miR-292a-3p	16	mmu-miR-302b-3p	16
mmu-miR-449a-5p	15	mmu-miR-292a-5p	16	mmu-miR-302b-5p	16
mmu-miR-449c-5p	15	mmu-miR-293-3p	16	mmu-miR-302c-3p	16
mmu-miR-497a-5p	15	mmu-miR-294-3p	16	mmu-miR-302d-3p	16
mmu-miR-291a-3p	16	mmu-miR-294-5p	16	mmu-miR-367-3p	16
mmu-miR-291a-5p	16	mmu-miR-295-3p	16		
mmu-miR-291b-3p	16	mmu-miR-302a-3p	16		

Table 2.1: List of differentially expressed miRNAs in mouse.

	miC1	miC2	miC3	miC4	miC5	miC6	miC7	miC8	miC9	miC10	miC11	miC12	miC13	miC14	miC15	miC16
forebrain	0	1	1	0	1	0	0	0	1	0	1	0	0	0	0	1
midbrain	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0
hindbrain	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1
craniofacee	1	0	1	1	1	0	0	0	1	0	0	1	1	1	0	1
neural.tube	1	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0
liver	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0
heart	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1
limb	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
stomach	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0
intestine	0	0	0	0	0	0	1	1	1	1	0	1	0	1	0	0
kidney	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0
lung	0	0	0	1	0	1	0	1	0	0	0	0	1	0	1	0

Table 2.2: MicroRNA clusters' tissue specificity values (scaled for each cluster).

Significant Int.	***I1***	***I2***	***I3***	***I4***	***I5***	***I6***	***I7***	***I8***	***I9***	***I10***	***I11***	***I12***	***I13***	***I14***	***I15***	***I16***	***I17***	***I18***
miRNA cluster	3	15	5	11	4	5	6	11	4	5	8	12	1	1	6	13	7	11
gene cluster	12	12	18	18	19	19	19	19	22	22	22	22	24	28	28	28	29	29

Table 2.3: List of all the significant miRNA-mRNA cluster interactions.

METHODS

Mouse tissue collection and total RNA isolation

For each of the embryonic stages assayed (Fig. 1A), a single pregnant female was euthanized and dissected for embryo removal. Tissues from embryos and a newborn mouse at day 0 were collected. Detailed protocol of tissue collection can be accessed at:

https://www.encodeproject.org/documents/631aa21c-8e48-467e-8cac-d40c875b3913/@@download/attachment/StandardTissueExcisionProtocol_02132017.pdf

Total RNA was extracted from each sample using mirVana miRNA isolation kit (Thermo Fisher Scientific Cat. #AM1561), followed by genomic DNA removal using TURBO DNA-free kit (Thermo Fisher Scientific Cat. #AM1907).

MicroRNA profiling of mouse embryonic and postnatal samples

Mouse microRNA-seq libraries were constructed by following the microRNA-seq protocol described previously (Alon et al.; Roberts et al.) without the blocking of highly abundant miRNAs and with some minor modifications. The library concentrations were measured by Library Quantification Kit (KAPA Biosystems Cat. #KK4824). The library loading concentration for the sequencing was determined using the concentration obtained by KAPA and the estimated fragment size of 140 bp (since we could not use bioanalyzer to determine the fragment sizes of these libraries). The microRNA-seq libraries were sequenced as 50 bp single-end reads on an Illumina HiSeq 2000 sequencer.

Mouse microRNA-seq read adapter trimming and mapping

We used Cutadapt v. 1.7.1 (Martin) with Python 2.7.10 to sequentially trim 5' and 3' adapters from raw reads. Trimmed reads were mapped to mouse miRBase v. 22 (Kozomara and Griffiths-Jones) mature miRNA sequences with STAR v. 2.4.2a (Dobin et al.). Counts of reads mapping to each miRBase mature miRNA were obtained from STAR output. The counts were normalized for sequencing depth and further TMM normalized using edgeR (Robinson et al.) to obtain counts per million (CPM). Furthermore, for cross-referencing with the high-confidence subset of miRNAs in miRBase (Kozomara et al., 2013), the miRNAs in MiRGeneDB (Fromm et al., 2018) and FANTOM project novel miRNAs (De Rie et al., 2017), the trimmed reads were mapped to their corresponding mature miRNA reference files and quantified using STAR v. 2.4.2a.

Tissue specificity analysis of individual miRNAs

MicroRNA tissue specificity was determined using a tissue specificity index as previously described (Ludwig et al.):

$$tsi_j = \frac{\sum_{i=1}^{N}(1 - x_{j,i})}{N - 1}$$

In order to prevent any biases introduced by multiple tissues of neural origin, we excluded the samples from hindbrain, midbrain and neural tube and used only the forebrain samples for tissue specificity study of individual miRNAs. Also since some of the organs were not assayed at the earlier stages of embryonic development (due to the fact that they start development later), we decided to restrict the number of organs considered for TSI calculations to cranioface, forebrain, heart, limb, and liver for the first 4 stages (E10.5-E13.5) and to cranioface, forebrain, heart, limb, liver, stomach, kidney, lung and intestine for the last 4 stages (E14.5-P0). Alternative methods for this calculation were employed as described in Fig. 7.

Clustering of mouse microRNA-seq data

Time-series analysis of the mouse microRNA-seq time-course was performed using maSigPro v. 1.48.0 (Nueda et al.) in R 3.4.4 (R Core Team, 2018). Briefly, each organ (12 in total) that were assayed in at least two developmental time points were analyzed using a degree 3 and maSigPro functions “p.vector(data, design = design.matrix, counts = TRUE)”, “T.fit(p.vector_output, alfa = 0.01)”, and “get.siggenes(T.fir_output, rsq=0.7, vars="all")”. Different numbers of clusters (k) were tested to obtain a robust clustering of miRNAs by comparing the clusters at each step of k with the previous ones using the command “see.genes(get$sig.genes, k = …)”. The best clustering of miRNAs was obtained with k=16. The median profiles of the genes were plotted using ggplot2 package (Wickham) in R. The R code used to generate these clusters and plots can be found at:

< https://github.com/sorenar/mouse_embryonic_miRNAs/blob/master/miRNA_maSigPro.R>

Analysis of the mouse mRNA-seq data

mRNA-seq reads were mapped to the mouse genome (assembly mm10) using STAR v. 2.4.2a. The alignments to the genome were assembled into ab initio transcripts using StringTie v. 1.2.4 (Pertea et al.). The expression levels of the GENCODE v. M10 and the StringTie model transcripts were obtained using STAR and RSEM v. 1.2.25 (Li and Dewey).

Time-series analysis of the mouse mRNA-seq time-course was performed similar to the clustering of miRNAs. In this case higher number of clusters (k = 20-35) were tested and k=30 was selected as the number that gave the best results. Similarly, the median profiles of these clusters were plotted using ggplot2 in R.

Target enrichment analysis

The R package, miRNAtap v. 1.10.0 (Pajak & Simpson 2018) was used as an ensemble method to compile the predicted targets for each miRNA in our data. miRNAtap uses five different sources to generate a list of predicted targets: TargetScan (Friedman et al.), DIANA (Maragkakis et al.), miRanda (Enright et al.), PicTar (Lall et al.), and miRDB (Wong and Wang). We used getPredictedTargets(miRNA, species = 'mmu',method = 'geom',min_src = 3) to obtain the list of predicted targets for miRNA. The parameter “min_src” indicates that if the miRNA has targets that are present in more than “min_src” value, the reported list would be only limited to those targets, otherwise the method will reduce the “min_src” until it gets a list of targets or no target at all.

For each significant interaction with a negative partial correlation, the list of the target genes in the interaction was compiled. The gene ontology analysis of each of these target lists was performed via Metascape (Zhou et al.) and the top ten most enriched terms were plotted for these analysis.

Identification of tissue specificity of the miRNA clusters:

The average expression of miRNAs in each cluster was calculated for each sampling point (for each tissue at each of the time points). Then the standard deviation of each cluster was calculated for each tissue across different time points. The standard deviations obtained for different tissues were then scaled for each miRNA cluster and the tissues with positive values were considered as the tissue specificity of the miRNA cluster. The code to create this tissue specificity matrix and the corresponding plot is available at:

< https://github.com/sorenar/mouse_embryonic_miRNAs/blob/master/PartialCorrelation.R >

Building the partial correlation matrix:

The average expression of mRNAs in each cluster was calculated for each tissue at all stages. For each pairs of miRNA-mRNA clusters only the sample points corresponding to the tissues identified as specific to the miRNA cluster were used to find the Pearson correlation. The code to generate this partial correlation matrix is provided at: < https://github.com/sorenar/mouse_embryonic_miRNAs/blob/master/PartialCorrelation.R >

Conservation analysis of microRNA targets

In order to investigate the targeting of gene clusters by the miRNA clusters, we looked at the conservation of miRNA target sites within 3'UTR of each gene cluster. We first identified the 8mer seeds of all the miRbase mature miRNAs and looked for their corresponding target sites in the 3'UTRs (with perfect complementarity). We used the PhastCons scores of the Euarchontoglires subset of mm10 multi-alignments (downloaded from UCSC Genome Browser) to determine the conservation of these target seeds. We extracted the 3' UTR regions with the PhastCons score of more than 0.9 and then counted the target sites that have more than 4 nucleotides within that region as conserved and the remaining target sites as non-conserved. Finally, we plotted the total number of conserved target sites and the fraction of target sites conserved for each microRNA-gene cluster pair.

REFERENCES

Alisch, Reid S., et al. “Argonaute2 Is Essential for Mammalian Gastrulation and Proper Mesoderm Formation.” *PLoS Genetics*, vol. 3, no. 12, Public Library of Science ({PLoS}), 2007, pp. 2565–71, doi:10.1371/journal.pgen.0030227.

Alon, Shahar, et al. “Barcoding Bias in High-Throughput Multiplex Sequencing of MiRNA.” *Genome Research*, vol. 21, no. 9, Cold Spring Harbor Laboratory, July 2011, pp. 1506–11, doi:10.1101/gr.121715.111.

Bartel, David P. “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.” *Cell*, vol. 116, no. 2, 2004, pp. 281–97, doi:10.1016/S0092-8674(04)00045-5.

---. “MicroRNAs: Target Recognition and Regulatory Functions.” *Cell*, vol. 136, no. 2, 2009, pp. 215–33, doi:10.1016/j.cell.2009.01.002.

Bernstein, Emily, et al. “Dicer Is Essential for Mouse Development.” *Nature Genetics*, vol. 35, no. 3, Springer Nature, Oct. 2003, pp. 215–17, doi:10.1038/ng1253.

Cai, Xuezhong, et al. “Human MicroRNAs Are Processed from Capped, Polyadenylated Transcripts That Can Also Function as MRNAs.” *RNA (New York, N.Y.)*, vol. 10, no. 12, 2004, pp. 1957–66, doi:10.1261/rna.7135204.

Cao, Dan Dan, et al. “MicroRNAs: Key Regulators in the Central Nervous System and Their Implication in Neurological Diseases.” *International Journal of Molecular Sciences*, vol. 17, no. 6, {MDPI} {AG}, May 2016, p. 842, doi:10.3390/ijms17060842.

Chiang, H. Rosaria, et al. “Mammalian MicroRNAs: Experimental Evaluation of Novel and Previously Annotated Genes.” *Genes and Development*, vol. 24, no. 10, 2010, pp. 992–1009, doi:10.1101/gad.1884710.

“Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways.” *Nature*, vol. 455, no. 7216, Springer Science and Business Media {LLC}, Sept. 2008, pp. 1061–68, doi:10.1038/nature07385.

De Rie, Derek, et al. “An Integrated Expression Atlas of MiRNAs and Their Promoters in Human and Mouse.” *Nature Biotechnology*, vol. 35, no. 9, 2017, pp. 872–78, doi:10.1038/nbt.3947.

Dobin, Alexander, et al. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics*, vol. 29, no. 1, 2013, pp. 15–21, doi:10.1093/bioinformatics/bts635.

Ehrenreich, Ian M., and Michael Purugganan. “MicroRNAs in Plants: Possible Contributions to Phenotypic Diversity.” *Plant Signaling & Behavior*, vol. 3, no. 10, 2008, pp. 829–30, doi:10.1101/gad.1004402.of.

Enright, Anton J., et al. “MicroRNA Targets in Drosophila.” *Genome Biology*, 2003, doi:10.1186/gb-2003-5-1-r1.

Fejes-Toth, K., et al. “Post-Transcriptional Processing Generates a Diversity of 5’-Modified Long and Short RNAs.” *Nature*, vol. 457, no. 7232, Nature Publishing Group, 2009, pp. 1028–32, doi:10.1038/nature07759.

Friedman, Robin C., et al. “Most Mammalian MRNAs Are Conserved Targets of MicroRNAs.” *Genome Research*, vol. 19, no. 1, Cold Spring Harbor Laboratory, Oct. 2009, pp. 92–105, doi:10.1101/gr.082701.108.

Fromm, B., Domanska, D., Hackenberg, M., Mathelier, A., Høye, E., Johansen, M., Hovig, E., Flatmark, K., Peterson, K. “MirGeneDB2.0: The Curated MicroRNA Gene Database Non-Coding RNAs (NcRNA), a Significant Part of the Increasingly Popular ‘.” *BioRxiv*, 2018.

Gao, Yan, et al. “Tissue-Specific Regulation of Mouse MicroRNA Genes in Endoderm-Derived Tissues.” *Nucleic Acids Research*, vol. 39, no. 2, 2011, pp. 454–63, doi:10.1093/nar/gkq782.

Geiss, Gary K., et al. “Direct Multiplexed Measurement of Gene Expression with Color-Coded Probe Pairs.” *Nature Biotechnology*, vol. 26, no. 3, 2008, pp. 317–25, doi:10.1038/nbt1385.

Han, Jinju, et al. “Molecular Basis for the Recognition of Primary MicroRNAs by the Drosha-DGCR8 Complex.” *Cell*, vol. 125, no. 5, Elsevier {BV}, June 2006, pp. 887–901, doi:10.1016/j.cell.2006.03.043.

He, Lin, and Gregory J. Hannon. “MicroRNAs: Small RNAs with a Big Role in Gene Regulation.” *Nature Reviews Genetics*, vol. 5, no. 7, Springer Nature, July 2004, pp. 522–31, doi:10.1038/nrg1379.

Jin, Daeyong, and Hyunju Lee. “A Computational Approach to Identifying Gene-MicroRNA Modules in Cancer.” *PLoS Computational Biology*, vol. 11, no. 1, 2015, doi:10.1371/journal.pcbi.1004042.

Kozomara, Ana, and Sam Griffiths-Jones. “MiRBase: Integrating MicroRNA Annotation and Deep-Sequencing Data.” *Nucleic Acids Research*, vol. 39, no. SUPPL. 1, Oxford University Press ({OUP}), Oct. 2011, pp. D152--D157, doi:10.1093/nar/gkq1027.

Ladewig, Erik, et al. “Discovery of Hundreds of Mirtrons in Mouse and Human Small RNA Data.” *Genome Research*, vol. 22, no. 9, 2012, pp. 1634–45, doi:10.1101/gr.133553.111.

Lagos-Quintana, Mariana, et al. “Identification of Tissue-Specific MicroRNAs from Mouse.” *Current Biology*, vol. 12, no. 9, 2002, pp. 735–39, doi:10.1016/S0960-9822(02)00809-6.

Lall, Sabbi, et al. “A Genome-Wide Map of Conserved MicroRNA Targets in C. Elegans.” *Current Biology*, vol. 16, no. 5, Elsevier {BV}, Mar. 2006, pp. 460–71, doi:10.1016/j.cub.2006.01.050.

Landgraf, Pablo, et al. “A Mammalian MicroRNA Expression Atlas Based on Small RNA Library Sequencing.” *Cell*, vol. 129, no. 7, 2007, pp. 1401–14, doi:10.1016/j.cell.2007.04.040.

Lee, Rosalind C., et al. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14.” *Cell*, vol. 75, no. 5, Elsevier {BV}, Dec. 1993, pp. 843–54, doi:10.1016/0092-8674(93)90529-Y.

Li, Bo, and Colin N. Dewey. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics*, vol. 12, 2011, doi:10.1186/1471-2105-12-323.

Lin, Shuibin, and Richard I. Gregory. “MicroRNA Biogenesis Pathways in Cancer.” *Nature Review Cancer*, vol. 15, no. 6, Nature Publishing Group, 2015, pp. 321–33, doi:10.1038/nrc3932.

Ludwig, Nicole, et al. “Distribution of MiRNA Expression across Human Tissues.” *Nucleic Acids Research*, vol. 44, no. 8, Oxford University Press ({OUP}), Feb. 2016, pp. 3865–77, doi:10.1093/nar/gkw116.

Maragkakis, Manolis, et al. “DIANA-MicroT Web Server Upgrade Supports Fly and Worm MiRNA Target Prediction and Bibliographic MiRNA to Disease Association.” *Nucleic Acids Research*, vol. 39, no. SUPPL. 2, Oxford University Press ({OUP}), May 2011, pp. W145--W148, doi:10.1093/nar/gkr294.

Martin, Marcel. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing

Reads." *EMBnet.Journal*, 2014, doi:10.14806/ej.17.1.200.

Mestdagh, Pieter, et al. "Evaluation of Quantitative MiRNA Expression Platforms in the MicroRNA Quality Control (MiRQC) Study." *Nature Methods*, vol. 11, no. 8, 2014, pp. 809–15, doi:10.1038/nmeth.3014.

Mineno, Junichi, et al. "The Expression Profile of MicroRNAs in Mouse Embryos." *Nucleic Acids Research*, vol. 34, no. 6, 2006, pp. 1765–71, doi:10.1093/nar/gkl096.

Morita, Sumiyo, et al. "One Argonaute Family Member, Eif2c2 (Ago2), Is Essential for Development and Appears Not to Be Involved in DNA Methylation." *Genomics*, vol. 89, no. 6, Elsevier {BV}, June 2007, pp. 687–96, doi:10.1016/j.ygeno.2007.01.004.

Mortazavi, Ali, et al. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods*, vol. 5, no. 7, 2008, pp. 621–28, doi:10.1038/nmeth.1226.

Ninova, Maria, et al. "Fast-Evolving MicroRNAs Are Highly Expressed in the Early Embryo of Drosophila Virilis." *Rna*, 2014, pp. 360–72, doi:10.1261/rna.041657.113.

Nueda, María José, et al. "Next MaSigPro: Updating MaSigPro Bioconductor Package for RNA-Seq Time Series." *Bioinformatics*, vol. 30, no. 18, 2014, pp. 2598–602, doi:10.1093/bioinformatics/btu333.

Park, Chong Y., et al. "Analysis of MicroRNA Knockouts in Mice." *Human Molecular Genetics*, vol. 19, no. R2, Oxford University Press ({OUP}), Aug. 2010, pp. R169--R175, doi:10.1093/hmg/ddq367.

Pertea, Mihaela, et al. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology*, vol. 33, no. 3, 2015, pp. 290–95, doi:10.1038/nbt.3122.

R Core Team. *R: A Language and Environment for Statistical Computing*.

Roberts, Brian S., et al. "Blocking of Targeted MicroRNAs from Next-Generation Sequencing Libraries." *Nucleic Acids Research*, vol. 43, no. 21, 2015, pp. 1–8, doi:10.1093/nar/gkv724.

Robinson, Mark D., et al. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics*, vol. 26, no. 1, 2009, pp. 139–40, doi:10.1093/bioinformatics/btp616.

Romaine, Simon P. R., et al. "MicroRNAs in Cardiovascular Disease: An Introduction for Clinicians." *Heart*, vol. 101, no. 12, BMJ, Mar. 2015, pp. 921–28, doi:10.1136/heartjnl-2013-305402.

Vidigal, Joana A., and Andrea Ventura. "The Biological Functions of MiRNAs: Lessons from in Vivo Studies." *Trends in Cell Biology*, vol. 25, no. 3, Elsevier {BV}, Mar. 2015, pp. 137–47, doi:10.1016/j.tcb.2014.11.004.

Wang, Yangming, et al. "DGCR8 Is Essential for MicroRNA Biogenesis and Silencing of Embryonic Stem Cell Self-Renewal." *Nature Genetics*, vol. 39, no. 3, Springer Nature, Jan. 2007, pp. 380–85, doi:10.1038/ng1969.

Wickham, Hadley. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

Wienholds, Erno, et al. "MicroRNA Expression in Zebrafish Embryonic Development." *Science*, vol. 309, no. July, 2005, pp. 310–11.

Wong, Nathan, and Xiaowei Wang. "MiRDB: An Online Resource for MicroRNA Target Prediction and Functional Annotations." *Nucleic Acids Research*, vol. 43, no. D1, Oxford University Press ({OUP}), Nov. 2015, pp. D146–52, doi:10.1093/nar/gku1104.

