DYNAMICS OF MRNA AND MICRORNA EXPRESSION IN THE ESTROGEN RESPONSE OF BREAST CANCER CELLS

Jeanette Baran-Gale

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2016

Approved by:

Praveen Sethupathy

Jeremy Purvis

Terry Furey

Jan Prins

Alain Laederach

Shawn Gomez

ABSTRACT

Jeanette Baran-Gale: Dynamics of mRNA and microRNA expression in the estrogen response of breast cancer cells
(Under the direction of Praveen Sethupathy and Jeremy Purvis)

Cellular signaling leads to broad changes in gene expression that reprogram the cell and alter cell state. Signaling often begins with cellular receptors binding a ligand and initiating a transcriptional response. One example of this is the estrogen receptor, which binds the ligand estrogen and translocates to the nucleus where it binds to estrogen response elements and regulates the expression numerous target RNAs. The regulatory network of both messenger RNAs (mRNAs) and microRNAs (miRNAs) responding to estrogen stimulation is a complex, dynamic and multilayered program that is critical to the etiology of breast cancer.

Estrogen receptor α (ERα) is an important biomarker of breast cancer severity and a common therapeutic target. Recent studies have demonstrated that in addition to its role in promoting proliferation, ERα also protects tumors against metastatic transformation. Current therapeutic strategies inhibit estrogen stimulated signaling and interfere with both beneficial and detrimental signaling pathways regulated by ERα. Additionally, ERα cyclically binds estrogen response elements and induces bursts of transcriptional activity. Together these observations suggest that ERα regulated genes and miRNAs may exhibit temporal variation in expression. Furthermore, it remains unclear if estrogen stimulated pathways exhibit the same temporal expression patterns, or if different pathways exhibit different temporal expression patterns.

By combining both RNA-sequencing and small RNA-sequencing of cells responding to estrogen, we uncover the dynamics of both mRNA and miRNA expression in response to

estrogen stimulation. Furthermore, we identify a regulatory circuit with potential therapeutic relevance to breast cancer that more specifically inhibits ERα-stimulated growth and survival pathways without interfering with its protective features. In response to estrogen stimulation, MCF7 cells (an estrogen receptor positive model cell line) exhibit induction of miR-503, and repression of the oncogene *ZNF217*. miR-503 inhibits proliferation in MCF7 cells, in part through its inhibition of the oncogene *ZNF217* and the cell-cycle gene *CCND1*. While numerous regulatory interactions can be mined from this temporal profile of estrogen responsive mRNAs and miRNAs, the induction of the anti-proliferative microRNA, miR-503, both highlights the protective aspects of estrogen signaling and indicates that miR-503 holds promise as a therapeutic for breast cancer.

ACKNOWLEDGEMENTS

My project would not have been possible without the support and guidance of members of the UNC Chapel Hill community. I would like to thank the members of both the Sethupathy and Purvis labs. I am grateful for the daily help and guidance of my mentors Praveen Sethupathy and Jeremy Purvis, without which I would not be the scientist I am today. Their support and guidance has helped me to grow as a scientist, and thanks to that guidance I am a much better writer, presenter and experimentalist today than I was at the beginning of this journey. I would also like to thank Lisa Kurtz, for her extensive help in teaching me the laboratory techniques used throughout this work. Also, I would like to thank Sam Wolff, Bailey Peck, Denise Davis and Phil Coryell for additional guidance on experimental techniques.

I also want to extend my thanks to the rest of my thesis committee: Terry Furey, Jan Prins, Alain Laederach and Shawn Gomez. Each of my committee members has provided invaluable feedback, constructive criticism, and guidance both during our official committee meetings and in one-on-one conversations. I am especially grateful to Terry Furey for (1) agreeing to be my committee chair, (2) keeping us all on track during committee meetings and (3) always having time to joke around with me. I am truly lucky to have the support of these amazing faculty members.