Wyman, Stacia K., et al. "Post-Transcriptional Generation of MiRNA Variants by Multiple

Nucleotidyl Transferases Contributes to MiRNA Transcriptome Complexity." *Genome Research*, vol. 21, no. 9, 2011, pp. 1450–61, doi:10.1101/gr.118059.110.

Zeng, Weihua, et al. "Single-Nucleus RNA-Seq of Differentiating Human Myoblasts Reveals the Extent of Fate Heterogeneity." *Nucleic Acids Research*, 2016, pp. 1–13, doi:10.1093/nar/gkw739.

Zhao, Wang, et al. "MicroRNA-143/-145 in Cardiovascular Diseases." *BioMed Research International*, vol. 2015, 2015, pp. 1–9, doi:10.1155/2015/531740.

Zhou, Yingyao, et al. "Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets." *Nature Communications*, vol. 10, no. 1, Springer US, 2019, p. 1523, doi:10.1038/s41467-019-09234-6.

CHAPTER 3

Long-TUC-seq is a robust method for quantification of metabolically labeled full-length isoforms

ABSTRACT

The steady state expression of each gene is the result of a dynamic transcription and degradation of that gene. While regular RNA-seq methods only measure steady state expression levels, RNA-seq of metabolically labeled RNA identifies transcripts that were transcribed during the window of metabolic labeling. Whereas short-read RNA sequencing can identify metabolically labeled RNA at the gene level, long-read sequencing provides much better resolution of isoform-level transcription. Here we combine thiouridine-to-cytosine conversion (TUC) with PacBio long-read sequencing to study the dynamics of mRNA transcription in the GM12878 cell line. We show that using long-TUC-seq, we can detect metabolically labeled mRNA of distinct isoforms more reliably than using short reads. Long-TUC-seq holds the promise of capturing isoform dynamics robustly and without the need for enrichment.

INTRODUCTION

Transcription is a dynamic process and different transcriptome profiles are indicative of different cellular states. While each cellular state can be identified by a set of quasi-steady state expression levels, all mRNA transcripts are transcribed and degraded at different rates (Munchel et al.; Lenstra et al.). The expression level of each gene isoform depends on its transcription rate, processing rate, and degradation rate. Although regular RNA-seq studies inform us of the steady state levels of each transcript, these lack any information on transcript stability or turnover rates. Transcription is controlled by cis-regulatory elements such as promoter and enhancer regions which play a role in determining the transcription rate of a transcript (Levine and Tjian). The binding of transcription factors as well as characterization of epigenetic marks from this category is primarily studied using ChIP-seq (Jiang and Mortazavi) and the chromatin accessibility can be measured by assays such as ATAC-seq (Buenrostro et al.). However, RNA degradation rates are just as important, and often times overlooked, when defining the steady state levels of expression (Maekawa et al.; Ghosh and Jacobson). Post-transcriptional regulatory factors such as miRNA and RNA binding proteins are the main players in regulating RNA stability and decay. Assays such as CLIP-seq and miRNA-seq have been developed to study the effects of each of these elements on gene expression (Ule et al.; Alon et al.). Overall, transcription is a complex process and using the expression profiles to understand the role of each of these regulators can be ambiguous and challenging.

Several new methods have been developed for genome-wide study of transcription dynamics. One category of these methods focuses on the study of nascent transcriptomes by profiling the RNA molecules instantaneously as they are being transcribed or processed. For instance, global run-on sequencing (GRO-seq) and precision nuclear run-on sequencing (PRO-seq) sequence the positions that the polymerase is residing at, providing information regarding active genes and the polymerase pausing dynamics (Core et al.; Mahat et al.). Another set of methods, such as native elongating transcript sequencing (NET-seq), report polymerase positions at the 3' ends of nascent transcripts (Nojima et al.; Churchman and Weissman). While GRO-seq, PRO-seq, and NET-seq investigate nascent transcripts, other methods focus on metabolic labeling of nascent RNA molecules that have been made over a window of time in order to study transcription and degradation rates. These methods use different nucleotide analogs to label the newly made RNA over a pulsing window followed by high throughput sequencing to detect the RNA molecules that incorporated the analog. A group of these methods such as bromouridine sequencing (Bru-seq), 4-thiouridine sequencing (4SU-seq) and transient transcriptome sequencing (TT-seq) rely on enrichment methods to recover signal from labeled transcripts (Paulsen et al.; Fuchs et al.; Schwalb et al.). Many of these methods suffer from enrichment biases and elution issues that lead to low yield and biases due to modified nucleotide identity used for enrichment.

More recently, additional methods have been developed that can still characterize modified nucleoside incorporation, but do not rely on enrichment. TimeLapse-seq, thiol(SH)-linked alkylation for the metabolic sequencing (SLAM seq), and thiouridine to cytidine conversion sequencing (TUC-seq) rely on chemical conversion of the metabolically incorporated analog. Modified positions are then identified in mutated cDNA in order to distinguish the metabolically labeled reads from pre-existing none-labeled reads (Schofield et al.; Herzog et al.; Gasser et al.). One of the challenges of this group of methods is the low incorporation rate of 4SU that results in under-estimation of recently transcribed genes (Russo et al.), especially when using short-read sequencing, which is still a long-standing challenge in transcriptomics, especially when interrogating more complex transcriptomes with large dynamic range.

All of these techniques rely on short-read Illumina sequencing, which even with higher sequencing depth cannot overcome these limitations. In addition, reconstructing different transcript models and quantifying the expression at the level of isoforms using short reads remains challenging and limited (Amarasinghe et al.). Long-read sequencing can improve the sensitivity of the assay by sequencing over the whole transcript, which would have a higher number of 4SU incorporated and makes it easier to detect over sequencing and biological noise. The two main long-read sequencing platforms are Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT). Despite the higher error rates in long-read technologies, the circular consensus technique implemented by PacBio has reduced the final error rate down to 1%

(Wenger et al.). Furthermore, long-read sequencing can unambiguously identify transcript isoforms using packages such as TALON (Wyman and Balderrama-Gutierrez et al., 2019), SQANTI (Tardaguila et al.) or FLAIR (Tang et al.).

In this work, we combine TUC metabolic labeling with long-read sequencing on the PacBio Sequel II platform to develop long-TUC-seq. We pulsed the GM12878 cells with 4 thio-Uridine (4SU) for 8 hours and then converted the incorporated 4SUs into cytidines using osmium tetroxide. We then built cDNA and libraries for sequencing on both Illumina NextSeq and PacBio platforms. We quantified the expression levels of each gene that correspond to the recently made RNA during the 8 hours pulsing window by quantifying the number of T→C substitutions identified in every read. We explored different thresholds to count the read with different levels of certainty as newly synthesized. We demonstrate that long-TUC-seq has higher sensitivity and lower FDR compared to the corresponding short-read version of TUC-seq. Finally, we count the reads in each category for all the isoforms to identify differences in transcription rates between isoforms of the same gene. Overall, long-TUC-seq is a robust protocol that would be widely applicable to a variety of settings were the metabolic labeling can be used to study transcriptome dynamics.

RESULTS

Identifying metabolically labeled RNA using long-TUC-seq

Our long-TUC-seq method relies on the incorporation of 4SU into the RNA and its further conversion to a regular cytidine (Fig. 3.1A.) We initially tested 4SU incorporation into recently synthesized transcripts by incubating GM12878 cells with 0.1mM and 1mM 4SU for a period of time between 2 to 24 hours and compared the amount of incorporation by dot blots (Fig. 3.2A). We then checked the RNA integrity after the treatment of the RNA samples with osmium tetroxide under different conditions (mainly time and temperature of the incubation). We compared the RNA Integrity Numbers (RINs) of the RNA samples after the treatment using a Bioanalyzer (Fig. 3.2B). Even with milder temperature (room temperature) we observed substantial degradation at 3 hours (RIN = 5.6). However, the integrity of the samples is improved with the addition of RNase inhibitor to the OsO_4 mix at this condition (Fig. 3.2D). Finally, we tested the conversion of incorporated 4SU by OsO_4 at this condition by checking the amount of 4SU remaining in the RNA sample before and after osmium treatment, using a dot blot assay (Fig. 3.2C). The dot blot shows complete conversion of 4SU with 3 hours of OsO_4 at room temperature.

We pulsed biological replicates of GM12878 cells with 1mM of 4SU for 8 hours and extracted the RNA, which were treated with osmium tetroxide. We also generated matching libraries of osmium treated samples without any 4SU pulsing. We built Illumina and PacBio libraries from these samples and sequenced them on their respective platforms and analyzed the data (Fig. 3.1B). Each of the PacBio libraries yielded between 3.4M - 6.2M raw sequencing reads (Table 1). After all the filtering, we are left with a minimum of 1.2M of reads for each sample that were mapped to human genome using minimap2 with an average of 99.65% mapping rate. In order to identify the reads that were synthesized during the 4SU pulse window, we counted the number of T→C substitutions for each read. We inspected the reads that mapped onto the MYC locus, which is known to be a fast turnover transcript (Fig. 3.1C). We observe that a high percentage (94%) of TUC-seq reads mapping to the MYC locus have at least 6 T→C events. By contrast, none of the reads mapping to MYC in the osmium control (sample without

4SU pulse and treated with OsO_4) or in publicly available PacBio ENCODE datasets are marked as labeled. We can therefore detect 4SU labeled reads based on the number of substitutions in a long read.

Distinct substitution profiles of long-TUC-seq at the level of base calls and reads

The nucleotide composition of the human genome is equally distributed between all the four nucleotides. There are some biological variations from multitude of SNPs that will introduce specific substitution events across the genome and some technical variation that is introduced via PCR or SBS. However, all of these substitutions should be distributed evenly between the 12 different possible substitution types globally. While this equal distribution is observed in the control PacBio RNA-seq from ENCODE and the osmium control, there is a very distinct profile in our long-TUC-seq samples with a much higher T→C counts as expected (Fig. 3.3A; Fig. 3.4). In order to asses our ability to call a read as labeled, we analyze the distribution of reads based on the number of T→C observed. We detect 34% of all the reads being labeled with more than 6 T→C in the TUC-seq samples compared to 0.4% in the osmium control and in the RNA-seq control (Fig. 3.3B). In addition, we detect 27% and 21% of the reads from the TUC-seq samples are labeled with a minimum of 20 and 30 T→C.

To ensure that the reads labeled by long-TUC-seq are not heavily biased by longer transcripts, we determined the correlation of the number of T→C with the length of each transcript. Although the number of observed T→ C in a read does correlate weakly with length of the transcript (Pearson correlation coefficient of 0.25), its distribution in the controls indicates that the transcript length is not a big driver of noise, which will therefore not hinder an accurate count of labeled transcripts (Fig. 3.3C). Finally, we counted the number of Ts in each transcript that has been converted to C in order to obtain an estimate of 4SU incorporation rate. Our 8-hour long-TUC-seq results indicate an average of 11.33% for 4SU incorporation in the transcription process, assuming a 100% conversion to C (Fig. 3.3D).

Robust detection of recently synthesized genes by long-TUC-seq

We used TranscriptClean (Wyman and Mortazavi) to correct the indels in our reads before running TALON V4.4.2 (Wyman and Balderrama-Gutierrez et al., 2019) to annotate the reads as known and novel transcripts, as well as to obtain accurate counts for each gene and transcript for each of our 4 datasets (1 RNA-seq control, 1 osmium controls and 2 TUC-seq samples). For the purpose of this study, we focused on known isoforms. We detect 21,496 known genes across the experiments and 32,250 (TPM > 0) known transcripts. We added the labeling information for each read to the TALON annotations and calculated the expression levels for each gene and transcript for the following 4 categories: all reads, permissive threshold (>6 T→C), intermediate threshold (>20 T→C) and conservative threshold (>30 T→C.) We detect an average of 9,270 genes labeled at permissive threshold with more than 2 TPM expression of labeled reads, in the TUC-seq samples compared to 35 genes out of 10,584 genes detected in the controls (FDR = 0.33%). This number drops to 8,169 in the conservative category of labeled reads in the TUC-seq samples (Fig. 3.5A). The detection of recently synthesized genes is very robust across the replicates, with 80% of detected labeled genes (> 2 TPM at permissive threshold) being confirmed by both replicates (Fig. 3.5B). There is also a high concordance amongst the expression levels of these recently synthesized genes across the replicates with 0.93 Pearson correlation (Fig. 3.5C). This correlation is still high for genes detected in the higher categories with Pearson correlation of 0.93 for intermediate labeled reads and 0. 92 for conservative reads (Fig. 3.6).

Comparison of long-TUC-seq with Illumina short-TUC-seq

Current methods using metabolic labeling for studying the dynamics of transcription rely on short read illumina sequencing. In order to benchmark our long-TUC-seq results we compared it with the short-read TUC-seq of the same samples. We built the Illumina Nextera libraries using the same cDNA materials that were used for PacBio libraries. We then sequenced these libraries on the Illumina NextSeq platform and mapped the reads to the human transcriptome reference using STAR with an average of 45M single end reads mapped per sample. We annotated each read with the number of observed substitutions and annotated the aligned reads with it. Here we also detect higher T→C substitution profile for TUC-seq samples compared to the controls (Fig. 3.7A). The TUC-seq samples contain more reads with higher T→C compared to the controls (Fig. 3.7B); based on the substitution profiles and the read distributions, we decided to used 2, 4 and 6 T→C as permissive, intermediate and conservative thresholds for calling the labeled reads. We detect 27% of total reads labeled with > 2 T→C in TUC-seq samples compared to 1.5 % in control samples. Although raising the threshold to 4 T→C reduces the percentage of false positive labeled reads in controls to 0.14%, it also reduces the percentage of labeled reads in the TUC-seq samples to 15%. Finally, we calculated the 4SU incorporation rate from Illumina short-read TUC-seq samples to be 17.22% which is 6% higher than what we have obtained using Pacbio long-TUC-seq data (Fig. 3.7C).

Using the permissive threshold of 2 T→C, we detect 57% of reads mapping to MYC in labeled samples (Fig. 3.8A), which is 37% lower than what was detected by PacBio. We then quantify the expression levels in each category using eXpress (Roberts and Pachter) as described in the methods. In order to compare the detection of labeled genes by each platform, we use the intermediate threshold for Illumina (4 T→C) which resulted in similar FDR (0.5%) to that of PacBio data with permissive threshold (FDR = 0.3%). Although Illumina TUC-seq detects twice as many genes across all the samples compared to PacBio (> 0 TPM), the number of detected genes at intermediate threshold is 5,511, which is much less than labeled genes in PacBio (Fig. 3.8B). When comparing genes (expressed > 2 TPM) detected as labeled in either platforms, we find that 47% are shared and the majority of the remainder (41% of all labeled genes) is detected only in PacBio (Fig. 3.5D). In general, the expression levels of the genes detected at 2 TPM or higher by only one of the platforms is lower than the expression of the commonly shared detected genes (Fig. 3.5E). Thus long-TUC-seq is more sensitive than its short-read equivalent at similar FDR thresholds.

Calculating degradation rates with long-TUC-seq

After annotating the detected genes with the different degrees of labeling, we focused on the dynamics of transcription for each gene and analyzed the rate at which each gene is transcribed. On average, 50% of the total expression of genes at the end of our 8-hour labeling window comes from newly synthesized RNA. MYC is one of the genes with faster turnover rate that is expressed at 111 TPM with 95% of its expression being labeled whereas GAPDH with a high expression of 11,378 TPM has only 5.6% labeled RNA (Fig. 3.9A).

Under a steady-state assumption that the overall expression level of a gene stays the same through the 8-hour pulsing window, the rates at which a gene is being transcribed and the rate at which it is degraded are constant. We calculated the degradation rates and the half-life of each gene using the total expression of the gene and its newly synthesized RNA. We obtained a degradation rate of 45 TPM/hour and a half-life of 1.7 hours for the MYC gene. The ranking of genes based on their half-life time is similar to what has been observed previously (Spearman correlation of 0.74 with timeLapse-seq ranking in K562 cells) (Schofield et al.). Although we used a long labeling time of 8 hours, the method could work with substantially shorter labeling time. Long-TUC-seq can be used to calculate degradation rates from 4SU labeling of transcripts and genes.

Analysis of isoform-specific expression and transcription rates

One of the advantages of long-read sequencing is that it inherently measures the expression levels of the isoforms of each gene. In our study, more than 58% of genes are expressed as multiple isoforms with an average of 2.5 isoforms per gene. GAPDH, which is one of the higher expressed genes, has 4 distinct isoforms. MYC, which is one of the higher turnover genes, has 2 isoforms detected. The highest number of isoforms belong to MSL3, with 15 isoforms detected. We can also take a step further and analyze the expression levels of each isoform to see if the gene is expressed through one isoform more than the other, or if it is expressed uniformly across different isoforms by calculating the isoform specificity index (ISI) for all genes as described in the methods. In the case of a gene that expresses all its isoforms equally, the ISI will be closer to zero and in the case of a gene that expresses primarily one of its multiple isoforms, the ISI will be closer to one. MYC and GAPDH each have an ISI of 0.67 and

0.99, respectively, which for MYC translates to the fact that its isoforms are expressed in a 3:1 ratio, and for GAPDH it means that its isoforms are expressed in approximately a 800:20:4:1 ratio (Fig. 3.10).

We can similarly define the isoform specificity index based on the expression levels of newly synthesized transcripts (ISI_{new}) and inspect the isoform specificity of the transcription machinery for a specific gene at a given time. The distribution of ISI_{total} and ISI_{new} for all the genes of GM12878 shows that majority of multi-isoform genes are expressed and being transcribed in a highly isoform-specific manner, and there is a Pearson correlation coefficient of 0.63 between total and labeled isoform specificity (Fig. 3.9B). Furthermore, we are interested in genes with ubiquitous isoform expression that are being transcribed in an isoform-specific manner. In order to obtain a list of such genes, we filter the genes with lower ISI_{total} (< 0.35) and higher ISI_{new} (>0.85). There are 9 genes in this category, all with two isoforms detected in our dataset, the expression of which are less than two-fold apart. However, the expression of recently synthesized isoforms is in some cases more than 70-fold different (Fig. 3.9C). One such gene is LRR1 that encodes for Leucine-rich repeat protein 1, which plays a role in protein ubiquitination and modification. This gene has five isoforms, two of which have been detected in our dataset with similar expression levels of about 5 TPM (201 and 202). These isoforms are protein coding and they differ only in one exon; however, the LRR1-202 isoform which has an extra exon compared to the 201 isoform has a much higher turnover, and about 73% of its expression has been made within the 8-hour pulse window.

DISCUSSION

Here, we introduce a method for detecting and quantifying metabolically labeled RNA at a single isoform resolution using PacBio long-read sequencing. To do so, we relied on the conversion of incorporated 4SU to C by TUC-seq chemistry. We demonstrated that even though short-read Illumina sequencing provides much higher depth in comparison to PacBio sequencing, we are able to recover higher number of labeled genes with PacBio. We also show that not only can PacBio detect the labeled RNA reproducibly, the quantification of these labeled RNAs is also highly concordant between the biological replicates. Furthermore, we took advantage of having T to C substitution data for full transcripts in order to calculate an accurate estimation of 4SU incorporation rate within each transcript. This estimation using illumina short-read technique would be in accurate and over-estimated due to the fact that many of the reads aligning to the T depleted regions are dis-missed as unlabeled.

We use long-TUC-seq data to obtain estimations of degradation rates of genes and consequently their half-lives. The caveats with these estimations are the two assumptions used in their calculations. First is the steady state assumption that the expression level, synthesis rate, and degradation rate of each gene is constant during the pulsing time, which can be closer to reality when the pulsing time is much shorter than 8 hours. The other assumption used in these calculations is that there is no doubling of cells during the 8-hour pulsing window. Although that might be the case with some of the cells, many of the cells would have inevitably doubled and the observed total and labeled RNA could be coming from different number of cells from beginning of the pulsing to the end point. However, all these limitations apply to the estimations obtained by short-read TUC-seq and similar labeling techniques. While in this study we focus on labeling newly synthesized RNA using pulse labeling with 4SU, we could have instead performed a chase experiment to obtain degradation rates in situations where the main assumption would not hold.

Finally, the main advantage of using long-read sequencing for detection and quantification of recently transcribed genes is that it allows us to annotate the recently synthesized transcripts at isoform levels. Using this feature of long-read sequencing, we were able to identify a representative set of genes that, despite having rather ubiquitous expression across their isoforms, have substantially different transcription dynamics across isoforms. This could reflect the fact that some isoforms are required for a faster dynamic of a response whereas other isoforms are required to be more stable in order to confer robustness to some pathways. Having such resolution, one can infer the degradation rate, synthesis rate and the half-life of each of the isoforms and study the regulatory mechanism that affect these rates by integrating this data with other genomic assays such as miRNA-seq and ChIP-seq, and assays that focus on poly-A tails and 3'/5'-UTRs. In summary, Long-TUC-seq can robustly identify and quantify recently transcribed genes at the level of individual isoforms to shed light on differential isoform transcription and degradation rates.

FIGURES

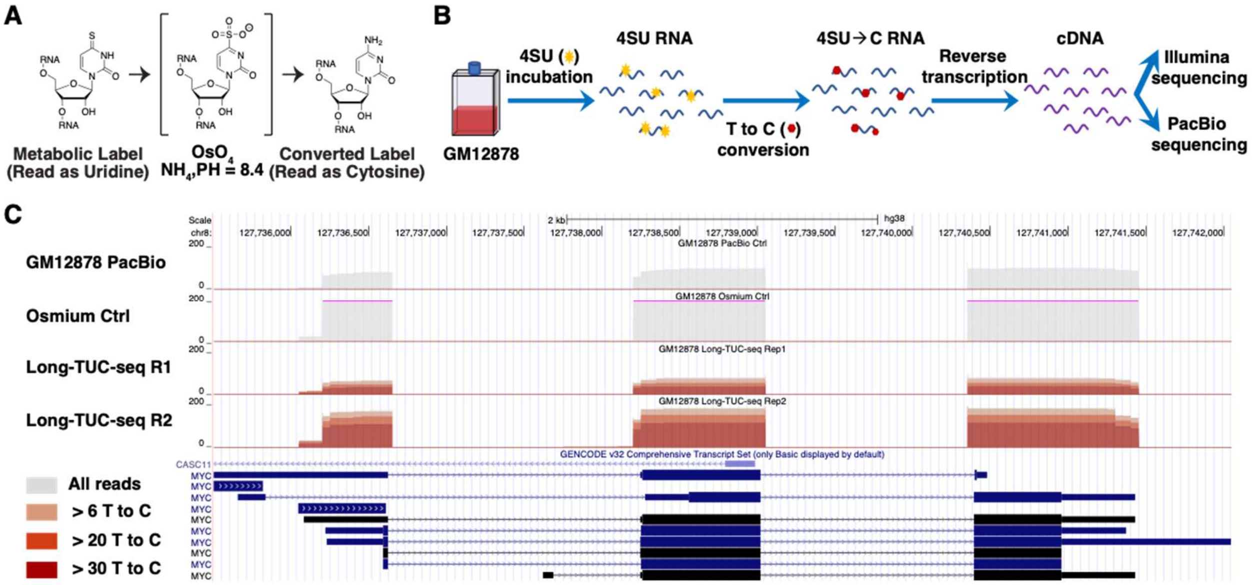

Figure 3.1. Identification of recently synthesized transcripts in GM12878 by long-TUC-seq. Osmium tetroxide converts an incorporated 4SU into a regular cytidine. **b)** Experimental layout of TUC-seq sample preparation, starting with the incorporation of 4SU into the GM12878 cells following by its conversion to C using OsO_4 and finally library building from cDNAs. **c)** Genome browser screenshot of PacBio data of GM12878 control from ENCODE, Osmium treated GM12878 without 4SU incorporation and two biological replicates of long-TUC-seq samples. The shot shows reads aligned to MYC, with increasing levels of labeled reads colored with darkening shades of red. The tracks are shown on a scale of 0 to 200 reads.

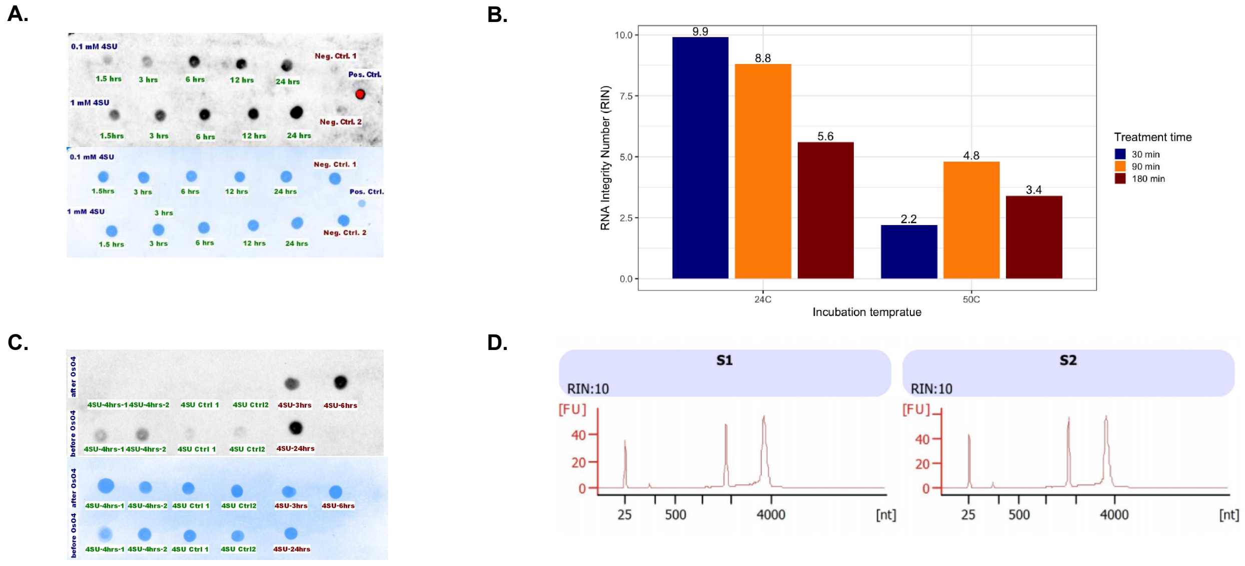

Figure 3.2. Optimization of TUC-seq protocol for GM12878. a) dot blot of GM12878 RNA extracted after different incubation times with 1mM or 100µM of 4SU **b**) The RIN scores of RNA samples from 4 hours 4SU incorporation samples after treatment with OsO_4 at room temperature or 50°C for 30, 90 and 180 mins. **c**) The dot blot of samples with and without 4 hours of 4SU before and after 3 hours treatment by OsO_4 at room temp. **d**) The bioanalyzer profiles of RNA from the 4hours 4SU incubations, treated with 3 hours of OsO4 at room temperature with addition of 1µl of Rnase inhibitor.

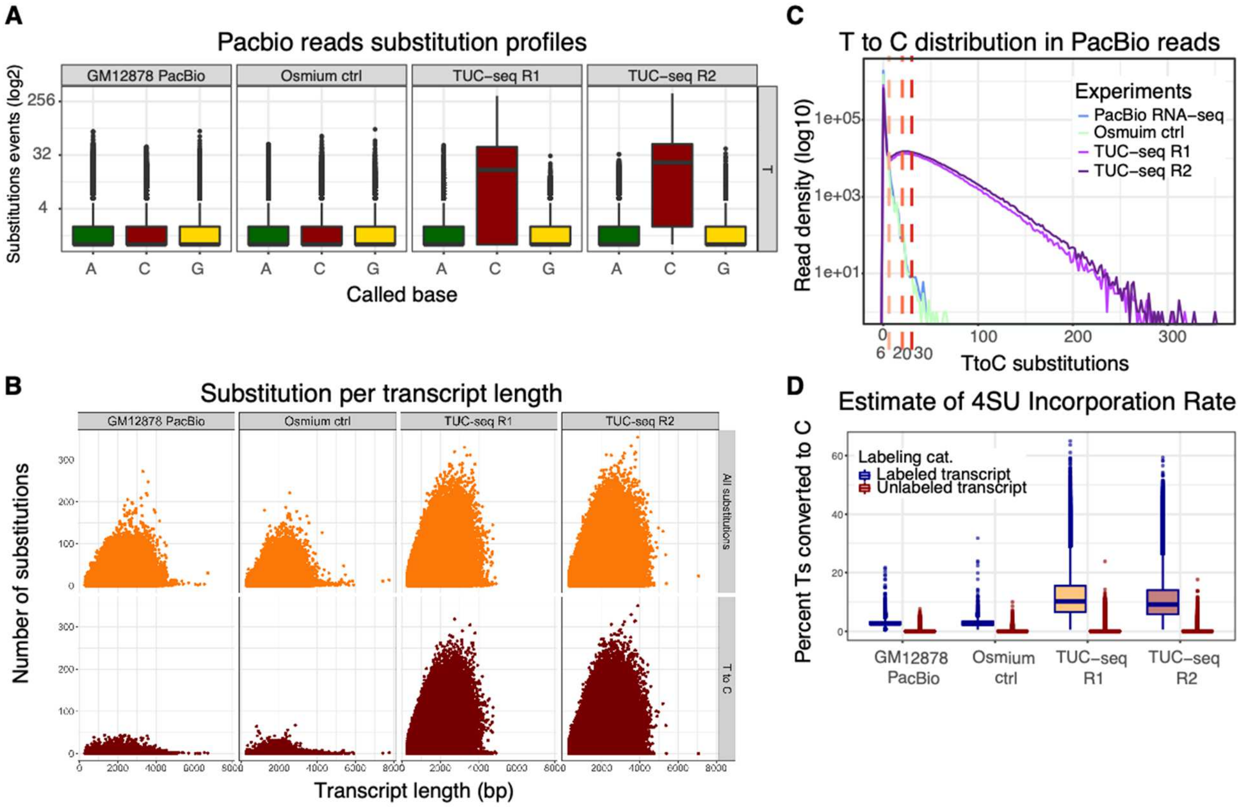

Figure 3.3. Identifying labeled reads in long-TUC-seq. **a)** Comparison of the profiles for all the possible substitutions at a reference T base across the TUC-seq samples and controls. **b)** The distribution of PacBio reads with respect to the number of T→C substitutions observed for each read. The three dotted lines indicate the thresholds we used to define lower, medium and higher labeled reads. **c)** Number of substitution events observed in each read with respect to the length of each read, showing slightly higher T→C events in the longer reads for TUC-seq samples. **d)** Average number of Ts in each read converted to C for labeled and unlabeled reads across the TUC-seq samples and controls.

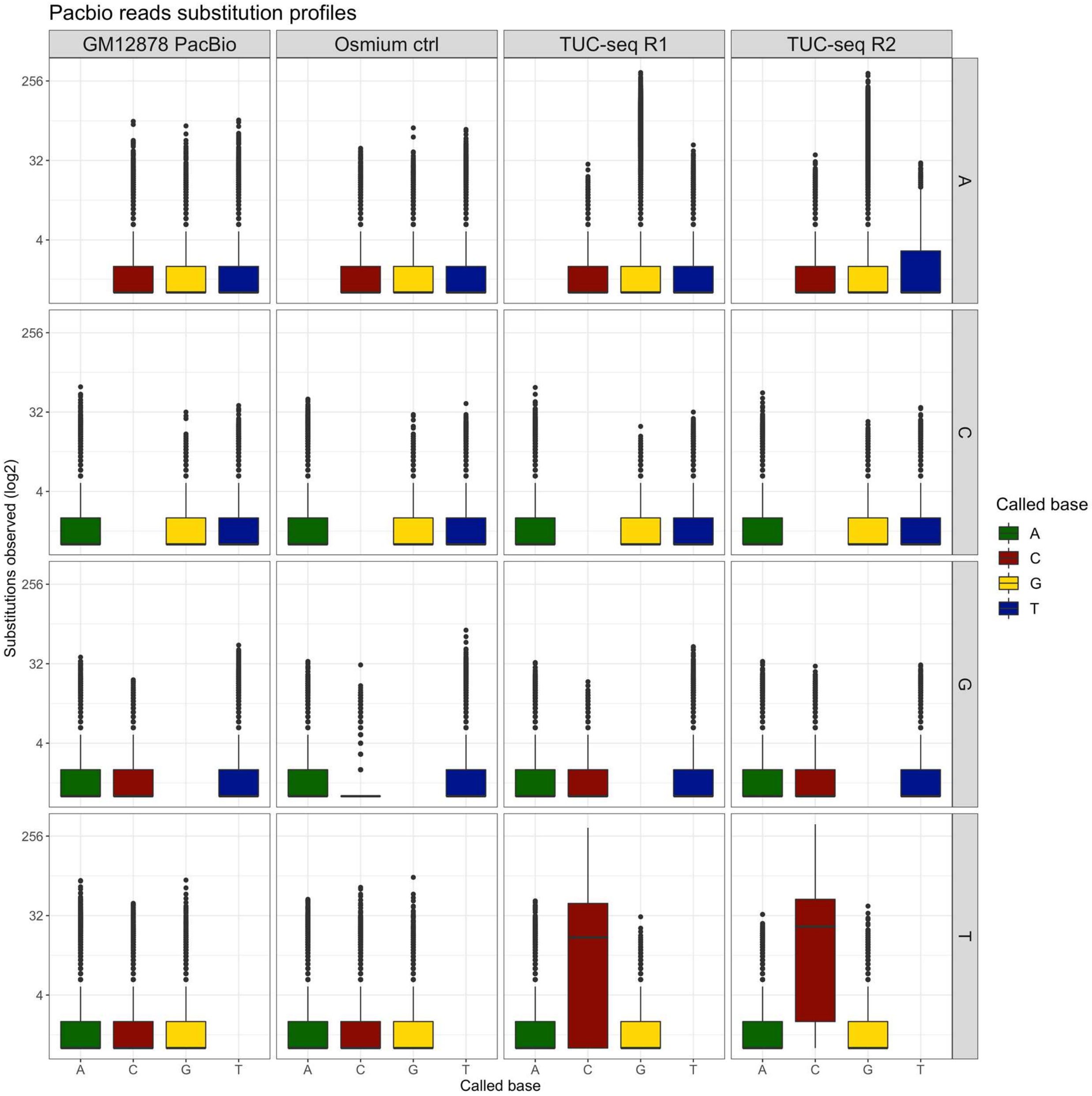

Figure 3.4. Long-TUC-seq substitution profiles. The number of all possible substitutions detected across TUC-seq samples and controls. The top label bases are the reference bases and the bottom labels corresponds to what base is identified in the read for that position.

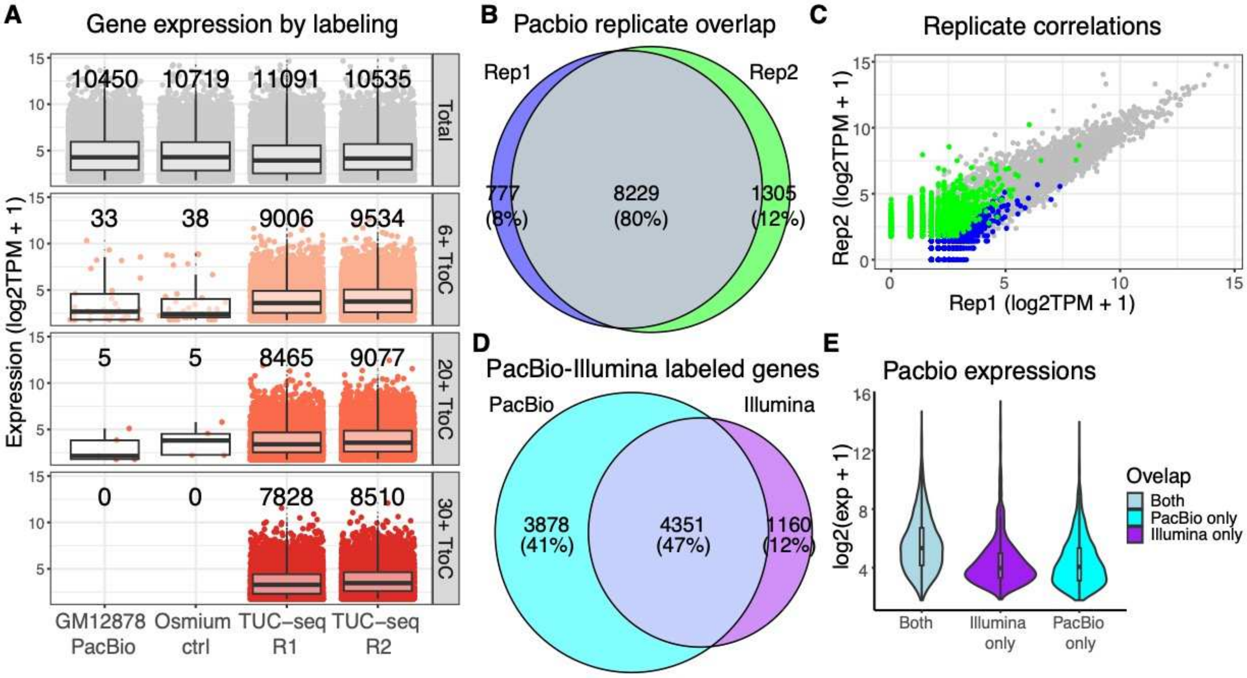

Figure 3.5. Robust identification and quantification of recently transcribed genes. a) Expression levels of genes with more than 2 TPM in each of the labeling categories. The total number of genes is indicated on top of each of the boxplots. **b)** The overlap of genes detected as lower labeled in either of the TUC-seq replicates showing the percentage of genes in each section of the Euler diagram. **c)** The correlation of lower labeled genes between the two replicates with a Pearson correlation coefficient of 0.93. **d)** Overlap of genes detected as lower labeled by both replicates on Illumina and those detected by both replicates of PacBio. **e)** The expression levels of the genes in each section of the Euler plot in section c. The expression levels of “both” and “PacBio only” groups are from PacBio and the expression levels of the middle group (“Illumina only”) is from Illumina data.

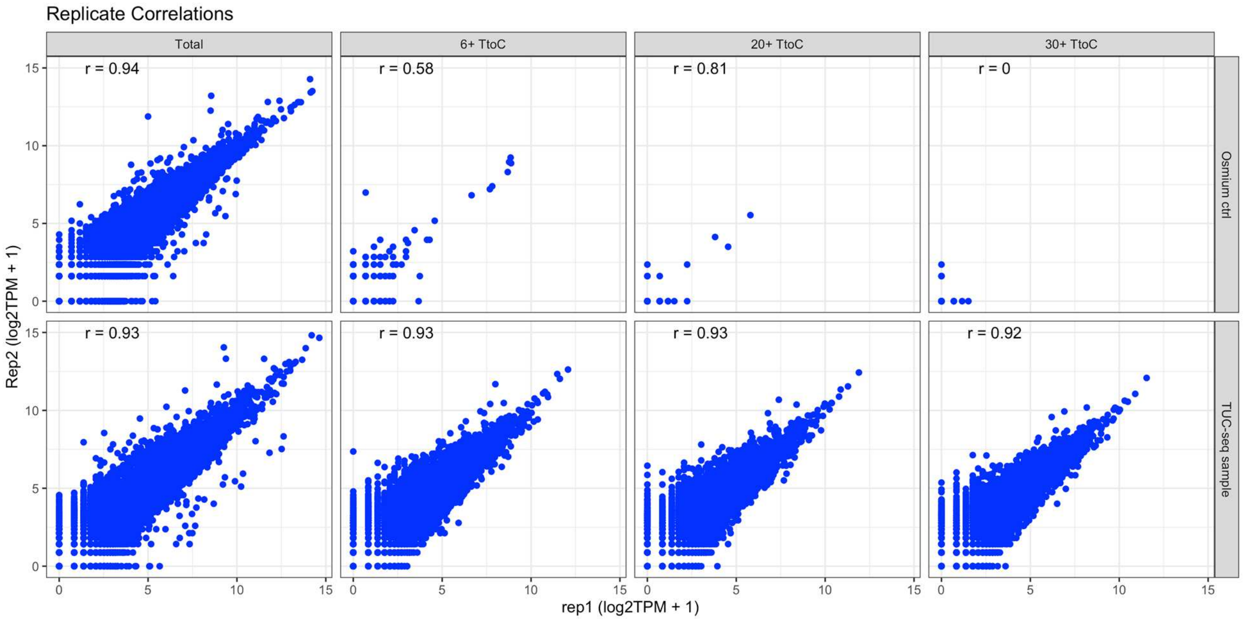

Figure 3.6. Long-TUC-seq expression concordance between biological replicates. The correlation of expression levels for different categories of labeled RNA and total RNA for TUC-seq samples and the osmium controls. The r value corresponds to the Pearson correlation between the log2TPM values.