I am truly grateful for all the support I have received from the entire community at UNC Chapel Hill. Interactions with my fellow Bioinformatics and Computational Biology graduate students, weekly Genetics department seminars, and numerous other scientific and professional development opportunities have all contributed to my scientific development. In particular, I

would like to thank Jessica Harrell and Bailey Peck for giving me the opportunity to co-develop and co-teach a Bioinformatics module to scholars from the UNC Postbaccalaureate Research Education Program. Lastly, I would like to thank my husband for his infinite patience, unending support and unswerving confidence in my eventual success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EGF	Epidermal growth factor
NGF	Neuronal growth factor
miRNA	microRNA
mRNA	Messenger RNA
pri-miRNA	Primary microRNA transcript
pre-miRNA	microRNA precursor
HTS	High throughput sequencing
T2D	Type 2 diabetes
ER	Estrogen receptor
ERα	Estrogen receptor α
ER+/-	Estrogen receptor positive / negative
ERE	Estrogen response element
ADAR	Adenosine deaminase acting on RNA
NTA	Non-templated nucleotide addition
RIN	RNA integrity number
RPMM	Reads per million mapped RNAs
RISC	RNA induced silencing complex
3'-UTR	3' untranslated region
lncRNA	Long non-coding RNA

CHAPTER 1

Introduction

Biological processes are dynamic in that they involve molecular changes over time. Information about a biological signal is often encoded not only in the level of protein expression, but also in the temporal pattern of expression of that protein. For example, the tumor suppressor p53 exhibits different dynamical patterns of expression in response to different stresses [1]. Furthermore, the temporal dynamics of p53 expression lead to activation of different cellular fates[2]. In another example, the MAPK pathway can be stimulated by either epidermal or neuronal growth factor (EGF or NGF) in a model of neuronal differentiation. These growth factors stimulate different dynamical patterns of behavior in Erk activation, and those differences in Erk dynamics encode different cellular fates (differentiation or proliferation for NGF or EGF respectively) [3]. These complex behaviors are possible due to the underlying regulatory architecture, and the most direct way to uncover that architecture is to use time-series experiments to measure not only levels of expression but also the dynamics of targets.

1.1 Dynamic RNA expression

Gene expression profiling of RNAs by high throughput sequencing enables us to measure the expression of thousands of genes in one experiment. By conducting multiple RNA-seq experiments we can get a measure of the average temporal pattern of expression of those genes. This temporal pattern of expression can be used to make inferences about the underlying regulatory architecture. Different network architectures can result in distinct temporal patterns of gene expression from simple activation or repression to oscillations, pulses and even

multi-stability[4,5]. For example, an incoherent feed-forward loop, in which two signaling pathways controlled by the same upstream factor exert opposite stimuli on a downstream target, can lead to a transient pulse in gene expression [4].

Regulatory networks can be controlled at both the transcriptional and post-transcriptional level. One mechanism of post-transcriptional regulation is by microRNA (miRNA) mediated translational inhibition or mRNA degradation[6]. miRNA mediated regulatory networks frequently take the form of incoherent feed forward loops that can fine-tune the level of expression of a target protein, or coherent feed forward loops that often act to repress “leaky” transcription [5]. Accordingly, the combined temporal profile of both mRNA and miRNA expression can greatly enhance our understanding of the regulatory architecture of biological signaling networks.

1.2 Post-transcriptional repression by microRNAs

1.2.1 microRNA introduction

miRNAs are short (~ 22 nucleotide) non-coding RNAs that act as post-transcriptional regulators of target RNAs (primarily mRNAs) [7]. The biogenesis of miRNAs begins with transcription of the primary transcripts (pri-miRNAs) by RNA pol II in the nucleus [8]. Next, the RNase III enzyme Drosha in conjunction with its partner DGCR8 cleaves the pri-miRNA[9]. These enzymes form the Microprocessor complex, which binds to the base of a stem-loop structure in the pri-miRNA and cleaves the transcript into a hairpin shaped miRNA precursor (pre-miRNA)[9]. The pre-miRNA is transported out of the nucleus with the aid of exportin 5 [9]. In the cytosol, pre-miRNAs undergo a second cleavage, by the RNase III enzyme Dicer[9]. This final cleavage produces a miRNA duplex containing a guide and passenger strand. At this point the duplex is loaded into the RNA Induced Silencing Complex (RISC), and either the 5’-arm or the 3’-arm of the duplex can be selected to guide RISC to target RNAs[9]. Selection of the 5’- or 3’-arm is determined by the stability of the RNA duplex at the cleavage sites, where the arm having relatively less stability at its 5’-end is typically selected as the guide strand[9]. Finally the

mature miRNA is used as a guide to tether RISC to target sites that exist primarily within the 3'-untranslated region of mRNAs, leading to repression of the target RNA [8]. Target sites are determined by reverse complementary binding of the "seed region" (nucleotides 2-8) of the miRNA to the target RNA. miRNA regulation of target mRNAs can lead to either the deadenylation and consequent degradation of the target mRNA or the inhibition of translation of the target mRNA[6]. In either case, miRNAs act as repressors of their targets. Emerging research has uncovered that a subset of miRNAs have an alternate function within the nucleus as global activators of gene expression[10]. Activation of gene expression by miRNAs remains a nascent field of study and is not considered further here.

1.2.2 Considerations for small RNA-sequencing

Quantification of miRNA expression by high throughput sequencing (HTS) is subject to several issues / challenges. miRNA species are (1) short, (2) subject to non-templated modifications that can occur anywhere within the miRNA sequence but are most prominent at the 3'-end [11], and (3) subject to significant bias in HTS [12-15]. Detailed quantification of sequenced miRNA species (including any variants, termed isomiRs) requires the use of a custom alignment pipeline; as such I developed the miRquant pipeline (discussed in Chapter 2). This pipeline was prototyped by quantifying the miRNAs and isomiRs expressed in common between human beta cells and a mouse insulinoma cell line (MIN6). Additionally, miRquant has been used to successfully quantify miRNA and isomiR expression in a wide variety of cell and tissue types including MCF-7 cells (discussed in Chapter 5), liver tissue from mice fed a high-fat or low-fat diet [16], colon tissue from patients with Crohn's disease [17], and liver tissue from patients with hepatitis B and C [18].

Significant bias has been observed in small RNA sequencing libraries depending on the library preparation kit used to prepare the samples [19]. One of the primary determinants of the observed bias is the variable efficiency of adapter ligation[12,14,15]. Many types of HTS

libraries require the ligation of fixed adapter sequences to the target RNA/DNA fragment, and this step leads to the preferential binding of certain fragments to the adapter sequences[14]. This issue is exacerbated in small RNA sequencing due to the fact miRNA species contain a single sequenced fragment that encompasses the whole of the sequenced target rather than randomly distributed fragments within a larger sequence (as in mRNA-seq). This results in individual miRNA species having less diversity at the adapter ligation boundaries than the pool of fragments derived from the larger transcripts that are sequenced in mRNA-seq. The lack of sequence diversity for a given miRNA species combined with the severity of the adapter ligation bias issue leads to certain miRNA species evading capture in most small RNA libraries[19]. Recent advances in small RNA library preparation protocols have led to the introduction of a protocol that utilizes adapters with degenerate bases at the ligation boundaries. Evaluation of this protocol (discussed in Chapter 3) shows that miRNA expression from libraries sequenced using this kit produce normalized read counts that most closely match results obtained by RT-qPCR. Additionally, using this kit allows us to capture miRNA species that we were previously unable to observe using the standard Illumina library preparation protocols that use fixed adapter sequences. This was particularly critical for our study because the miRNA we found to be a key regulator of the cellular response to estrogen, miR-503, is not efficiently captured by standard library preparation protocols that use fixed adapter sequences.

Finally, miRNAs and their isomiRs can repress a large number of mRNA targets [6], but our studies have shown us that there exist certain miRNA or isomiRs that target a list or network of genes more than expected by chance [20,21]. We term these miRNAs “master regulators” of the target gene list / network. Identification of miRNA “master regulators” can be performed using a custom target site enrichment algorithm called miRhub (discussed in Chapter 4). To prototype the miRhub algorithm we examined genes subject to single nucleotide polymorphisms (SNPs) in type 2 diabetes (T2D) patients, which allowed us to uncover an isomiR of miR-375 that acts as potential master regulator of T2D. Additionally, miRhub has been used successfully

in numerous studies in the lab including the identification of miR-29 as a candidate master regulator of *Foxa2* target genes in the mouse liver [16], the identification of miR-31-5p and miR-203 as candidate master regulators of genes differentially expressed in the colon of patients with Crohn's disease compared to those with non inflammatory bowel diseases [17], and the identification of miR-503 as a candidate master regulator of estrogen responsive genes in MCF-7 cells (discussed in Chapter 5).