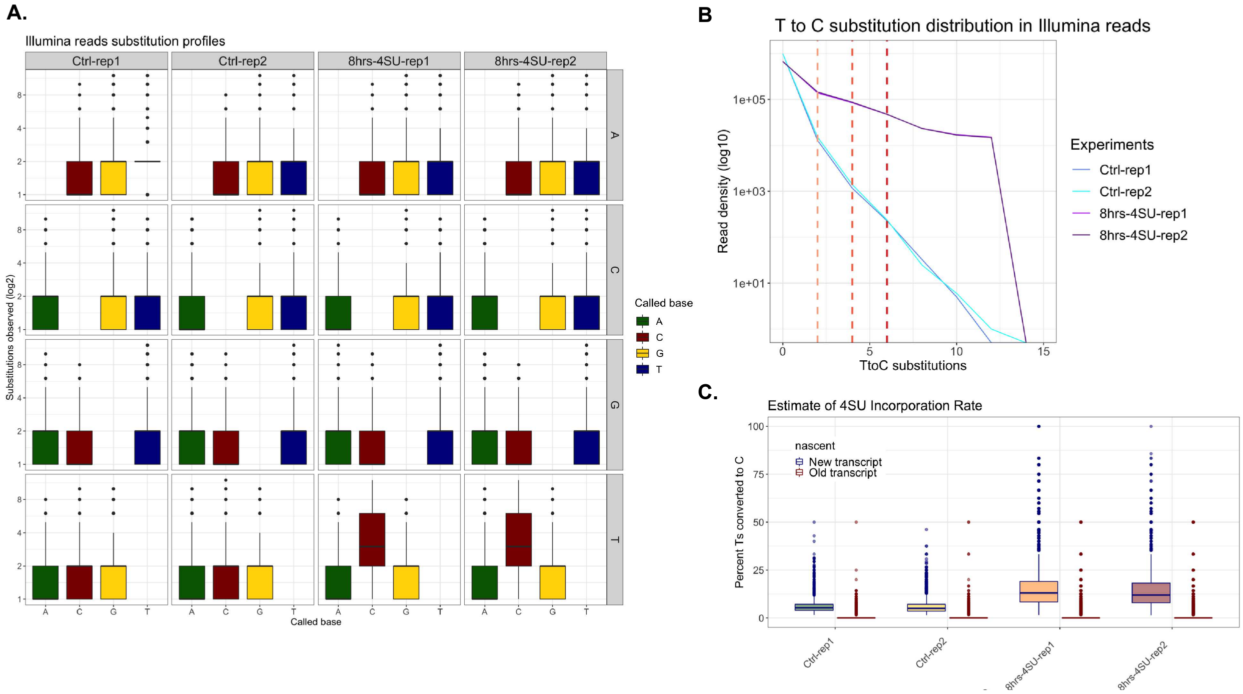

Figure 3.7. short-TUC-seq performance. a) The number of all possible substitutions detected across TUC-seq samples and controls. The side label bases are the reference bases and the bottom labels corresponds to what base is identified in the read for that position. **b**) Distribution of the reads based on the number of T→C substitutions identified. **c**) Average number of Ts in each read converted to C for labeled (new) and unlabeled (old) reads across the TUC-seq samples and controls.

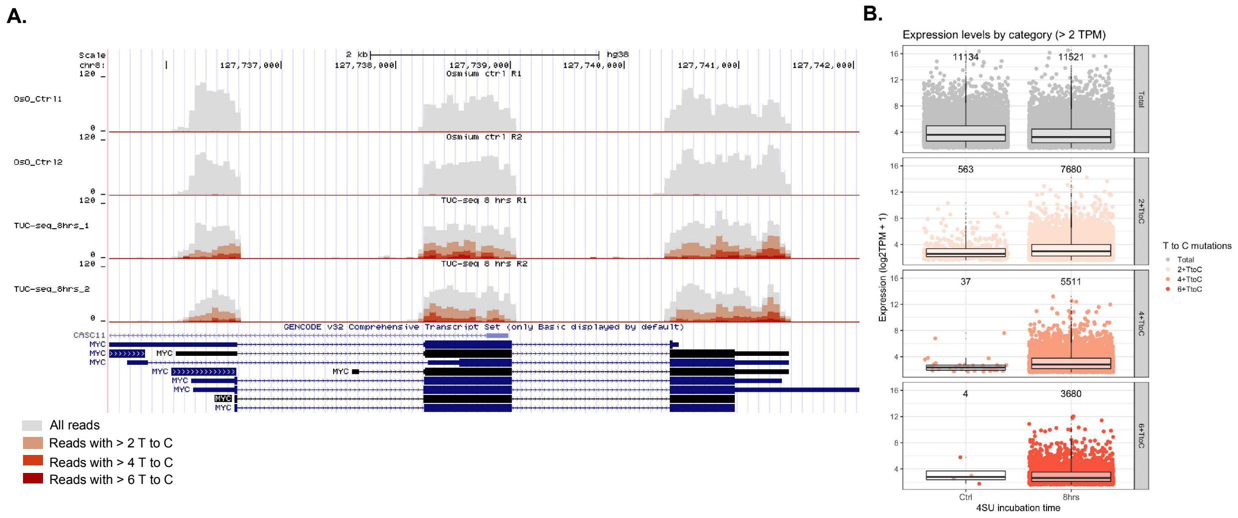

Figure 3.8. Short-TUC-seq identification of recently transcribed genes. a) Genome browser shot of mapped reads to MYC colored by the level of T→C substitutions in each read. **b**) Expression levels of genes with more than 2 TPM expression in each of the labeling categories. The total number of genes is indicated on top of each of the boxplots.

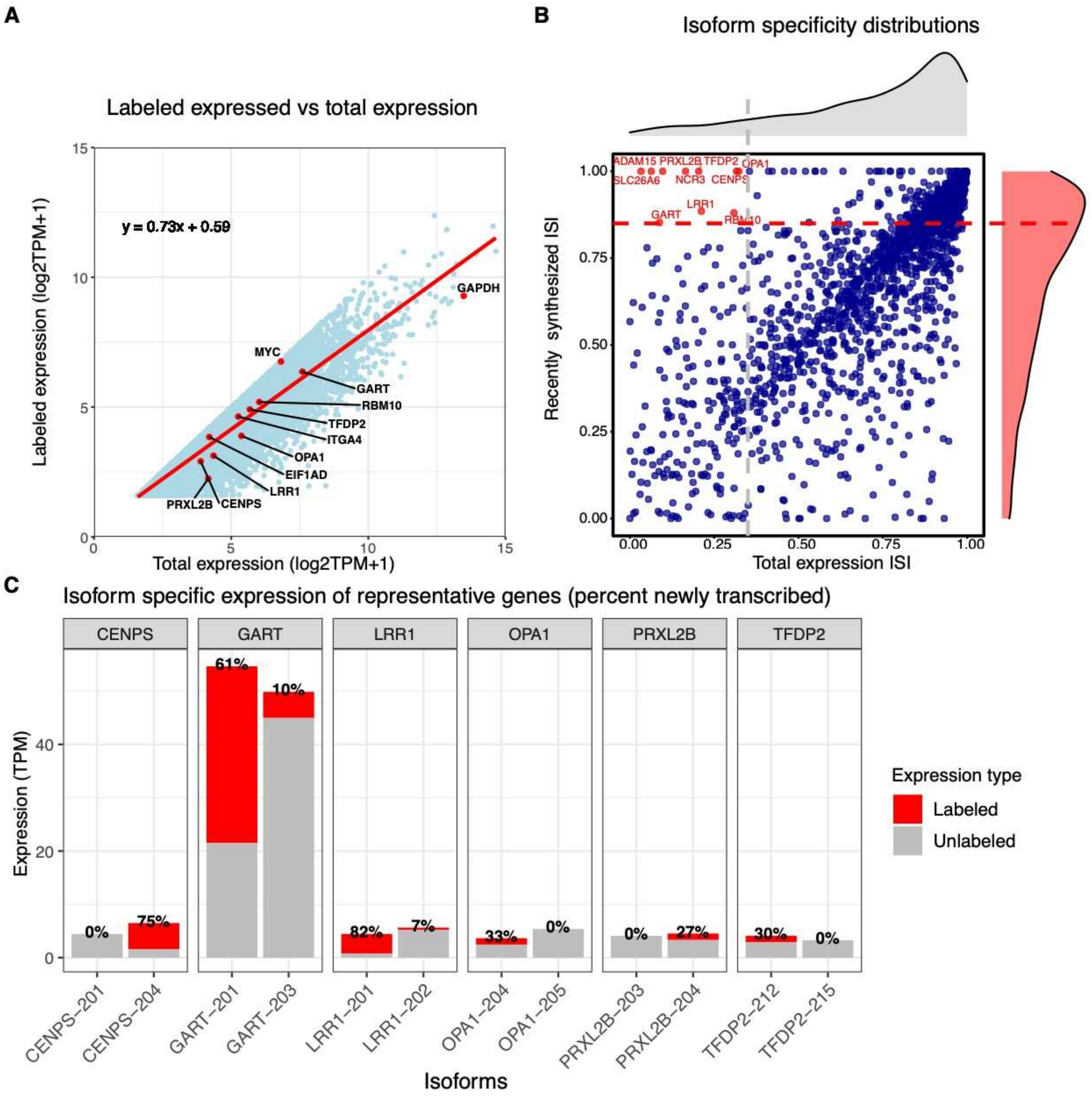

Figure 3.9. Dynamics of expression at the level of individual isoforms. **a)** Expression levels of recently transcribed genes (labeled at permissive threshold) with respect to the total expression level of that gene (for genes >2 TPM). The equation corresponds to the regression line drawn in red. Two example genes (MYC and GAPDH) are highlighted in red. **b)** The distribution of isoform specificity indices for all of the genes calculated from total expression (in grey) and from recently made transcripts (in red). The dotted lines indicate the thresholds used to

find genes with lower ISI_{total} (< 0.35) and higher $ISI_{labeled}$ (> 0.85). **c)** Expression levels of the isoforms corresponding to representative genes from the set defined in b. The grey portion of the bars corresponds to the expression level of pre-existing RNA and the red portion corresponds to the recently synthesized transcripts. Finally, the percentages on top of the bars are representing the percentage of total expression of the isoform that is transcribed recently.

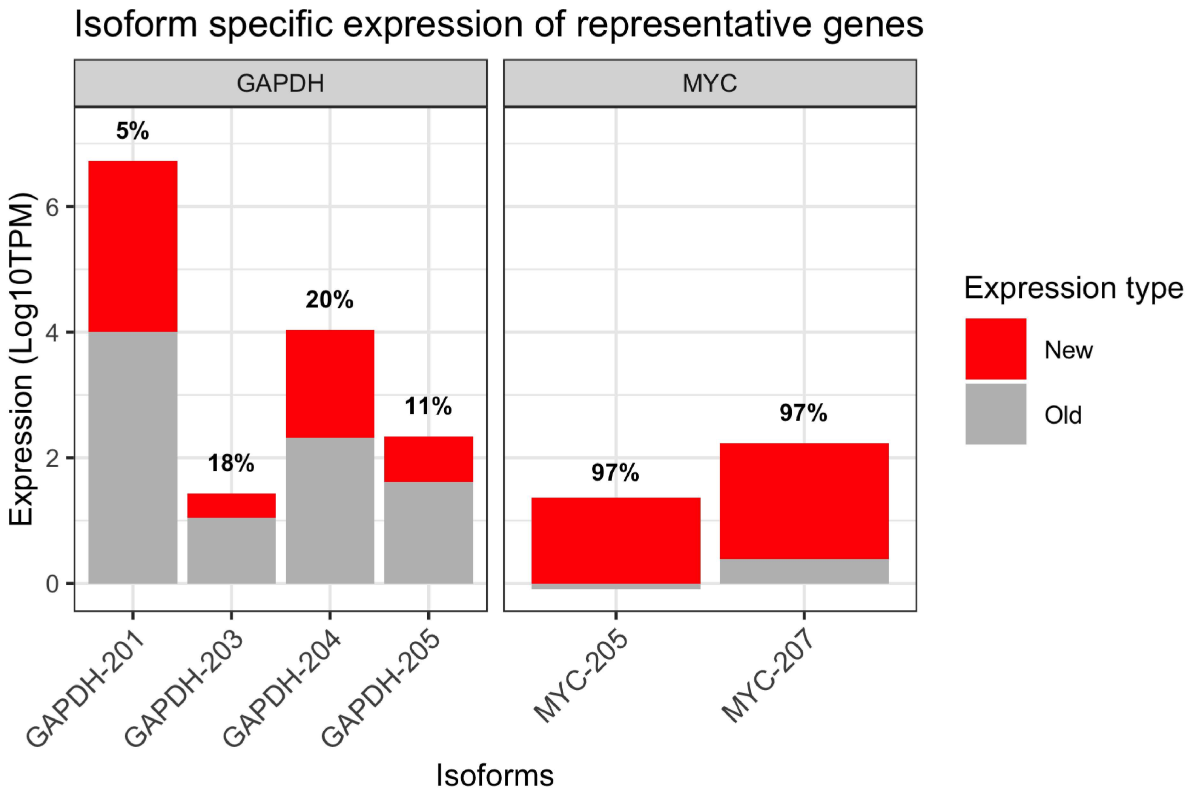

Figure 3.10. Expression dynamics of MYC and GAPDH at isoform levels. Expression levels of GAPDH and MYC isoforms. The grey portion of the bars corresponds to the expression level of pre-existing RNA and the red portion corresponds to the recently synthesized transcripts. Finally, the percentages on top of the bars are representing the percentage of total expression of the isoform that is transcribed recently.

TABLES

Description	Raw Reads	N50 (bp)	CCS Reads	Mapped reads	Mapping rate
GM12878 PacBio	6.1M	1,857	3.8M	2.1M	99.8%
Osmium ctrl	4.7M	2,006	2.8M	1.7M	99.7%
Long-TUC-seq R1	4.4M	1,891	2.1M	1.3M	99.3%
Long-TUC-seq R2	6.2M	1,978	3.9M	1.2M	99.8%

Table 3.1. Sequencing Statistics

METHODS

Sample collection and RNA extraction

GM12878 cells were obtained from Corriell Institute and were cultured in accordance with ENCODE protocols (www.encodeproject.org). The cells were passed every two to three days at 200k-500k cells/mL density and were harvested for the experiments at 500k-1M cells/mL. The RNA was extracted using QIAGEN RNeasy Plus kit (Cat. No. 74134).

TUC-seq sample preparation

4-thiouridine was obtained from Sigma Aldrich (T4609) and used fresh at a working concentration of 200 mM. For each TUC-seq experiment, 10-15M cells were spun down and resuspended in 10-15 mL of fresh media with added 4SU at a final concentration of 1mM (no 4SU was added for the osmium controls). The cells were incubated with 4SU for 8 hours and harvested for RNA extraction. The RNA was then treated with OsO_4 solution for 3 hours at room temperature in dark. The osmium solution was prepared fresh every time by mixing 20 µl of 1mM OsO_4 (Sigma Aldrich, 201030) with 4µl of 2M NH_4Cl at pH 8.8 and 1µl of RNasin Plus RNase inhibitor (Promega, N2615) for every 10µg of RNA. The RNA was then purified using Zymo RNA cleanup kit (R1015). Finally, the RNA was treated with 1U of exonuclease from epicenter (Terminator™ 5′-Phosphate-Dependent Exonuclease, TER51020) for 1 hour at 30^{o}C and neutralized by 1µl 100mM EDTA. Then, the RNA was once more purified with Zymo RNA cleanup kit.

PacBio library preparation and sequencing

The set III of SIRV controls were spiked into the RNA samples at a level of 0.03% of the total RNA. The cDNA was generated using a modified version of SMART-seq2 protocol. We then followed SMRTbell Template Prep Kit 2.0 to build PacBio libraries using 1-2µg of input RNA. We checked the quality of the libraries using the Bioanalyzer 2000 and Qubit to get the final concentrations. Finally, the libraries were delivered for sequencing on a Sequel II platform at UCI sequencing core facility, using 1 SMRT cell per library.

Illumina library preparation and sequencing

Starting from 30-50ng of the same cDNA, we followed the Illumina tagmentation protocol using Nextera DNA Flex Library Prep Kit to generate Illumina short-read libraries. We checked the concentration of the libraries with Qubit and got the average length of the library using the Bioanalyzer. We then performed a 2x43 paired-end sequencing on our NextSeq 500 instrument.

PacBio data processing

Raw reads from Sequel II machine were processed by PacBio circular consensus package (CCS v4.0.0) to filter any reads with less than 3 passes (parameters: --noPolish --minLength=10 --minPasses=3 –min-rq=0.9 –min-snr=2.5). Then reads with misconfigured adapters were filtered using PacBio lima package (v1.10.0; parameters: --isoseq --num-threads 12 --min-score 0 --min-end-score 0 --min-signal-increase 10 --min-score-lead 0). Finally, full-length non-chimeric (FLNC) reads were extracted using the PacBio Refine package (v3.2.2; parameters: --min-polya-length 20 --require-polya). The bam files processed by Refine were then converted to fastq files and they were all deposited to GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149551) with the exception of PacBio GM12878 control sample which has been previously deposited onto ENCODE portal (https://www.encodeproject.org/experiments/ENCSR838WFC/).

The FLNC reads were then aligned to a modified version of human genome reference (GRCh38 with added SIRV and ERCC references) using minimap2 (v2.17; parameters: -ax splice:hq -t 16 --cs -uf). We then used TransciptClean (v2.0.2; parameters: -m False --primaryOnly) for reference-based error correction of the reads. We provided TranscriptClean with splice junctions reference derived from the GENCODE annotations using TranscriptClean accessory script get_SJs_from_gtf.py. We also provided it with VCF-formatted NA12878 truth-set small variants from Illumina Platinum Genomes. We first initialized the TALON database with GENCODE v29 + SIRVs/ ERCC annotations using talon_initialize_database and finally annotated the reads by running TALON V4.4.2 module on all the datasets. We obtained the table of annotated reads from all the datasets by running the talon_summarize module. All the scripts used for analysis of long-TUC-seq Pacbio data can be accessed on the mortazabilab github. (https://github.com/mortazavilab/long-TUC-seq)

PacBio labeling of the reads

We used a custom python script (mismatch_analysis_PB.py) to annotate the reads with their corresponding substitutions. The script uses the CS tag option from minimap2 to count different types of substitutions and to generate a text file containing each read name and its corresponding substitution tally. The script also breaks down the alignment file into subfiles, each containing one of one category of reads (> 0 , > 6 , > 20 and > 30 T→C) for visualization on the UCSC genome browser. The information on different substitutions was added to the annotations obtained from TALON. We then calculate the TPM and counts for each of the categories for each gene and transcript.

Illumina data processing

The reads from illumina runs were mapped to human transcriptome reference (GRCh38.p12, gencode.v29.primary_assembly.annotation) using STAR aligner (v2.6.0c; parameters: --

outFilterMismatchNmax 15 --outFilterMismatchNoverReadLmax 0.07 --outFilterMultimapNmax 10 --outSAMunmapped None --outSAMattributes MD NM --alignIntronMax 10 --alignIntronMin 20 --outSAMtype BAM SortedByCoordinate). The raw fastq files for each sample is available on GEO database under GSE149551 accession. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149551).

Illumina calling of the labeled reads

We ran a custom python script (mismatch_analysis_ill.py) to annotate each of the mapped reads with the number different substitution events. The script uses the MD tag to tally the number of substitutions for each read. The script also breaks down the alignment file into sub-files of reads with > 0, > 2, > 4 and > 6 T→C substitutions. Finally, we count the reads in each category using eXpress (v1.5.1; parameters: --no-bias-correct). The quantification can be accessed under GSE149551 accession in GEO database. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149551). All the scripts used to process illumina TUC-seq data can be accessed on the Mortazavilab github. (https://github.com/mortazavilab/TUC-seq)

Degradation rate and half-life calculations:

Assuming steady-state and doubling rate of zero during the pulsing time, we can calculate the degradation rate (λ_i) and consequently the half-life (hl_i) of gene i:

$$\lambda_i = \frac{-\ln(1-\frac{L_i}{R_i})}{t_L},$$

$$hl_i = \frac{ln2}{\lambda_i}$$

Here R refers to the steady state expression of the specific mRNA, L stands for the expression of labeled RNA, and t_L is the labeling time.

Isoform specificity analysis

In order to help us understand the isoform specificity of each gene and its dynamics, we introduce an index for isoform specificity of a gene (ISI) as follows:

$$ISI_i = \frac{\sum_1^j (1 - X_{ij})}{N_i - 1}$$

Here, index *i* corresponds to each gene and index *j* represents each corresponding isoform for gene *i*. X is the expression level of the isoform normalized to the expression level of the highest expressed isoform of the gene *i*. Finally, N_i is the number of isoforms corresponding to gene *i*. We calculated the isoform specificity indices for each of the genes using the total and labeled RNA. Then we filter for the genes with more than 2 isoforms that has an $ISI_{total} < 0.35$ and $ISI_{labeled} > 0.85$. We then plot the expression of each isoform of a representative set of these genes and color the portion of the expression that corresponds to the labeled reads.