1.3 Breast cancer and the estrogen response

Breast cancer is one of the most prevalent causes of cancer-related death in women world-wide[22], and is categorized into at least five molecular subtypes[23]. In a clinical setting, these subtypes are defined primarily by a marker for proliferation (Ki67) and the expression of three cellular receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2)[24]. The five most commonly studied molecular subtypes, which are luminal A (LA), luminal B (LB), basal-like (BL), HER2-enriched (HER2), and normal-like (NL), are clinically relevant in terms of treatment response and are used to inform new therapies for breast cancer[25]. Both luminal subtypes express the ER, only LA expresses the PR, and the remaining subtypes display little to no expression of either hormone receptor[25]. Current recommendations for breast cancer treatment are based on an approximation of the molecular subtype using immunohistochemistry on tumor biopsies. In essence, tumors expressing hormone receptors are treated with hormone therapies, those overexpressing HER2 are treated with anti-HER2 therapies, and those lacking either are treated with chemotherapy alone[24]. Subtypes expressing a combination of these features receive a combined therapy; for example, ER+/HER2+ tumors (clinically classified as luminal B/HER2+) would receive all three therapies [24].

By far the most prevalent forms of breast cancer are those that stain positive for estrogen receptor-α (ERα)[24]. The estrogen receptor, in particular ERα (encoded by the *ESR1*

gene), has been widely studied in breast cancer [26-28]. ERα binds to estrogen (usually estradiol or E2), dimerizes, and translocates to the nucleus where it recruits co-activators or co-repressors to estrogen response elements (EREs)[28]. Although the estrogen receptor can form either ERα and ERβ homodimers, or ERα/ERβ heterodimers, ERα is thought to be the primary receptor involved in the estrogen response of both normal and breast cancer cells [29] and therefore is the focus of our study. ERα is extensively regulated at the transcriptional, post-transcriptional and post-translational levels, and its activity is pivotal to both the promotion of proliferation and prevention of transformation [30].

1.3.1 miRNAs in breast cancer

In breast cancer, several miRNAs have been identified that target and degrade *ESR1*[28], and numerous miRNAs have been identified as potential biomarkers or therapeutic targets [31]. At least five miRNAs (miR-22, miR-222, miR-221, miR-18a and miR-206) are both repressors of the *ESR1* mRNA and regulated by ERα [28]. Post-transcriptional regulation of ERα is of significant interest in breast cancer because although ER(-) cancers lack the estrogen receptor, many ER(-) tumors still express the *ESR1* mRNA without accumulating mature ERα protein [32]. One example of the importance of miRNA-mediated regulation in estrogen signaling is miR-18a, a miRNA that has been shown to target the *ESR1* gene, and is involved in a feedback loop in which ligand-bound ERα induces expression of the miR-17-92 cluster (of which miR-18a is a member) through activation of c-MYC. Post-maturation, miR-18a represses the translation of *ESR1* mRNA in breast cancer cells [33]. Additionally, studies have shown that this miRNA is more highly expressed in ER(-) breast cancer, leading to the hypothesis that altered miR-18a expression may play a role in the loss of ER signaling in breast cancer [28].

1.3.2 Estrogen responsive coding RNAs

Estrogen regulates the expression of a large number of mRNAs with different functions [30,34]. Firstly, estrogen stimulates both proliferation [28] and survival [35] of breast cancer cells. One mechanism by which ERα stimulates cell proliferation is through the estrogen-dependent activation of MAPK [36], while estrogen stimulated survival is mediated in part by the induction of the Inhibitor of Apoptosis family member, cIAP2 [35]. Additionally, ERα promotes an epithelial phenotype through its interaction with the transcription factors GATA3 and FOXA1, both of which are important transcriptional regulators [30,37]. Finally, ERα opposes the epithelial to mesenchymal transition (EMT) through down regulation of the transcription factors Snail 1 and Snail 2[38]. ER+ tumors are treated with anti-estrogen therapies that inhibit all estrogen-mediated signaling. Thus while such treatment may reduce tumor cell mass, it may also remove some of the estrogen-mediated safeguards against EMT.

1.4 Dynamic expression of coding and non-coding RNAs in breast cancer

Recent studies of the dynamics of ERα-mediated transcription in MCF-7 breast cancer cells have demonstrated that estrogen treatment induces transcriptional bursts by linking ERα transcriptional activity to proteasome-mediated degradation of ERα[39]. However, few studies have investigated how estrogen-stimulated regulatory networks change over time [34,40]. Additionally, although miRNAs have been profiled at one time point after estrogen stimulation [41] and one study investigated the dynamics of both mRNAs and miRNAs in response to estrogen [40], there are no sequencing based studies in which matched mRNA and miRNA expression has been measured at multiple time points in response to estrogen stimulation. The study of regulatory networks is greatly enhanced by the inclusion of temporal data, as it expands static interaction diagrams into dynamic models that can uncover complex behaviors,

such as the generation of expression thresholds[42], or the existence of stable points that allow the cell to maintain expression in the absence of continued stimulation [43].

The response of ER+ breast cancer cells to estrogen stimulation is well studied, but the matched temporal profile of both mRNAs and miRNAs has not yet been established. This system is an excellent model for investigating the regulatory architecture of cells responding to cellular signals (in this case the hormone estrogen). Utilizing the tools and techniques that I developed and which I have described in Chapters 2-4, I sought to identify both mRNA and miRNA species that are both regulated by estrogen stimulation in breast cancer and have a potential impact on the regulatory networks critical to the etiology of breast cancer (Chapter 5).

CHAPTER 2: QUANTIFICATION OF MIRNAS AND THEIR ISOMIRS FROM HIGH THROUGHPUT SEQUENCING[1]

2.1 Overview

microRNAs (miRNAs) are small non-coding RNAs that bind to and regulate the stability and/or translation of RNA molecules. Next-generation deep sequencing of small RNAs has unveiled the complexity of the miRNA transcriptome, which is in large part due to the diversity of miRNA sequence variants ("isomiRs"). The miRNA 5'-end sequence, referred to as the "seed" region, is a critical determinant of miRNA targeting and function. As such, changes at the 5'-end of a miRNA, including shifted start positions, can redirect targeting to a dramatically different set of RNAs and alter biological function. miRNA 5'-end diversity remains uncharacterized in beta cells. Therefore, we performed deep sequencing of small RNA from mouse insulinoma (MIN6) cells (widely used as a surrogate for the study of pancreatic beta cells) and developed a bioinformatic analysis pipeline (miRquant) to profile isomiR diversity. Additionally, we applied the pipeline to recently published small RNA-seq data from primary human beta cells and whole islets and compared the miRNA profiles with that of MIN6. These analyses led to the following three key observations:

(1) The miRNA expression profile in MIN6 cells is highly correlated with those of primary human beta cell and whole islet samples.

[1] Portions of this chapter were previously published as an article in the journal PLoS ONE. The original citation is as follows: Baran-Gale J, Fannin EE, Kurtz CL, Sethupathy P. Beta Cell 5′-Shifted isomiRs Are Candidate Regulatory Hubs in Type 2 Diabetes. PLoS ONE; 2013;8: e73240.

(2) miRNA loci can be classified as either: (a) homogenous – which generate isomiRs with a single dominant 5'-start, or (b) heterogeneous – which generate multiple highly expressed isomiRs with different 5'-start positions (5'-isomiRs);

(3) IsomiRs with shifted 5'-start positions (5'-shifted isomiRs) are highly expressed in MIN6, primary human beta cells and whole islets, and can be as abundant as their unshifted counterparts (5'-reference miRNAs).

2.2 Introduction

miRNAs are short regulatory RNAs that are processed from variable length primary transcripts through consecutive ribonuclease-mediated cleavage events [44,45]. miRNAs guide and tether the RNA induced silencing complex (RISC) to specific RNAs in order to regulate their stability and/or translation [6]. Numerous studies have identified miRNAs as important modulators of a wide variety of biological pathways[46,47]; for example, miR-375-mediated gene regulation is critical for both beta cell development and function. [48,49].