REFERENCE

Alon, Shahar, et al. “Barcoding Bias in High-Throughput Multiplex Sequencing of MiRNA.” *Genome Research*, vol. 21, no. 9, Cold Spring Harbor Laboratory, July 2011, pp. 1506–11, doi:10.1101/gr.121715.111.

Amarasinghe, Shanika L., et al. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology*, vol. 21, no. 1, Genome Biology, 2020, pp. 1–16, doi:10.1186/s13059-020-1935-5.

Buenrostro, Jason D., et al. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology*, vol. 2015, no. January, 2015, pp. 21.29.1-21.29.9, doi:10.1002/0471142727.mb2129s109.

Churchman, L. Stirling, and Jonathan S. Weissman. “Native Elongating Transcript Sequencing ({NET}-Seq).” *Current Protocols in Molecular Biology*, vol. 98, no. 1, Wiley, Apr. 2012, pp. 14.4.1--14.4.17, doi:10.1002/0471142727.mb0414s98.

Core, Leighton J., et al. “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.” *Science*, vol. 322, no. 5909, American Association for the Advancement of Science ({AAAS}), Dec. 2008, pp. 1845–48, doi:10.1126/science.1162228.

Fuchs, Gilad, et al. “4sUDRB-Seq: Measuring Genomewide Transcriptional Elongation Rates and Initiation Frequencies within Cells.” *Genome Biology*, vol. 15, no. 5, 2014, p. R69, doi:10.1186/gb-2014-15-5-r69.

Gasser, Catherina, et al. “Thioguanosine Conversion Enables MRNA-Lifetime Evaluation by RNA Sequencing Using Double Metabolic Labeling (TUC-Seq DUAL).” *Angewandte Chemie - International Edition*, 2020, doi:10.1002/anie.201916272.

GHOSH, SHUBHENDU, and ALLAN JACOBSON. “MRNA Decay Modulates Gene Expression and Controls Its Fidelity.” *Wiley Interdisciplinary Reviews: RNA*, vol. 1, no. 3, 2010, pp. 351–61, doi:10.1007/springerreference_35999.

Herzog, Veronika A., et al. “Thiol-Linked Alkylation of {RNA} to Assess Expression Dynamics.” *Nature Methods*, vol. 14, no. 12, Springer Nature, Sept. 2017, pp. 1198–204, doi:10.1038/nmeth.4435.

Jiang, Shan, and Ali Mortazavi. “Integrating ChIP-Seq with Other Functional Genomics Data.” *Briefings in Functional Genomics*, vol. 17, no. 2, 2018, pp. 104–15, doi:10.1093/bfgp/ely002.

Lenstra, Tineke L., et al. “Transcription Dynamics in Living Cells.” *Annual Review of Biophysics*, vol. 45, no. 1, 2016, pp. 25–47, doi:10.1146/annurev-biophys-062215-010838.

Levine, Michael, and Robert Tjian. “Transcription Regulation and Animal Diversity.” *Nature*, vol. 424, no. 6945, 2003, pp. 147–51, doi:10.1038/nature01763.

Maekawa, Sho, et al. “Analysis of RNA Decay Factor Mediated RNA Stability Contributions on RNA Abundance.” *BMC Genomics*, vol. 16, no. 1, 2015, pp. 1–19, doi:10.1186/s12864-015-1358-y.

Mahat, Dig Bijay, et al. “Base-Pair-Resolution Genome-Wide Mapping of Active RNA Polymerases Using Precision Nuclear Run-on (PRO-Seq).” *Nature Protocols*, vol. 11, no. 8, Nature Publishing Group, 2016, pp. 1455–76, doi:10.1038/nprot.2016.086.

Munchel, Sarah E., et al. “Dynamic Profiling of MRNA Turnover Reveals Gene-Specific and System-Wide Regulation of MRNA Decay.” *Molecular Biology of the Cell*, vol. 22, no. 15, 2011, pp. 2787–95, doi:10.1091/mbc.E11-01-0028.

Nojima, Takayuki, et al. “Mammalian NET-Seq Reveals Genome-Wide Nascent Transcription Coupled to RNA Processing.” *Cell*, vol. 161, no. 3, 2015, pp. 526–40, doi:10.1016/j.cell.2015.03.027.

Paulsen, M. T., et al. “Coordinated Regulation of Synthesis and Stability of RNA during the Acute TNF-Induced Proinflammatory Response.” *Proceedings of the National Academy of Sciences*, 2013, doi:10.1073/pnas.1219192110.

Roberts, Adam, and Lior Pachter. “Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments.” *Nature Methods*, vol. 10, no. 1, 2013, pp. 71–73, doi:10.1038/nmeth.2251.

Russo, Joseph, et al. “Metabolic Labeling and Recovery of Nascent RNA to Accurately Quantify MRNA Stability.” *Methods*, vol. 120, 2017, pp. 39–48, doi:10.1016/j.physbeh.2017.03.040.

Schofield, Jeremy A., et al. “TimeLapse-Seq: Adding a Temporal Dimension to RNA Sequencing through Nucleoside Recoding.” *Nature Methods*, vol. 15, no. 3, Nature Publishing Group, 2018, pp. 221–25, doi:10.1038/nmeth.4582.

Schwalb, Björn, et al. “TT-Seq Maps the Human Transient Transcriptome.” *Science*, vol. 352, no. 6290, 2016, pp. 1225–28, doi:10.1126/science.aad9841.

Tang, Alison D., et al. “Full-Length Transcript Characterization of SF3B1 Mutation in Chronic Lymphocytic Leukemia Reveals Downregulation of Retained Introns.” *Nature Communications*, vol. 11, no. 2020, Springer US, 2020, pp. 1–12, doi:10.1038/s41467-020-15171-6.

Tardaguila, Manuel, et al. “Corrigendum: SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification (Genome Research (2018) 28 (396-411) DOI: 10.1101/Gr.222976.117).” *Genome Research*, vol. 28, no. 7, 2018, p. 1096, doi:10.1101/gr.239137.118.

Ule, Jernej, et al. “CLIP: A Method for Identifying Protein-RNA Interaction Sites in Living Cells.” *Methods*, vol. 37, no. 4, 2005, pp. 376–86, doi:10.1016/j.ymeth.2005.07.018.

Wenger, Aaron M., et al. “Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome Aaron.” *Nature Biotechnology*, vol. 37, no. 10, 2020, pp. 1155–62, doi:10.1038/s41587-019-0217-9.Accurate.

Wyman, Dana, et al. “A Technology-Agnostic Long-Read Analysis Pipeline for Transcriptome Discovery and Quantification.” *BioRxiv*, 2019, p. 672931, doi:10.1101/672931.

Wyman, Dana, and Ali Mortazavi. “TranscriptClean: Variant-Aware Correction of Indels, Mismatches and Splice Junctions in Long-Read Transcripts.” *Bioinformatics*, vol. 35, no. 2, 2019, pp. 340–42, doi:10.1093/bioinformatics/bty483.

CHAPTER 4

Investigating transcriptome dynamics during HL-60 macrophage differentiation using metabolic labeling

ABSTRACT

MicroRNAs (miRNAs) are post-transcriptional regulatory elements that play a key role in differentiation and developmental processes. Although miRNAs are known to be fairly stable, recent studies indicate that their stabilities are differentially regulated, and this could shed light into their role as a post-transcriptional regulator of expression. Previous attempts at studying these dynamics either used low throughput molecular techniques or relied on enrichment of labeled miRNA that produces some biases. Here we develop micro-TUC-seq, a new technique for reproducibly measuring miRNA transcription dynamics using 4SU labeling and its conversion to a cytidine. We then apply this method to study the dynamics of miRNAs during HL60 differentiation to macrophages and to decipher their role in this process.

INTRODUCTION

miRNAs are small pieces of RNA that regulate the expression levels of their target mRNAs via degradation. MiRNAs play a key role in development and their mis-regulation can lead to variety of diseases (Chandra et al.; Sayed and Abdellatif). Historically the discovery of miRNAs was tied with a developmental process. The first miRNA discovered in *C. elegans* Lin-4 was known to be involved in temporal regulation of “larval to adult switch” (Ambros; Lee et al.). Since then, many miRNAs have been studied as key regulators of developmental and differentiation processes. Combined loss of miRNAs by deletion of DICER or DGCR8 leads to embryonic lethality and arrest at E7.5 and E6.5 (Yangming Wang et al.; Bernstein et al.). In addition, many individual miRNAs regulate different differentiation processes: miR-27 in myogenesis (Crist et al.), miR-124 and miR-9 in neurogenesis (Visvanathan et al.; Coolen and Bally-cuif), and miR-30 in nephrogenesis (Agrawal et al.). Although some individual miRNAs are key to the development of certain tissues, others might play a softer role as the fine tuners of development and endow the process with robustness (Alberti and Cochella). The study of miRNAs is essential to complete our understanding of the dynamics of developmental and differentiation processes.

MicroRNA regulation of mRNAs involves two different mechanisms, which are translational repression and mRNA degradation (Valencia-Sanchez et al.; Huntzinger and Izaurralde). Although translational repression precedes the degradation mechanism, the latter is responsible for the majority of overall regulation (Iwakawa and Tomari). Some studies utilize the changes in target expression levels for validation of functional miRNA targeting in a gain or loss of function fashion (Garzon et al.). More recent studies take advantage of genome-wide expression data of mRNAs and miRNAs to infer functional relationships and to validate the predicted targets of miRNAs (Le et al.; Rahmanian et al.; Ovando-Vázquez et al.). However, since mRNA expression levels are regulated by many other transcriptional regulatory mechanisms, it might be more beneficial to constrain the correlation analysis to the changes in degradation of target expression rather than the overall changes. Decoupling of transcription and mRNA decay can be achieved through expression data (Alkallas et al.) or other methods that directly measure decay rates by metabolic labeling. Hence, the kinetic information of mRNA transcription and decay can be used to improve functional analysis of miRNA target prediction.

Primary miRNAs, just like regular mRNAs, are capped and poly-adenylated and go through similar mechanisms of degradation and decay. The pre-miRNAs and mature miRNAs do not share those features that control the half-life of mRNA. However, mature miRNAs are known to be much more stable than an average mRNA, perhaps due to their shorter length. Nevertheless, the stability of miRNAs is also under differential regulations (Li et al.; Bail et al.).

In order to comprehensively understand the regulatory effects of miRNAs, we need to also consider the stability and dynamics of miRNAs themselves (Zhou et al.).

Many methods have been developed to study mRNA transcription and decay dynamics. The majority of these methods relies on metabolic labeling of RNA and its identification by enrichment such as in Bru-Seq (Paulsen et al.) and TT-seq (Schwalb et al.), or by chemical conversion of the analog to a different base such as in SLAM-seq (Herzog et al.) and TUC-seq (Lusser et al.). While the study of mRNA dynamics is well-established and many tools have been developed for it, there are only a couple of studies that have used 4SU labeling for studying miRNA biogenesis and decay dynamics (Marzi and Nicassio; Duffy et al.). Furthermore, both of these methods rely on enrichment of labeled miRNA, which can introduce biases. Therefore, there is still room for improved methods to study miRNA dynamics. Here, we study the dynamics of miRNAs by investigating their transcription and decay rates using a new method called micro-TUC-seq that we initially establish in human GM12878 lymphoblastoid cells.

HL-60 is promyelocytic cell line that can differentiate into monocytes (Mangelsdorf et al.), macrophages (Murao et al.) and neutrophils (Breitman et al.) through simple induction using different reagents. The transcriptional dynamics of the HL-60 cell line during these differentiations have been previously well studied using parallel RNA-seq and ATAC-seq (Ramirez et al.). Furthermore, miRNA expression profiles of HL-60 and throughout its TPA-induced differentiation to monocytes/macrophages (Kasashima et al.) as well as its DMSO-induced differentiation towards neutrophils (Dakir and Mollinedo) have been studied. All these comprehensive studies make HL-60 differentiation a good target for studying the coupling of miRNA and mRNA dynamics throughout differentiation.

In order to better decipher the role of miRNAs during HL-60 differentiation, we performed a concordant dynamic study of miRNAs and mRNAs during a five-day timecourse of HL-60 differentiation into macrophages by phorbol 12-myristate 13-acetate (PMA) using Illumina TUC-seq and micro-TUC-seq. We observe a slightly negative correlation between MYC, which is one of the transcription factors downregulated during HL-60 differentiation, and the miRNAs that are known to target it. We also detect an additional number of miRNAs and mRNAs that are only differentially detected at the level of their labeled expression.

RESULTS

Identifying metabolically labeled miRNA using micro-TUC-seq

We labeled GM12878 cells using 4SU with varying incubation times from 2 to 12 hours followed by osmium treatment to study the dynamics of miRNA biogenesis and decay. We sequenced the resulting miRNA libraries on the Illumina NextSeq platform with a minimum of 8.4M raw reads for each of the libraries. We mapped the data to the miRBase miRNA database and obtained an average of 4.1M uniquely mapped reads per sample (within standard range of ENCODE miRNA-seq, Table 4.1). The majority of miRNAs annotated in miRBase have five or more Ts in their sequence (Fig. 4.1A), however due to low efficiency of 4SU incorporation into RNA, we decided to categorize the miRNA reads into three categories of lowly (with 1 or more T→C), medium (with 2 or more T→C) and highly (with 3 or more T→C) labeled reads. This way, we are able to detect most of labeled miRNAs, with the T-rich miRNAs showing up in the highly labeled category (Fig. 4.1B). For miR155-3p, which is one of the highly expressed miRNAs in our data with 7 Ts in its sequence, we detect an average of 53% labeled reads and a maximum as high as 10% of the reads with more than 2 T→C at 2- and 6-hours pulse time. On the other hand, for let-7a-5p, which is another highly expressed miRNA with 9 Ts in its sequence, we detect only 3% labeled reads (Fig. 4.1C.) Hence, we are able to identify the reads that come from recently labeled miRNAs and quantify the percentage of labeled reads during the pulsing window.

We detect an average of 344 miRNAs expressed across the samples (with >= 2 CPM in each replicate). The number of miRNAs detected as lowly labeled (>= 2 CPM; >=1 T→C in each replicate) varies from 93 to 177 depending on the time of 4SU pulsing (Fig. 4.1D). Out of the lowly labeled miRNAs, 57% are medium labeled and 25% are highly labeled. We observe an average Pearson correlation of 0.93 between the miRNA expression levels of replicates across our samples. We also observe high concordance between the expression levels of the labeled miRNAs across the duplicates with average Pearson correlation of 0.95, 0.98 and 0.98 for labeled categories from low to high (Fig. 4.1E). Our current thresholds result in 9.7% FDR for the lowly labeled miRNAs in our controls and less than1% FDR for highly labeled miRNAs. The miRNAs called as false positives in the control runs have an average of 7.5 Ts in their sequence and have a median total expression of 4,153 CPM. At lower threshold micro-TUC-seq can detect 30-42% of the miRNA as labeled miRNA with a rather high FDR, however using a medium threshold we can detect 15-30% labeled miRNAs with a much more reduced FDR of 1.5%.

In general, we detect a growing number of labeled miRNAs by increasing the pulsing time window, however we expect to continue detecting a miRNA labeled at a shorter pulse time in the longer pulse times as well. Out of 93 miRNAs detected as labeled within 2 hours, 83% are consistently detected as labeled with longer time points. Although at each incubation time we observe a new group of miRNAs detected as labeled for the first time, a variant percentage of these miRNAs are consistently detected in the following incubations times (Fig. 4.2A). To further investigate this inconsistency, at each incubation time, we split the group of miRNAs detected as labeled for the first time into consistence and inconsistence. We observe that across all the incubation time points, the median expression of the inconsistent group of miRNAs is lower than the consistent subgroup (Fig. 4.2B). Furthermore, it appears that longer incubation times allow us to detect the transcription of miRNAs with lower overall expression.

Exploring the dynamics of miRNA production and decay in GM12878

In addition to the detection of labeled miRNAs, we can quantify the expression of labeled miRNA and use that as a measure for transcription dynamics of each miRNA. One simple measure of transcription is the ratio of labeled miRNA to the total expression of the miRNA. On average, 14% to 16% of the expression of labeled miRNAs comes from labeled reads, but at each incubation time there are a number of miRNAs such as miR-155-3p, miR-20a-3p and miR-378i where the majority of their reads are labeled (Fig. 4.3A). There are 34 unique miRNAs that have more than 50% labeled expression in at least one of the incubation time points. Out of these 34 miRNAs, the majority of these are detected as labeled miRNAs within 4 hours of 4SU incubation, however there are five of these miRNAs that are not detected until a longer incubation time (miR-139-5p, miR-1296-5p, miR-26a-2-3p, miR-579-3p and let-7f-2-3p). It is important to note that two of these miRNAs, miR-4455 and miR-12135, are also called as lowly labeled in the control runs and hence are considered as false positives in our experiment. Finally, a heatmap of labeled miRNA expression, clustered by the timepoint that they are first detected as labeled, shows a steady increased of labeled expression through incubation times for all the miRNAs since their first detection (Fig. 4.3B).

Detection of recently made mRNA and miRNA during the HL-60 differentiation timeline

We differentiated HL-60 cells into macrophages with the addition of PMA as described previously (Ramirez et al.). We collected RNA samples along the time course (3 hours to 96 hours in duplicates) and conducted TUC-seq and micro-TUC-seq experiments in parallel on the same RNA samples with 3 hours of 4SU pulsing prior to collection. At least one of the replicates at each time point was successfully sequenced at sufficient depth to be considered for our study (Table 4.2; Table 4.3). We detect an average of 9,258 genes expressed in our HL60 samples and an average of 50% of the genes being labeled in the 3 hours of 4SU incubation (Fig. 4.4A). The miRNA data shows an average of 362 miRNAs detected per timepoint samples and 30% of miRNAs are detected as labeled on average per sample (Fig. 4.4B). Although the percentage of labeled genes increases from 20-30% to around 50% by 12 hours and then stays the same, the total number of detected genes continues to increase up to 72 hours. On the other hand, although the total number of expressed miRNAs does not change significantly across the samples, the percentage of labeled miRNAs decreases through time, especially when considering the medium labeled category.

The gene expression levels of total reads and lowly labeled mRNAs are highly reproducible across the samples with available replicate data (average Pearson correlation of 0.98 and 0.97 for total and lowly labeled expressions, respectively). Similarly, the miRNA total and labeled expression levels are highly concordant between replicates (average Pearson correlation of 0.97 and 0.96, respectively). We calculated the fraction of mRNA and miRNA total expression that is labeled as a measure of transcription rate and dynamics. In general, we observe a surge of higher transcription dynamics in the earliest time point with an average label RNA fraction of 0.41, which then levels off to an average fraction of 0.22 (Fig. 4.5A). There is a group of 102 genes at 3 hours adherent time point with labeled fraction of more than 0.9, however the median expression of these genes is only 3 TPM. One of the higher expressed genes had a rather high labeled fraction of 0.74 at 3 hours and 0.64 at 0 hour, nevertheless the labeled fraction of this gene varies throughout our time course. Unlike MYC, the labeled fraction of GAPDH stays invariant through our time course (Fig. 4.5B). We do not detect a significant change in global changes of miRNA transcription dynamics throughout the time course (Fig. 4.6A). Four miRNA that are more than 50% labeled in at least half of the samples and not the control runs are: miR-26a-2-3p, miR-520g-3p, miR-4448 and let-7f-2-3p. However, different group of miRNAs are revealed as higher turnover miRNAs across time points (Fig. 4.6B).

Integrative analysis of miRNA and mRNA expressions during HL-60 differentiation

MYC is one of the known regulators of HL-60 proliferation that is downregulated during macrophage differentiation. We can detect this down-regulation especially at the earlier time points within 12 hours (Fig. 4.7A). We searched for miRNAs that are known to target MYC and repress it. Most of these miRNAs have a surge of expression at some point of differentiation. Out of these miRNAs, miR-144-3p, miR-145-3p, miR-451a and miR-494-3p reach their peak expression in 3 hours adherent cells, miR-145-5p, miR-148a-3p and miR-148-5p reach their peak expression at 12 hours, and finally miR-34-5p and miR-375-3p increase their expression at 24 hours and 48 hours respectively (Fig. 4.7B). The expressional changes of these miRNAs and their target MYC are negatively correlated, which could indicate the role of these miRNAs in differentiation via repressing MYC.

Differential expression of recently made mRNA and miRNAs during the differentiation time course

We further performed a differential expression analysis across our timeline with 4 sets of data: total and labeled expression of mRNA and miRNA. Overall, there are more genes being upregulated than downregulated. We observed that 6,552 genes are upregulated and 4,901 genes are downregulated at least in one of the differentiation time points. Also, there is an increase in the number of genes that are upregulated and downregulated during the differentiation time course. We observe a similar trend in labeled genes, with 4,298 genes with faster dynamics and 3,545 genes with lower dynamics in at least one of the differential time points. There is also an interesting split in the population of genes that are being differentially regulated, which might indicate two different processes. The miRNAs behave similarly with 190 miRNAs being upregulated and 143 downregulated in at least one time point. However, we observe a distinctly higher number of upregulated miRNAs (102) in the earliest time point of 3 hours compared to the rest of the time course. There are many genes and miRNAs that are affected by or are orchestrating the differentiation of HL-60 into macrophages.

DISCUSSION

We initially established micro-TUC-seq, a method for detecting and quantifying recently made mature miRNAs in GM12878, and further applied it to HL-60 macrophage differentiation as a case study. We used 4SU labeling followed by osmium treatment, miRNA-seq library building and Illumina sequencing. In as short as 2 hours of labeling, we were able to detect 2% of the reads as labeled 10 times more than the noise we detected in our control runs. Although there are very few miRNAs with three or less Ts in their sequences, our method is able to detect reads with as low as 1 T in their sequence, showing that micro-TUC-seq has a good coverage across the miRNome and potentially would only miss less than 1% of miRNAs from miRBase database (Fig. 4.1A). Not only can we detect the number of miRNAs that are expressed at minimum 2 CPM during our pulse time, we can also quantify the amount of transcription for each miRNA during this time by measuring the fraction of labeled miRNA with respect to the total expression of that miRNA.

We performed micro-TUC-seq on cycling GM12878 cells for varying labeling times and observed an increase number of miRNAs detected as labeled. The transcription dynamics of miRNA can be estimated based on the shortest labeling time that the miRNA is detected in as labeled and is consistently detected in the longer labeling times. This measure furthermore corresponds well with our alternative estimation of transcription rate by considering the fraction of mapped reads to each miRNA that are labeled (Fig. 4.3). Considering the high reproducibility of total and labeled expression between the replicates in micro-TUC-seq, the fraction of them can present a robust way of estimating measurements of transcription dynamics.

We then decided to apply micro-TUC-seq to study the dynamics of miRNA and mRNA during HL-60 differentiation into macrophages. To this end, we performed micro-TUC-seq with 3 hours labeling time on a differentiation time course of 3 to 96 hours. In order to understand the dynamics of miRNA in the context of their functionality, we also collected regular TUC-seq data at these time points. By investigating MYC as one of the key regulators we can confirm its downregulation throughout the time course, however the labeled expressions also follow the same trend. We observe upregulation of a few miRNAs that are known to target MYC throughout the differentiation time course. Furthermore, there is a group of early risers amongst these miRNAs whereas another group of miRNAs increase their expression in the later time points. Finally, we performed differential expression analysis and observed a high number of genes and miRNAs with differential expression throughout the differentiation time course. Addition of the labeled differential expression analysis helps us identify an additional 7,843 genes and 117 miRNAs that have differential transcription rates. Although the TUC-seq and micro-TUC-seq data from HL-60 differentiation time course provide us with many layers of information, many more quality checks and further analysis is needed to truly decipher dynamics and regulatory modules involving miRNAs from this data.

FIGURES

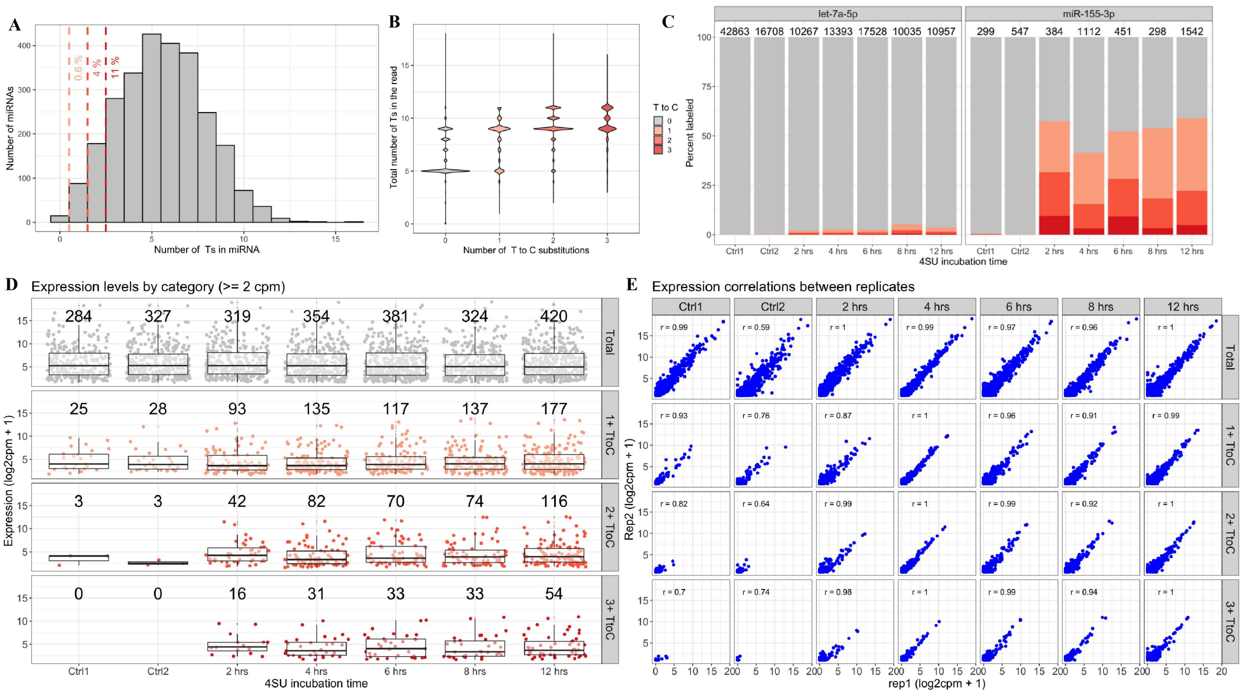

Figure 4.1. Identification of recently synthesized miRNA in GM12878 by micro-TUC-seq. **a)** Distribution of human mature miRNAs annotated by miRBase V22 based on the number of Ts in the sequence of miRNA. **b**) Violin plot of mapped reads from one of the control replicates showing the number of Ts in the miRNA it mapped to, with respect to its labeling category (number of T→C). **c**) Expression levels of two representative miRNAs, let-7a-5p (highly

expressed) and miR-155-3p (medium expressed) with the percentage of reads in each labeling category. The number on the top is the total expression level in CPM in each sample. **d**) Expression levels of miRNAs with more than or equal to 2 CPM in each of the labeling categories. The total number of genes is indicated on top of each of the boxplots. **e**) Expression correlation between the replicate of each sample grouped in their labeling categories. The number represents the Pearson correlation coefficient between the log2 expressions.

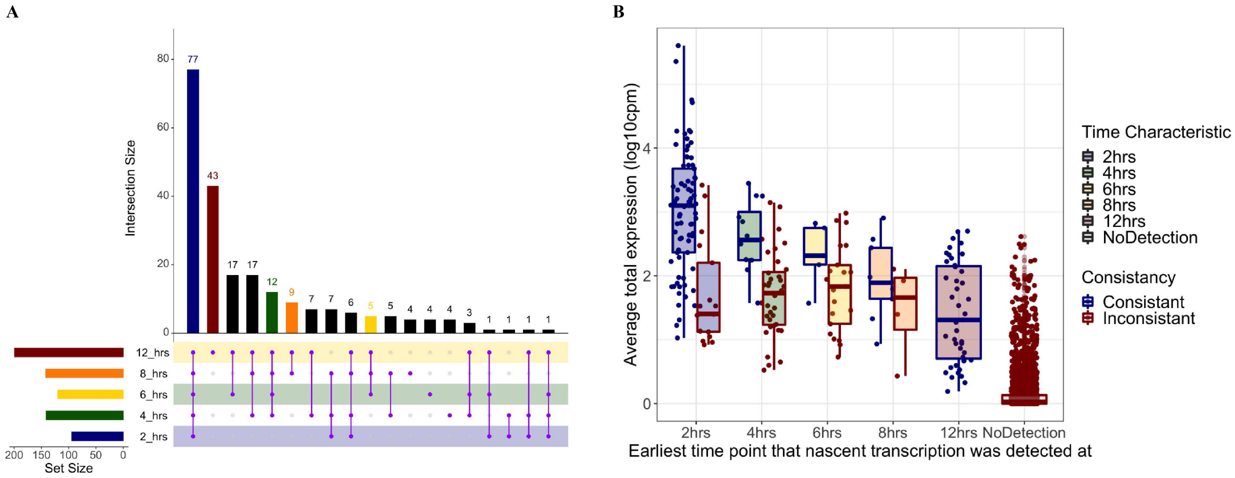

Figure 4.2. Grouping of miRNAs based on the pulse time required for the detection of their labeled expression. **a)** An upset plot from the overlap of groups of miRNAs detected in different pulse times (represented by different colors). The horizontal bars represent the number of labeled miRNAs detected in each sample. The vertical bars show the number of miRNAs in the unique combination of samples represented by the points underneath. **b**) Average expression of miRNAs across samples for different groups of miRNAs, based on the pulse time in which they were first detected as labeled, and colored by the whether they have been detected consistently in all the subsequent times or not.

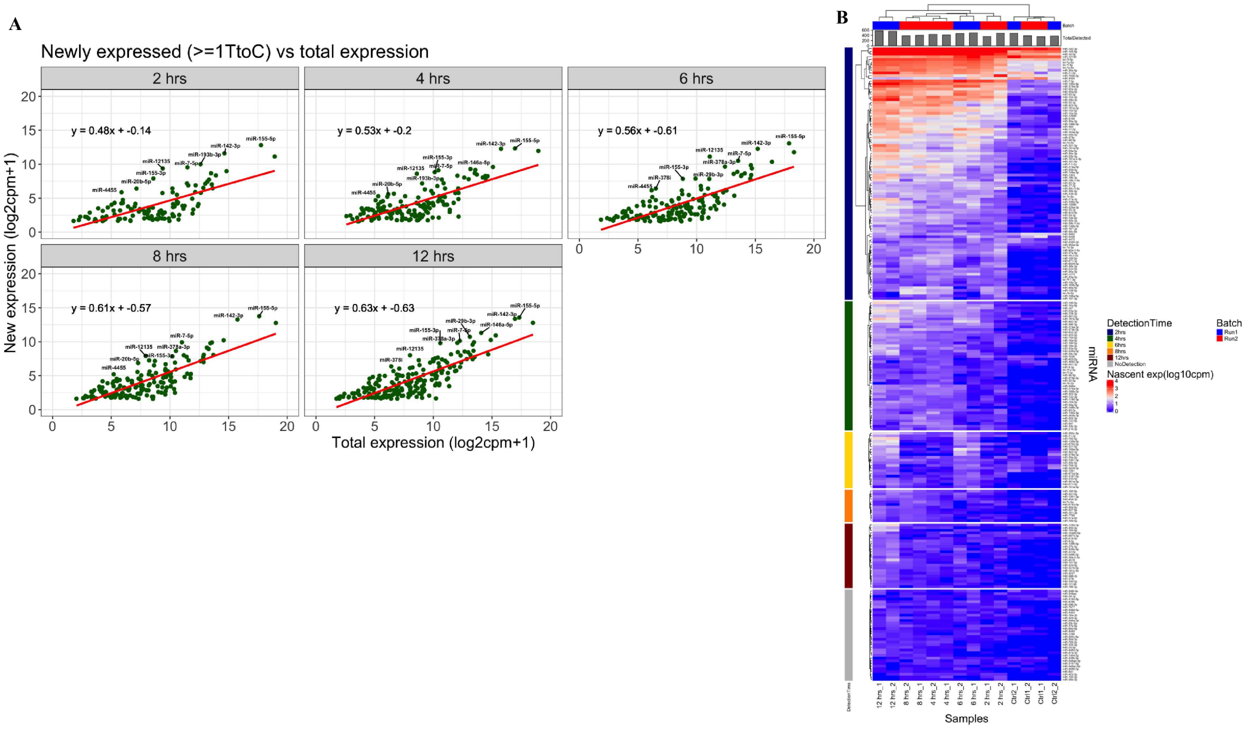

Figure 4.3. Expression dynamics of miRNAs in GM12878. **a)** Scatter plots of total and labeled expression for the miRNAs detected as labeled at each pulse time. The red line shows the regression line of the point with its corresponding equation written above it. **b)** Heatmap of labeled expression levels with miRNAs in rows grouped by the first pulse time they were detected in.

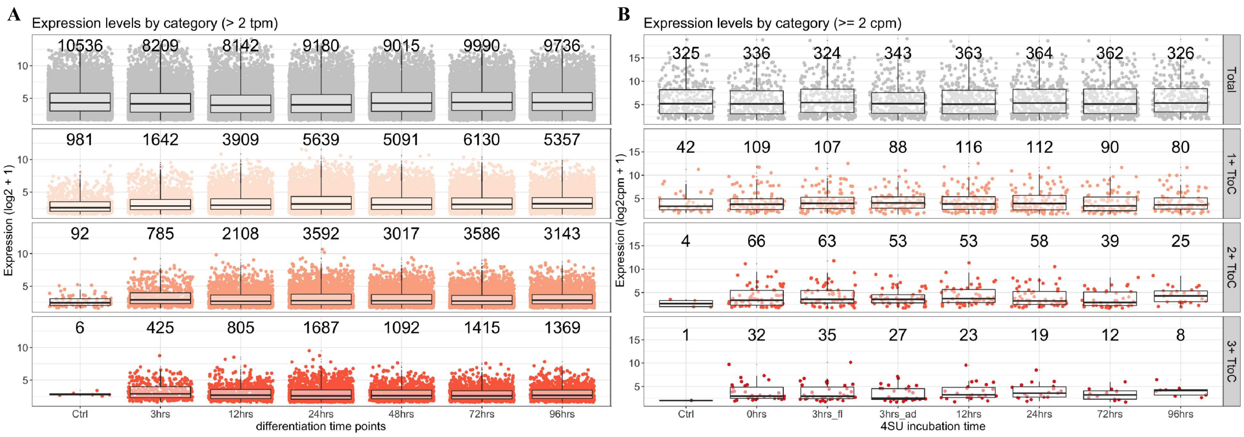

Figure 4.4. mRNAs and miRNAs detected as expressed or labeled and expressed through the HL-60 differentiation time-course. **a**) Expression levels of mRNAs with more than or equal to 2 CPM in each of the labeling categories. The total number of genes is indicated on top of the boxplots. **b**) Same graph as in a for miRNAs in HL-60.

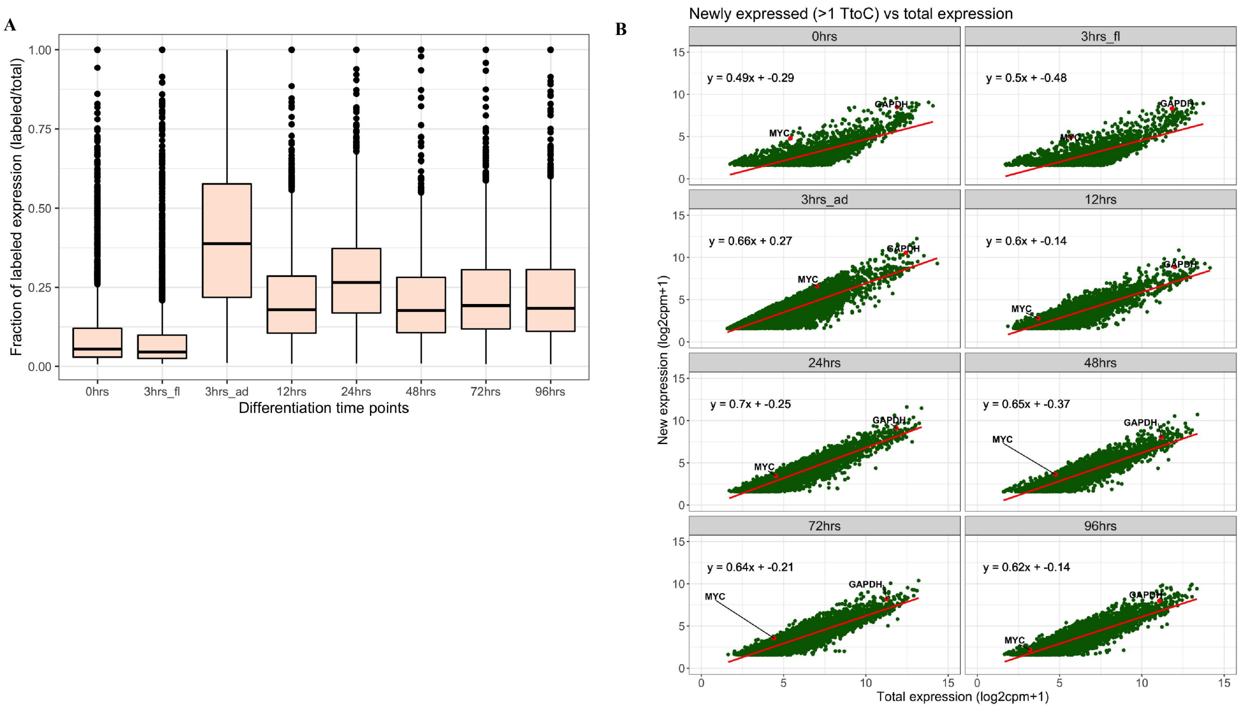

Figure 4.5. Dynamics of mRNA expression during HL-60 differentiation. **a**) Boxplot of expression rates (as the fraction of labeled reads to total reads) for each time point. **b**) Scatter plots of labeled RNA expression vs. total expression. MYC and GAPDH genes are labeled for reference.

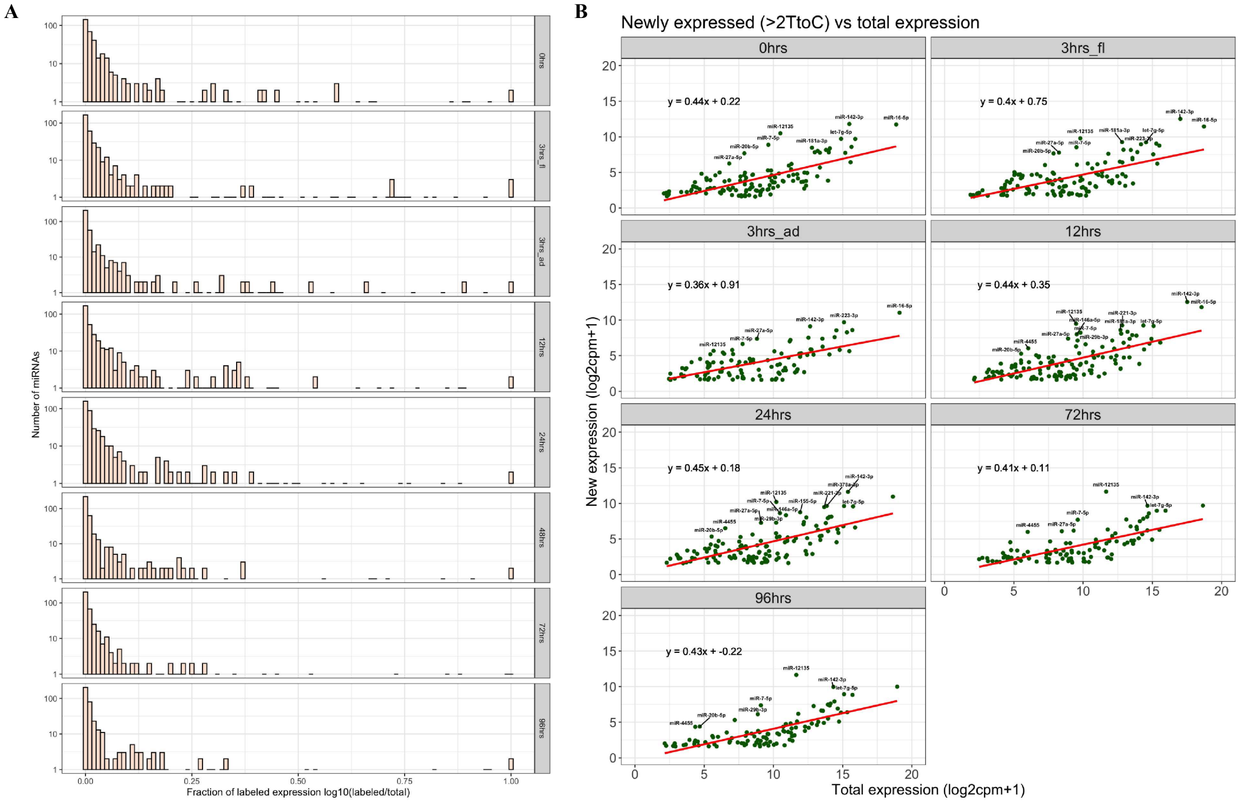

Figure 4.6. Dynamics of miRNA expression during HL-60 differentiation. **a**) Histogram of expression rates (as the fraction of labeled reads to total reads) for each time point. **b**) Scatter plots of labeled miRNA expression vs total expression. MiRNAs with higher transcription rates are labeled.