Similar to protein coding genes, miRNAs are present in multiple isoforms, called isomiRs [50-52]. IsomiRs are sequence variants, generated from a single miRNA locus, that consist of one or both of two types of variations: templated and non-templated [51,53,54](Figure 2.1). Templated variants match the genomic sequence, but have differing 5'-start and/or 3'-end positions, likely due to processing heterogeneity by Drosha/Dicer [44,45] and/or exonuclease-mediated nucleotide trimming [55,56]. Non-templated isomiRs are diverged from the genomic sequence due to post-transcriptional enzymatic processes that add, remove, or edit specific nucleotides[52]. Nucleotide additions are catalyzed by a class of enzymes called ribonucleotidyl transferases, which modify miRNAs by covalent addition of nucleotides to the 3'-end [57]. The most prevalent form of RNA editing is the adenosine-to-inosine edit, which is mediated by the double-stranded RNA adenosine deaminase (ADAR) family of enzymes [58].

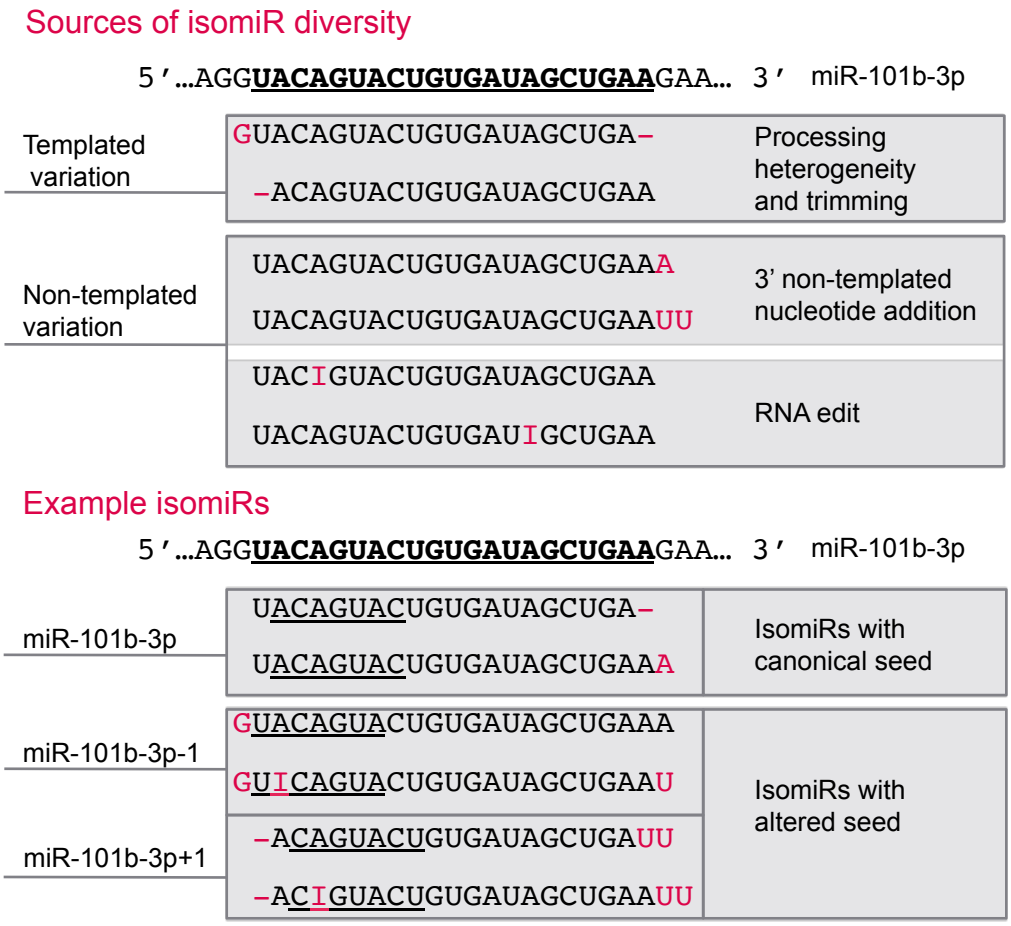


Figure 2.1: Sources of isomiR diversity. The top panel illustrates sources of isomiR diversity, which stratify into two classes: templated and non-templated variations. As illustrated by the bottom panel, an isomiR may contain one or both types of variations, and both templated (e.g. 5′-shifts) and non-templated (e.g. RNA edits) variations can create an isomiR with an altered "seed". The "seed" region of each isomiR is underlined.