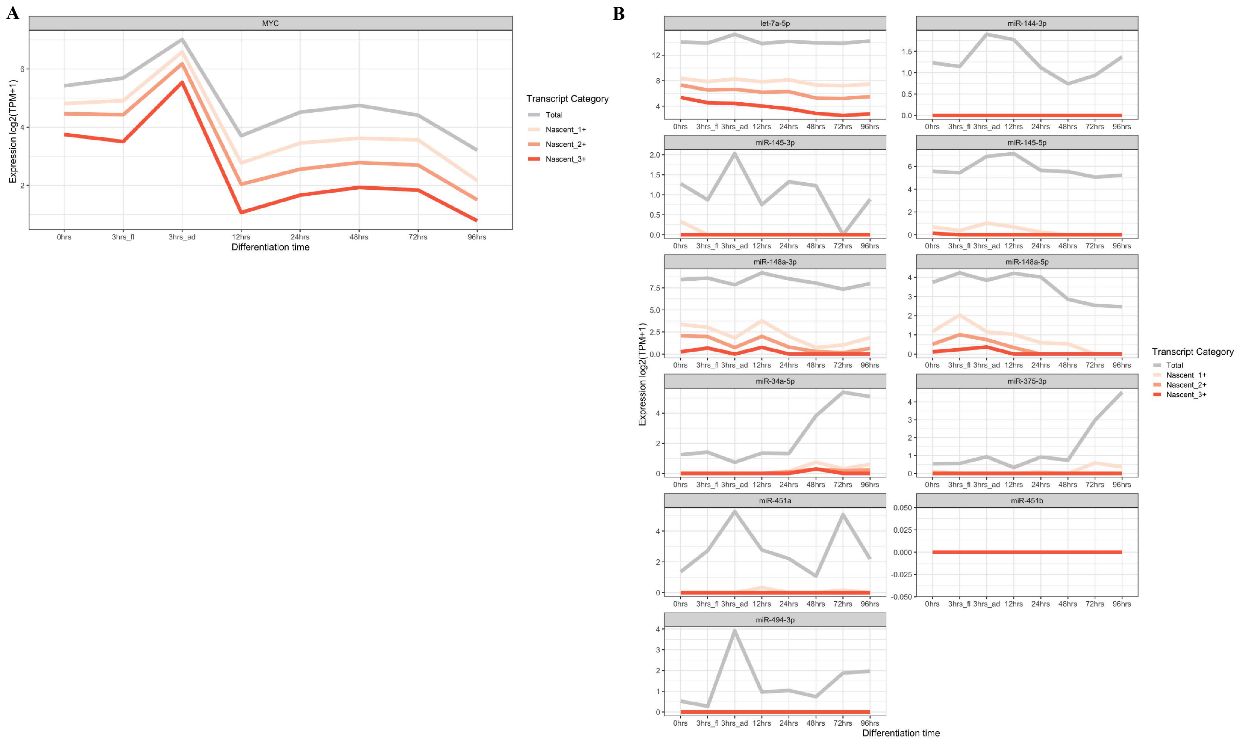

Figure 4.7. Dynamic expression of MYC and the miRNAs targeting MYC. **a**) Expression profile of MYC during the HL-60 differentiation time course. **b**) Expression profile of 11 miRNAs known to target MYC.

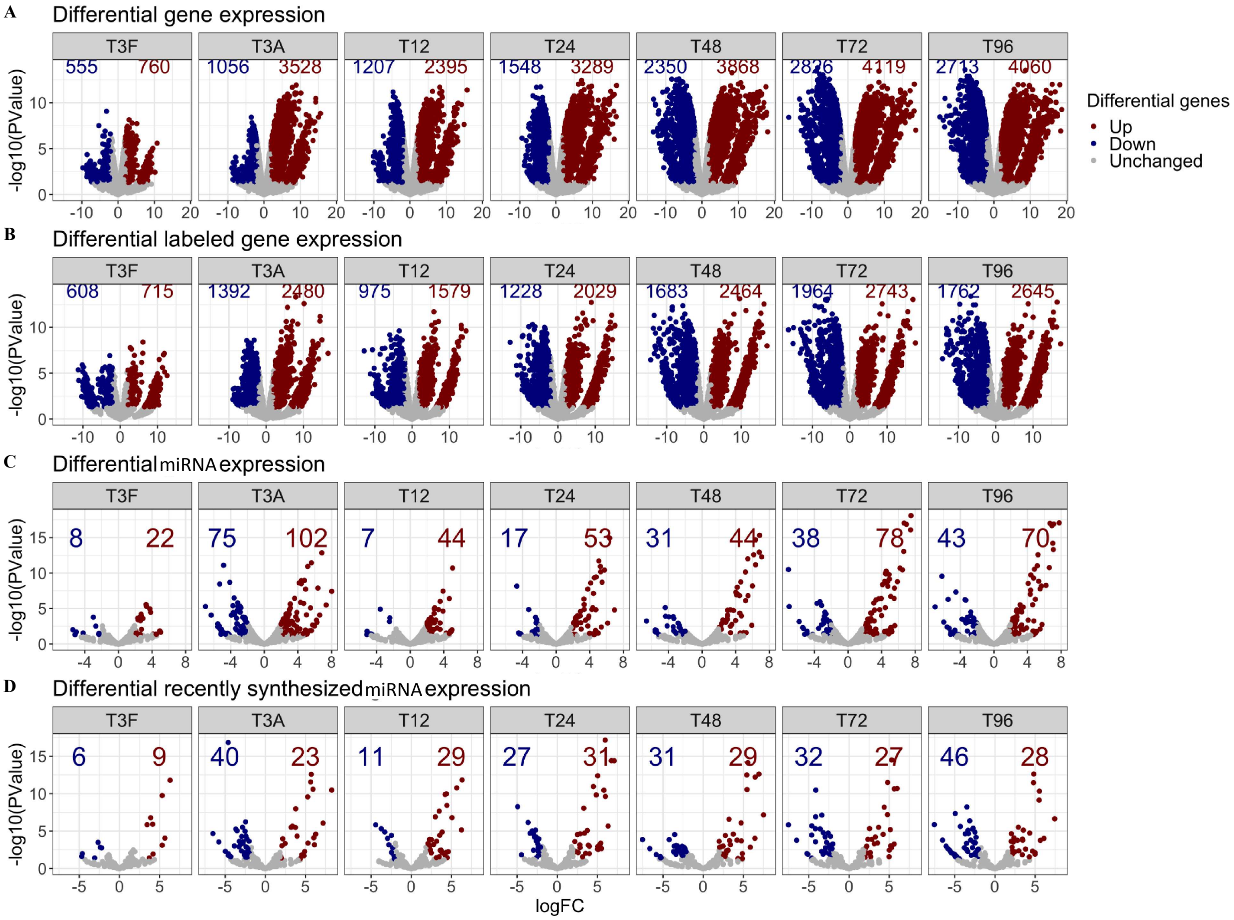

Figure 4.8. Differential expression analysis of labeled and total mRNA and miRNA.

Volcano plots showing the differential expressed **a**) genes, **b**) labeled genes, **c**) miRNAs and **d**) labeled miRNAs throughout the differentiation of HL-60 into macrophages. A cut off of |logFC| > 2 and P-value of < 0.05 has been used to call upregulated (in red) and downregulated (in blue) genes. The T3F refers to HL-60 cells that were washed out at T3, and T3A refers to the cells that remained adherent to the dish.

TABLES

Sample	Description	Raw reads	Mapped reads	Mapping rate%	MiRNAs detected (>=2CPM)
GM81	Ctrl1 R1	14,411,581	3,672,613	25%	430
GM82	Ctrl1 R2	8,927,976	3,977,265	45%	354
GM121	Ctrl2 R1	10,837,826	5,712,634	53%	305
GM122	Ctrl2 R2	10,147,358	4,756,803	47%	355
GM123	2 hrs 4SU R1	11,994,529	5,383,160	45%	330
GM124	2 hrs 4SU R2	15,373,420	5,865,665	38%	428
GM125	4 hrs 4SU R1	8,954,299	3,890,722	43%	387
GM126	4 hrs 4SU R2	11,211,463	4,964,013	44%	387
GM83	6 hrs 4SU R1	11,135,583	3,759,681	34%	444
GM84	6 hrs 4SU R2	8,603,977	3,060,238	36%	448
GM127	8 hrs 4SU R1	9,749,793	5,059,814	52%	374
GM128	8 hrs 4SU R2	8,467,359	3,969,710	47%	352
GM85	12 hrs 4SU R1	8,388,856	1,451,977	17%	516
GM86	12 hrs 4SU R2	9,286,815	1,935,784	21%	487

Table 4.1. Summary of sequencing data for GM12878 micro-TUC-seq experiments

Sample	Description	Raw reads	Mapped reads	Mapping rate%	Genes detected (>=2TPM)
HL10	Osmium Ctrl R1	24,782,990	17,554,803	71%	11,361
HL11	TUC-seq T0 R1	8,909,459	5,758,528	65%	7,875
HL12	TUC-seq T3_fl R1	14,016,636	9,100,487	65%	8,957
HL15	TUC-seq T12 R1	6,890,460	4,240,296	62%	8,704
HL16	TUC-seq T24 R1	30,775,846	20,600,435	67%	10,411
HL17	TUC-seq T48 R1	10,086,772	6,290,270	62%	9,678
HL18	TUC-seq T72 R1	9,161,481	6,254,731	68%	10,835
HL19	TUC-seq T96 R1	5,106,747	3,252,068	64%	10,096
HL20	Osmium Ctrl R1	18,615,821	13,580,556	73%	11,040
HL22	TUC-seq T3_fl R1	9,903,352	6,346,492	64%	8,526
HL23	TUC-seq T3_ad R1	28,276,681	16,086,591	57%	11,495
HL25	TUC-seq T12 R1	14,415,128	8,938,905	62%	9,150
HL26	TUC-seq T24 R1	17,894,263	11,803,778	66%	9,455
HL27	TUC-seq T48 R1	10,574,253	7,354,824	70%	9,818
HL28	TUC-seq T72 R1	8,042,910	5,692,294	71%	10,596
HL29	TUC-seq T96 R1	21,476,245	15,381,842	72%	11,232

Table 4.2. Summary of sequencing data for HL-60 TUC-seq experiments

Sample	Description	Raw reads	Mapped reads	Mapping rate%	MiRNAs detected (>=2CPM)
HL10	Osmium Ctrl R1	9,706,684	5,065,878	52%	387
HL11	micro-TUC-seq T0 R1	11,086,777	5,367,505	48%	364
HL12	micro-TUC-seq T3_fl R1	9,882,070	4,870,202	49%	351
HL13	micro-TUC-seq T3_ad R1	14,950,910	5,026,454	34%	364
HL15	micro-TUC-seq T12 R1	4,249,855	1,984,208	47%	415
HL16	micro-TUC-seq T24 R1	10,313,949	4,678,523	45%	438
HL18	micro-TUC-seq T72 R1	9,036,498	4,020,960	44%	394
HL19	micro-TUC-seq T96 R1	8,097,510	3,121,225	39%	394
HL20	Osmium Ctrl R1	13,010,187	7,286,860	56%	333
HL21	micro-TUC-seq T0 R1	12,882,134	6,388,488	50%	366
HL22	micro-TUC-seq T3_fl R1	6,662,759	2,847,467	43%	378
HL23	micro-TUC-seq T3_ad R1	9,917,602	3,681,771	37%	410
HL25	micro-TUC-seq T12 R1	13,160,916	6,192,145	47%	374
HL26	micro-TUC-seq T24 R1	15,812,399	9,501,188	60%	375
HL27	micro-TUC-seq T48 R1	11,593,505	4,645,989	40%	390
HL28	micro-TUC-seq T72 R1	10,318,663	4,804,339	47%	334
HL29	micro-TUC-seq T96 R1	8,045,630	4,688,429	58%	366

Table 4.3. Summary of sequencing data for HL-60 micro-TUC-seq experiments

METHODS

Sample collection and RNA extraction

GM12878 cells were obtained from Corriel Institute and were cultured in accordance with ENCODE protocols (www.encodeproject.org). The cells were passed every two to three days at 200k-500k cells/mL density and were harvested for the experiments at 500k-1M cells/mL. HL-60 cells were obtained from ATCC (CCL-240) and cultured according to ENCODE protocols. The cells were passed every two to three days at >1M cells/ml. We extracted the RNA using QIAGEN RNeasy Plus kit (Cat. No. 74134) following the protocol 1 for capturing total RNA including miRNAs.

TUC-seq sample preparation

4-thiouridine was obtained from Sigma Aldrich (T4609) and used fresh at a working concentration of 200 mM. For each TUC-seq experiment, 10-15M cells were spun down and resuspended in 10-15 mL of fresh media with 4SU added at a final concentration of 1mM (no 4SU was added for the osmium controls). The cells were incubated with 4SU for 2-12 hours and harvested for RNA extraction. The RNA was then treated with OsO_4 solution for 3 hours at room temperature in dark. The osmium solution was prepared fresh every time by mixing 20 µl of 1mM OsO_4 (Sigma Aldrich, 201030) with 4µl of 2M NH_4Cl at pH 8.8 and 1µl of RNasin Plus RNase inhibitor (Promega, N2615) for every 10µg of RNA. The RNA was then purified using Zymo RNA cleanup kit (R1015).

HL-60 differentiation into macrophages

The HL-60 cells were grown to 1-2 million/mL cell density, then they were sub-cultured in a dish with differentiation media of DMEM with a final concentration of 10µM PMA (Sigma Aldrich, P8139). After three hours the floating cells were collected, and fresh differentiation media was added to the remaining adherent cells. We collected samples at 0, 3, 12, 24, 48, 72 and 96 hours after adding PMA. For all the TUC-seq samples, the 4SU was added to the media 3 hours before the collection time point. We also collected a no-4SU 0 hour time point for the osmium control.

mRNA library preparation and sequencing

Starting from 30-50ng of the same cDNA, we followed the Illumina tagmentation protocol using the Nextera DNA Flex Library Prep Kit to generate Illumina short-read libraries.

We checked the concentration of the libraries with Qubit and got the average length of the library using BioAnalyzer. We then performed a 2x43 paired-end sequencing on our NextSeq 500 instrument.

miRNA library preparation and sequencing

Starting from 1-1.5 µg of total RNA, we followed the modified version of multiplexed miRNA-seq protocol from ENCODE (Alon, 2011). We checked the concentration of the libraries with Qubit and used 140bp as the average length of the library for determining the loading concentration of the library on the sequencer. We then performed a 2x43 paired-end sequencing on our NextSeq 500 instrument.

mRNA data processing

The raw reads from Illumina were first filtered for any PCR duplicates using fastuniq, trimmed the adaptors by Cutadapt (v2.5; parameters: -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -m 20). Then the reads were mapped to human genome reference (GRCh38.p12) using STAR aligner (v2.6.0c; parameters: --outFilterMismatchNmax 15 --outFilterMismatchNoverReadLmax 0.07 --outFilterMultimapNmax 10 --outSAMunmapped None --outSAMattributes MD NM --alignIntronMax 1000000 --alignIntronMin 20 --alignMatesGapMax 1000000 --outSAMtype BAM Unsorted SortedByCoordinate --quantMode TranscriptomeSAM --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1) using the gencode.v29.primary_assembly.annotation.gtf for sjdbGTFfile option.

mRNA calling of the label reads

We ran a custom python script (mismatch_analysis_ill.py) to annotate each of the mapped reads with the number of different substitution events. The script uses the MD tag to tally this number for each read and report them for each paired-end read. The script also breaks down the alignment file into sub-files of reads with > 0, > 1, > 2 and > 3 T→C substitutions. Finally, we quantify each category as follows: first we used HTSeq (V0.11.2; parameters: -r pos --nonunique all -s no -i transcript_id) to count the reads in each category. In addition, we used Kallisto (V0.43.1; parameters: --single -l -s) as a single-read option with the mean and standard deviation of the libraries approximated from STAR alignment to the transcriptome. We calculate a normalization factor by dividing the sub-categories of reads (>1,>2 and >3 T→C) by the counts from total reads (>0 T→C). We then apply these normalization factors to the Kallisto TPMs for the total RNA (>0 T→C) to obtain TPMs for each of the subcategories.

miRNA data processing

The raw fastq data was first demultiplexed to obtain the reads corresponding to each miRNA-seq sample. Then the 3' and 5' adaptors were removed using Cutadapt (V2.5; parameters: -e 0.25 --match-read-wildcards for the 3' adapter and -e 0.34 --match-read-wildcards --no-indels -m 15 -O 6 -n 1 for the 5' adaptors – a set of 4). Then the reads were aligned to the miRBase human mature miRNA reference (V22) using STAR (V2.7.3a ; parameters: --alignEndsType EndToEnd --outFilterMismatchNmax 3 --outSAMattributes MD NM --outFilterMultimapNmax 10 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0

--outFilterMatchNmin 16 --alignIntronMax 1). The mapped reads were filtered for including primary alignments with minimum 2 mapping quality scores.

miRNA calling of the label reads

We ran a custom python script (mismatch_analysis_miR.py) to annotate each of the mapped reads with the number different substitution events. The script uses the MD tag to tally this number for each read and report them for each single-end read. The script also breaks down the alignment file into sub-files of reads with > 0, > 1, > 2 and > 3 T→C substitutions. Finally, we quantify each category using eXpress (V1.5.1; parameters: --no-bias-correct). The counts obtained from eXpress are then normalized into CPMs.

REFERENCES

Agrawal, Raman, et al. "The MiR-30 MiRNA Family Regulates Xenopus Pronephros Development and Targets the Transcription Factor Xlim1/Lhx1." *Development*, vol. 136, no. 23, 2009, pp. 3927–36, doi:10.1242/dev.037432.

Alberti, Chiara, and Luisa Cochella. "A Framework for Understanding the Roles of MiRNAs in Animal Development." *Development (Cambridge)*, vol. 144, no. 14, 2017, pp. 2548–59, doi:10.1242/dev.146613.

Alkallas, Rached, et al. "Inference of RNA Decay Rate from Transcriptional Profiling Highlights the Regulatory Programs of Alzheimer's Disease." *Nature Communications*, vol. 8, no. 1, Springer US, 2017, pp. 1–11, doi:10.1038/s41467-017-00867-z.

Ambros, Victor. "A Hierarchy of Regulatory Genes Controls a Larva-to-Adult Developmental Switch in C. Elegans." *Cell*, vol. 57, no. 1, 1989, pp. 49–57, doi:10.1016/0092-8674(89)90171-2.

Bail, Sophie, et al. "Differential Regulation of MicroRNA Stability." *Rna*, vol. 16, no. 5, 2010, pp. 1032–39, doi:10.1261/rna.1851510.

Bernstein, Emily, et al. "Dicer Is Essential for Mouse Development." *Nature Genetics*, vol. 35, no. 3, Springer Nature, Oct. 2003, pp. 215–17, doi:10.1038/ng1253.

Breitman, T. R., et al. "Induction of Differentiation of the Human Promyelocytic Leukemia Cell Line (HL-60) by Retinoic Acid." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 5 I, 1980, pp. 2936–40, doi:10.1073/pnas.77.5.2936.

Chandra, Swati, et al. "Role of MiRNAs in Development and Disease: Lessons Learnt from Small Organisms." *Life Sciences*, vol. 185, no. 5, Elsevier Inc., 2017, pp. 8–14, doi:10.1016/j.lfs.2017.07.017.

Coolen, Marion, and Laure Bally-cuif. *MicroRNAs in Brain Development and Physiology*. 2009, doi:10.1016/j.conb.2009.09.006.

Crist, Colin G., et al. "Muscle Stem Cell Behavior Is Modified by MicroRNA-27 Regulation of Pax3 Expression." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 32, 2009, pp. 13383–87, doi:10.1073/pnas.0900210106.

Dakir, El Habib, and Faustino Mollinedo. "Genome-Wide MiRNA Profiling and Pivotal Roles of MiRs 125a-5p and 17-92 Cluster in Human Neutrophil Maturation and Differentiation of Acute Myeloid Leukemia Cells." *Oncotarget*, vol. 10, no. 51, 2019, pp. 5313–31, doi:10.18632/oncotarget.27123.