IsomiRs were initially dismissed as byproducts of technical (e.g. sequencing errors) or biological noise [59,60]. However, recent studies have shown that isomiRs interact with the RISC and are present in polysomes [61-64], suggesting that they may be biologically relevant. Several studies have demonstrated that 3'-non-templated nucleotide additions (3'-NTAs), most commonly uridylation or adenylation [11,61,64-68], affect miRNA stability and/or loading onto the RISC [52,64,65] and are physiologically regulated [57,69]. Also, several studies have identified isomiRs generated by RNA edits at the 5'-end of the miRNA [70-73], referred to as the "seed" region, which is a critical determinant of stable miRNA targeting [6]. Modifications to the

"seed" region have the potential to redirect a miRNA to a vastly different set of target RNAs, thereby potentially altering its biological function [70]. Perhaps the best-studied example of this phenomenon is the A-to-I editing of the miR-376 primary transcript leading to the expression of a 5'-isomiR of miR-376 with a modified "seed" [70]. The canonical version of miR-376 and its "seed"-altered isomiR were shown to have highly distinct target sets [70], highlighting the biological importance of 5'-isomiRs.

5'-isomiRs are not limited to those generated by RNA edits; they can also be produced by processing heterogeneity and/or 5'-end nucleotide trimming, which can shift the 5'-start positions of miRNAs (Figure 2.1). 5'-shifted isomiRs have been identified in a few recent studies[11,61,64,74-76]; however, they are often reported to be lowly expressed [11,61] and continue to be perceived as rare [51]. As such, they are often overlooked by deep sequencing studies, including those performed in pancreatic beta cells. Because the 5'-end of a miRNA is so critical for function, it is of substantial interest to characterize comprehensively the prevalence and physiological relevance of miRNA 5'-diversity.

To that end, we developed an in-house bioinformatic analysis pipeline called miRquant to quantify miRNAs and their isomiRs. We applied miRquant to small RNA-seq libraries prepared from mouse insulinoma cells (MIN6), which are widely used as a surrogate for pancreatic beta cells [77]. Further, we applied miRquant to published small RNA-seq libraries from primary human beta cells and whole islets. Strikingly, we found not only that the miRNA expression profile in MIN6 cells correlates very well with those of the primary human beta cells and islets ($r^2 > 0.98$), but also that 7 highly expressed 5'-shifted isomiRs in MIN6 cells are also abundant in human beta cells and whole islet.