Duffy, Erin E., et al. "Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry." *Molecular Cell*, vol. 59, no. 5, Elsevier Inc., 2015, pp. 858–66, doi:10.1016/j.molcel.2015.07.023.

Garzon, Ramiro, et al. "MicroRNA 29b Functions in Acute Myeloid Leukemia." *Blood*, vol. 114, no. 26, 2009, pp. 5331–41, doi:10.1182/blood-2009-03-211938.

Herzog, Veronika A., et al. "Thiol-Linked Alkylation of RNA to Assess Expression Dynamics." *Nature Methods*, vol. 14, no. 12, Springer Nature, Sept. 2017, pp. 1198–204, doi:10.1038/nmeth.4435.

Huntzinger, Eric, and Elisa Izaurralde. "Gene Silencing by MicroRNAs: Contributions of Translational Repression and MRNA Decay." *Nature Reviews Genetics*, vol. 12, no. 2, Nature Publishing Group, 2011, pp. 99–110, doi:10.1038/nrg2936.

Iwakawa, Hiro oki, and Yukihide Tomari. "The Functions of MicroRNAs: MRNA Decay and Translational Repression." *Trends in Cell Biology*, vol. 25, no. 11, Elsevier Ltd, 2015, pp. 651–65, doi:10.1016/j.tcb.2015.07.011.

Kasashima, Katsumi, et al. "Altered Expression Profiles of MicroRNAs during TPA-Induced Differentiation of HL-60 Cells." *Biochemical and Biophysical Research Communications*, vol. 322, no. 2, 2004, pp. 403–10, doi:10.1016/j.bbrc.2004.07.130.

Le, Thuc Duy, et al. "Ensemble Methods for MiRNA Target Prediction from Expression Data." *PLoS ONE*, vol. 10, no. 6, 2015, doi:10.1371/journal.pone.0131627.

Lee, Rosalind C., et al. "The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14." *Cell*, vol. 75, no. 5, Elsevier {BV}, Dec. 1993, pp. 843–54, doi:10.1016/0092-8674(93)90529-Y.

Li, Yang, et al. "Genome-Wide Analysis of Human MicroRNA Stability." *BioMed Research International*, vol. 2013, 2013, doi:10.1155/2013/368975.

Lusser, Alexandra, et al. "Thiouridine-to-Cytidine Conversion Sequencing (TUC-Seq) to Measure MRNA Transcription and Degradation Rates." *The Eukaryotic RNA Exosome*, 2020, pp. 191–211.

Mangelsdorf, David J., et al. "In a Human Promyelocytic Leukemia Cell Line (HL-60) : Receptor-Mediated Maturation to Macrophage-like Cells." *The Journal of Cell Biology*, vol. 3, 1984, pp. 1–8.

Marzi, Matteo J., and Francesco Nicassio. "Uncovering the Stability of Mature MiRNAs by 4-Thio- Uridine Metabolic Labeling." *MiRNA Biogenesis*, 2018, pp. 141–52.

Murao, Shin Ichi, et al. "Control of Macrophage Cell Differentiation in Human Promyelocytic HL-60 Leukemia Cells by 1,25-Dihydroxyvitamin D3 and Phorbol-12-Myristate-13-

Acetate." *Cancer Research*, vol. 43, no. 10, 1983, pp. 4989–96.

Ovando-Vázquez, Cesaré, et al. "Improving MicroRNA Target Prediction with Gene Expression Profiles." *BMC Genomics*, vol. 17, no. 1, BMC Genomics, 2016, pp. 1–13, doi:10.1186/s12864-016-2695-1.

Paulsen, Michelle T., et al. "Use of Bru-Seq and BruChase-Seq for Genome-Wide Assessment of the Synthesis and Stability of RNA." *Methods*, 2014, doi:10.1016/j.ymeth.2013.08.015.

Rahmanian, Sorena, et al. "Dynamics of MicroRNA Expression during Mouse Prenatal Development." *Genome Research*, vol. 29, no. 11, 2019, pp. 1900–09, doi:10.1101/gr.248997.119.

Ramirez, Ricardo N., et al. "Dynamic Gene Regulatory Networks of Human Myeloid Differentiation." *Cell Systems*, vol. 4, no. 4, Elsevier Inc., 2017, pp. 416-429.e3, doi:10.1016/j.cels.2017.03.005.

Sayed, Danish, and Maha Abdellatif. "Micrornas in Development and Disease." *Physiological Reviews*, vol. 91, no. 3, 2011, pp. 827–87, doi:10.1152/physrev.00006.2010.

Schwalb, Björn, et al. "TT-Seq Maps the Human Transient Transcriptome." *Science*, vol. 352, no. 6290, 2016, pp. 1225–28, doi:10.1126/science.aad9841.

Valencia-Sanchez, Marco Antonio, et al. "Control of Translation and MRNA Degradation by MiRNAs and SiRNAs." *Genes and Development*, vol. 20, no. 5, 2006, pp. 515–24, doi:10.1101/gad.1399806.

Visvanathan, Jaya, et al. "The MicroRNA MiR-124 Antagonizes the Anti-Neural REST/SCP1 Pathway during Embryonic CNS Development." *Genes and Development*, vol. 21, no. 7, 2007, pp. 744–49, doi:10.1101/gad.1519107.

Yangming Wang, et al. "DGCR8 Is Essential for MicroRNA Biogenesis and Silencing of Embryonic Stem Cell Self-Renewal." *Nature Genetics*, vol. 39, no. 3, 2007, pp. 380–85, doi:10.1038/ng1969.

Zhou, Li, et al. "Importance of MiRNA Stability and Alternative Primary MiRNA Isoforms in Gene Regulation during Drosophila Development." *ELife*, vol. 7, 2018, pp. 1–31, doi:10.7554/eLife.38389.

CHAPTER 5

Future directions

Advances to miRNA-mRNA integrative analysis for mouse embryonic timeline

We used our miRNA-mRNA analysis pipeline for mouse embryonic development to predict mRNA targets for each miRNA cluster by enrichment analysis and scoring of the miRNA-mRNA cluster-interactions. The current implementation used miRNAtap which is an ensemble utility that combines predictions from 5 different sources with a set of default parameters (Pajak and Simpson). However, the field of miRNA target prediction is growing quickly and some of our underlying algorithms and databases have been updated since our original analysis (Kozomara et al.; Chen and Wang). For example, a newer ensemble tool called miRwalk2 utilizes 13 different algorithms in conjunction with experimental databases (Sticht et al.). We could improve our target enrichment analysis by switching to miRwalk2 and by carefully choosing the experimental databases to be included as well as by fine tuning of the parameters to use for each algorithm.

In our initial approach we first defined the clusters of miRNA and mRNA based on their expression profiles. We then looked for clusters of miRNAs with mean expression that is partially negatively correlated with mRNA clusters to identify potential mRNA targets for the given miRNAs. One major assumption for the correlation analysis is that changes in the expression of mRNA clusters in respective tissues are solely driven by post-transcriptional regulation by miRNA. We know this to be false as the majority of changes in gene expression are regulated by a number of critical transcriptional regulatory mechanisms such as methylation, chromatin remodeling, and transcription factor binding to cis-regulatory elements. A more thorough approach to assessing the dynamics of expression during embryonic development would take into consideration other major transcriptional regulators of expression. We could better model mRNA expression levels by integrating data on regulatory elements that have been collected for these same samples, which are available from ENCODE portal (Gorkin et al.). The available data includes ChIP-seq data for eight major histone modifications, open chromatin features (ATAC-seq and DNase-seq) and DNA methylation data (WGBS) for the majority of the samples. We will use this additional data to train a regression model of mRNA expression based on these epigenomic features for the tissues in which changes in miRNA cluster expression are not significantly anti-correlated with mRNA expression. Then we will apply the model to establish a baseline of predicted mRNA expression in the other tissues where the miRNA clusters are significantly anti-correlated. Finally, we will use this baseline correction to find residual mRNA expression attributable to post-transcriptional regulation by miRNA.

Once more taking advantage of the fact that we have multi-omics data available for this developmental time-course, we could look for regulatory modules that integrate the different layers of regulation. This analysis could also help us understand the underlying regulations that lead to differential expression of miRNAs. One approach to this analysis could be to used linked self-organizing maps to create SOMs based on each epigenetic feature, and then to generate a draft regulatory networks by linking the SOMs together (Jansen et al.). The resulting networks would integrate both microRNAs as well as other pre-transcriptional regulators into these networks.

Finally, we could use long-read RNA sequencing on these samples to refine and improve our analysis. The majority of miRNA functional target sites are known to be in the 3'UTR (Xu et al.). However, we are unable to distinguish between transcripts with variable 3' UTR regions using short-read RNA-sequencing data. Furthermore, some miRNA target sites may fall in alternatively spliced exons thereby causing different isoforms of a gene to be regulated by different miRNAs. All these issues can be resolved by utilizing long read-read single isoform resolution sequencing of these samples to move the analysis from gene-level to transcript-level associations.

Improvements to long-TUC-seq protocol and analysis pipeline

Although 4SU is a naturally occurring derivative of uridine, studies have shown that higher concentrations and longer incubation times can inhibits rRNA synthesis and be toxic to some cells (Burger et al.). The toxicity of 4SU varies from cell line to cell line, and another study showed no toxicity effect on expression after 12 hrs of 1mM 4SU in the Hek293 cell line (Hafner et al.). With the current concentration of 4SU and labeling time (1mM for 1hr), we have not observed any significant change in the viability of GM12878 cells. However, reducing both the incubation time and the concentration of 4SU for our study may be beneficial in order to mRNAs with higher turnover. In order to conduct our proof of concept experiments we decided on a higher concentration of 4SU and longer incubation time. Now that we have established a working protocol for long-TUC-seq, we will look into optimization of our protocol with regards to labeling time and concentration in favor of lower toxicity and to reduce any potential interference with the underlying biology of the cell. Additionally, a shorter pulse time will allow for more accurate estimation of the rates of labeled transcripts with faster turnover.

A major advantage of using long-read sequencing for detecting labeled RNA is that it significantly increases signal to noise ratio (SNR) compared to the short-read sequencing. This allows us to detect labeled RNA molecules with much higher sensitivity and without losing any specificity. Our static lower SNR threshold of 20 T→C undermines our ability to detect labeled transcripts using long-TUC-seq. We are potentially missing labeled transcripts due to length or number of uridines in their sequence. We can improve our calling rate and sensitivity by setting a dynamic threshold that considers other factors such as the length or the U rich content of the transcript. Finally, a more comprehensive model could be implemented after careful cataloging of all the possible features that affect the detection of a labeled transcript. These features may include transcript model complexity (i.e. splicing events), 4SU incubation time and 4SU concentration, especially for samples or studies that vary these parameters to maintain more physiological conditions. Finally, the model can be trained on long-TUC-seq data obtained from a set of synthesized labeled ERCC or SIRV reference transcripts.

Although Pacbio sequencing has improved significantly over the past years by reducing its final error rates and increasing its theoretical output to 8 million reads per SMRT cell, the library preparation remains laborious and the cost per experiment is still rather high. Furthermore, the time it takes for sequencing and pre-processing the data is cumbersome. Similar to Pacbio technology, Oxford Nanopore Technologies (ONT) direct RNA sequencing has improved significantly (Workman et al.). Recent kits have enabled sequencing of 1 million full-length reads on one flowcell within 48 hours, with slightly improved error rates of 92% sequence identity (Parker et al.). Once the long-TUC-seq basic protocol and analysis pipeline is established for PacBio, it should be easily transferable to Nanopore directRNA sequencing. The main steps of 4SU incorporation and its conversion to C happens at the level of RNA which would stay the same for both platforms. For Nanopore, we can build the library for direct RNA sequencing instead of making cDNA after conversion. One advantage of this approach would be to avoid errors introduced by a reverse transcriptase or DNA polymerase. However, the detection of substitution events would be more difficult due to the higher noise of Nanopore data.

A further step in improving the long-TUC-seq protocol implementation for Nanopore would be to bypass the conversion step by directly sequencing the 4SU labeled RNA since the technology is voltage based. Previous studies have shown successful identification of modified nucleotides in direct RNA sequencing (Lorenz et al.; Parker et al.; Liu et al.). The detection of labeled reads can occur post-basecalling by looking for specific error profiles and miscall signatures, but the high error rate of direct RNA sequencing makes this task quite challenging. Another option would be to re-train the neural net used for RNA basecalling to include an additional base (i.e. 4SU). This could be done by training the neural net on a set of in-vitro synthesized labeled control RNAs and then applying the trained neural net on the actual samples for 5-base mode base calling as shown by the Nano-ID study (Maier, Gresse, Cramer, & Schwalb, 2019)

Many recent studies have shown the importance of single cell variation in gene expression in order to understand the underlying dynamics of cellular processes such as differentiation and response to viruses (Griffiths et al.; Wyler et al.). Recent studies have been exploring the possibilities of merging long-read sequencing with single-cell RNA-seq techniques. One group developed a technique to utilize UMIs in conjunction with the 10X platform to perform long-read sequencing on ONT (Lebrigand et al.). Currently, my colleagues in the Mortazavi lab are exploring the possibility of combining a newer single cell technique called split-seq (Rosenberg et al.) with long-read sequencing on PacBio. This technique is attractive as it avoids the complication of droplet fluidics and for this reason might be a good candidate to explore the possibility of single-cell long-TUC-seq. Although an establish protocol for single cell label RNA sequencing (SLAM-seq) exists, additional technologies to study the dynamics of transcription at a isoform level in single cells are needed (Erhard et al.).

Finally, the ability to identify labeled RNA with long-read sequencing opens the horizon for many alternative applications. For example, if we can couple the labeling of the RNA with the solvent accessibility of the RNA structure, we would be able to characterize the secondary structure of the RNA. This would require a chemical method for inducing a nucleotide modification via the solvent that can be altered later to introduce a specific substitution at solvent accessible regions. Another example of alternative applications could be cellular localization-specific labeling that could help resolve the spatial transcriptomics of the RNA. A third example would be to introduce an RNA-labeling mechanism via a fusion RBP in order to identify the RNA regions associated with the RBP of interest. While these methods would utilize different chemistries and molecular biology, they could leverage the nucleotide substitution and detection technique of long-TUC-seq for vast exploration of nucleic acids.

Advances to Short-TUC-seq and Micro-TUC-seq and the dynamics of HL60 differentiation:

Currently, both short-TUC-seq and micro-TUC-seq have high levels of false positives detected in the control runs. This is most likely due to the biological and technical noise that introduces T→C into the sequence. One way to avoid this high rate of false positive calling of the labeled reads would be to use a medium threshold of 2 T→C per read. Although this cut off will reduce the non-specific calling of labeled reads, it will also reduce our sensitivity especially in micro-TUC-seq where we do not have a high signal of T→C in our true positive reads because of the shortness of microRNAs. Another solution is to build a baseline of T→C calls using the control runs, then we would apply this baseline correction to our sample reads before calling them as labeled or unlabeled.

Our HL60 differentiation study is designed as a multi-omics experiment where we can integrate the mRNA and miRNA expression data with the labeled mRNA and miRNA data. Although we have analyzed each of these datasets separately with basic integration, there is much more room for improvement of our integrative analysis. This analysis can help us to further understand the role of miRNAs during HL60 differentiation. Similar to our pipeline proposed in chapter 2 of my thesis, we can cluster miRNAs and mRNAs each separately based on their expression changes through the differentiation timeline. Then we would perform a correlation analysis of the miRNA and mRNA cluster expression to find candidate miRNA clusters that could interact with specific mRNA clusters. Finally, this analysis can be complemented by target enrichment analysis of miRNA cluster targets in mRNA clusters. This analysis pipeline should lead us to miRNA clusters that significantly affect and regulate mRNA clusters through HL60 differentiation into macrophages.

Ultimately, we could perform a higher resolution analysis of miRNA targets and the dynamics of transcription if we had the isoform-level RNA expression in our differentiation time-points. We have already established a long-TUC-seq protocol that can be run on the same sample as short-TUC-seq, thus we can apply the same protocol on the samples obtained from the differentiation timeline in order to increase the resolution of our study.

References

Burger, Kaspar, et al. “4-Thiouridine Inhibits RRNA Synthesis and Causes a Nucleolar Stress Response.” *RNA Biology*, vol. 10, no. 10, 2013, pp. 1623–30, doi:10.4161/rna.26214.

Chen, Yuhao, and Xiaowei Wang. “MiRDB: An Online Database for Prediction of Functional MicroRNA Targets.” *Nucleic Acids Research*, vol. 48, no. D1, Oxford University Press, 2020, pp. D127–31, doi:10.1093/nar/gkz757.

Erhard, Florian, et al. “ScSLAM-Seq Reveals Core Features of Transcription Dynamics in Single Cells.” *Nature*, vol. 571, no. 7765, 2019, pp. 419–23, doi:10.1038/s41586-019-1369-y.

Gorkin, David U., et al. “Systematic Mapping of Chromatin State Landscapes during Mouse Development David.” *BioRxiv*, 2014.

Griffiths, Jonathan A., et al. “Using Single□cell Genomics to Understand Developmental Processes and Cell Fate Decisions.” *Molecular Systems Biology*, vol. 14, no. 4, 2018, pp. 1–12, doi:10.15252/msb.20178046.

Hafner, Markus, et al. “Transcriptome-Wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP.” *Cell*, vol. 141, no. 1, Elsevier Ltd, 2010, pp. 129–

41, doi:10.1016/j.cell.2010.03.009.

Jansen, Camden, et al. "Building Gene Regulatory Networks from ScATAC-Seq and ScRNA-Seq Using Linked Self Organizing Maps." *PLoS Computational Biology*, vol. 15, no. 11, 2019, pp. 1–22, doi:10.1371/journal.pcbi.1006555.

Kozomara, Ana, et al. "MiRBase: From MicroRNA Sequences to Function." *Nucleic Acids Research*, vol. 47, no. D1, Oxford University Press, 2019, pp. D155–62, doi:10.1093/nar/gky1141.

Lebrigand, Kevin, et al. "High Throughput, Error Corrected Nanopore Single Cell Transcriptome Sequencing." *BioRxiv*, vol. 1, 2019, p. 831495, doi:10.1101/831495.

Liu, Huanle, et al. "Accurate Detection of M6A RNA Modifications in Native RNA Sequences." *Nature Communications*, vol. 10, no. 1, 2019, pp. 1–9, doi:10.1038/s41467-019-11713-9.

Lorenz, Daniel A., et al. "Direct RNA Sequencing Enables M6A Detection in Endogenous Transcript Isoforms at Base-Specific Resolution." *Rna*, vol. 26, no. 1, 2020, pp. 19–28, doi:10.1261/rna.072785.119.

Maier, Kerstin C., et al. "Native Molecule Sequencing by Nano-ID Reveals Synthesis and Stability of RNA Isoforms." *BioRxiv*, 2019, pp. 1–32, doi:10.1101/679670.

Pajak, Author Maciej, and T. Ian Simpson. *Package ' MiRNAtap .'* 2015.

Parker, Matthew T., et al. "Nanopore Direct RNA Sequencing Maps the Complexity of Arabidopsis MRNA Processing and M6A Modification." *ELife*, vol. 9, 2020, pp. 1–35, doi:10.7554/eLife.49658.

Rosenberg, Alexander B., et al. "Single-Cell Profiling of the Developing Mouse Brain and Spinal Cord with Split-Pool Barcoding." *Science*, vol. 360, no. 6385, 2018, pp. 176–82, doi:10.1126/science.aam8999.

Sticht, Carsten, et al. "Mirwalk: An Online Resource for Prediction of Microrna Binding Sites." *PLoS ONE*, vol. 13, no. 10, 2018, pp. 1–6, doi:10.1371/journal.pone.0206239.

Workman, Rachael E., et al. "Nanopore Native RNA Sequencing of a Human Poly(A) Transcriptome." *Nature Methods*, vol. 16, no. 12, Springer US, 2019, pp. 1297–305, doi:10.1038/s41592-019-0617-2.

Wyler, Emanuel, et al. "Single-Cell RNA-Sequencing of Herpes Simplex Virus 1-Infected Cells Connects NRF2 Activation to an Antiviral Program." *Nature Communications*, vol. 10, no. 1, Springer US, 2019, pp. 1–14, doi:10.1038/s41467-019-12894-z.

Xu, Wenlong, et al. "Identifying MicroRNA Targets in Different Gene Regions." *BMC Bioinformatics*, vol. 15, no. Suppl 7, BioMed Central Ltd, 2014, p. S4, doi:10.1186/1471-2105-15-S7-S4.