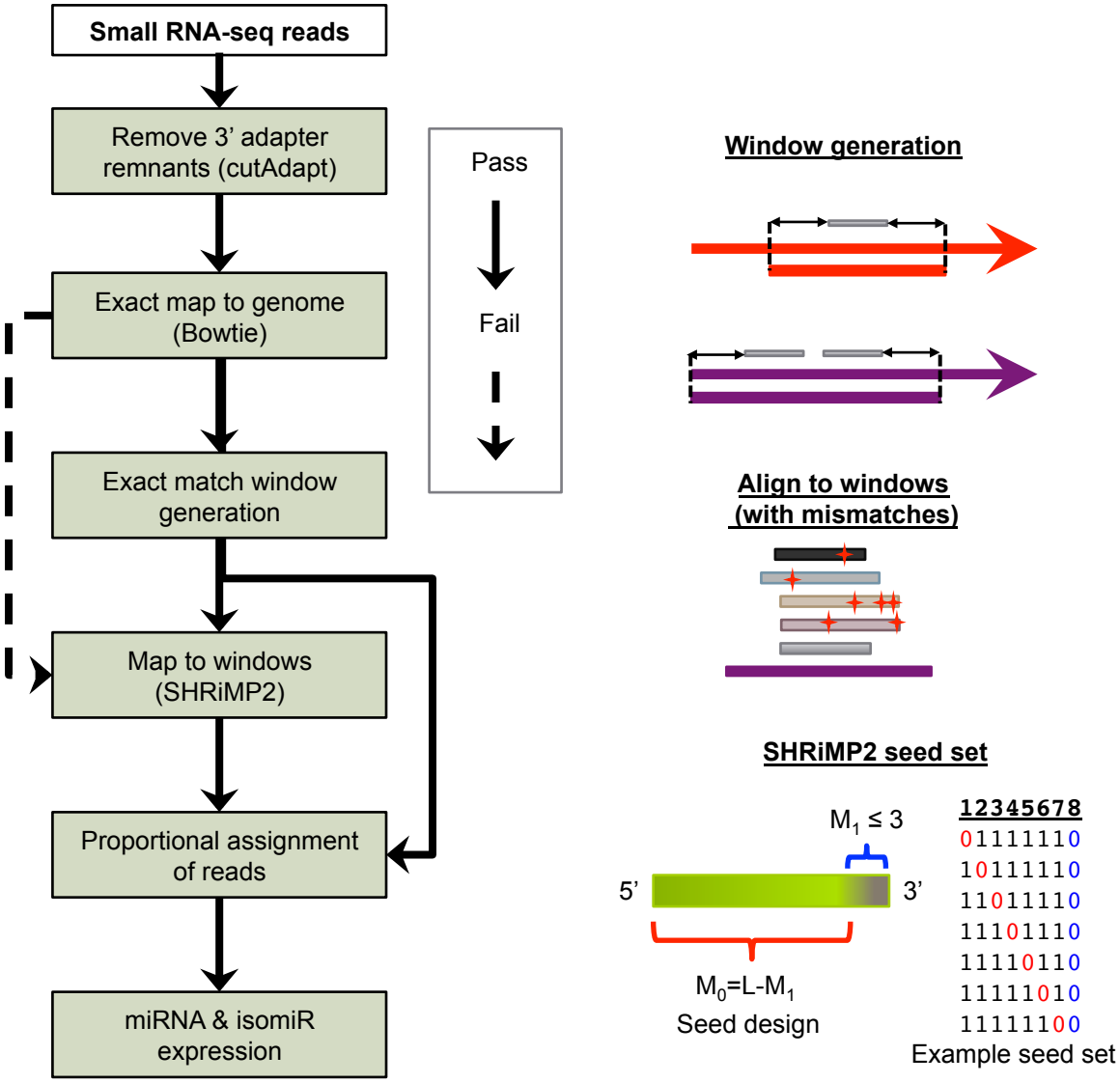


Figure 2.2: Overview of miRquant alignment pipeline. Trimmed reads are mapped in a tiered fashion to the genome. First, reads that map exactly to the genome, and these reads are used as a surrogate to define transcriptionally active regions of the genome (upper right). Next, any remaining un-mapped reads are allowed to map imperfectly only to the genomic windows generated in the previous stage, as opposed to the entire genome, thereby drastically reducing the mapable space. One mismatch is allowed in the beginning of the read and up to three mismatches are allowed at the 3'-end of the read, depending on length (lower right). An example alignment seed set is shown, where zeros mark the locations where mismatches are allowed. All reads that map equally well to multiple loci are proportionally assigned to all those loci.

2.3 Results

2.3.1 A pipeline for detailed quantification of miRNAs and their isomiRs

To capture the diversity of expressed miRNAs and their isomiRs, we developed an in-house bioinformatics pipeline (miRquant) to process and quantify small RNA-seq reads that align to both known annotated miRNAs, and potential novel miRNAs (Figure 2.2). miRNA sequences can contain several mismatches to the genome due to (1) RNA edits and (2) NTAs. To align short reads with many potential mismatches, we utilize a multi-stage alignment strategy. First reads are aligned with no mismatches allowed. Next, in a second alignment stage, alignments with mismatches are allowed to regions of the genome where other reads align perfectly. Finally, miRNAs and their isomiRs can be quantified and sequence variations can be cataloged.

<u>miRquant small RNA-sequencing alignment pipeline:</u>

Step 1: Sequencing reads are trimmed to remove 3' adapter remnants.

Small RNAs are typically shorter than the length of reads from high throughput sequencing. miRNAs have an average length of 22 nucleotides, and typical HTS reads lengths are 50 nucleotides. Therefore, roughly half of the read contains a remnant of the 3' adapter sequence. In order to accurately align the RNA fragment to the genome, the 3' adapter remnant is trimmed using cutAdapt (Preferred parameters: Overlap = 10, number of errors=1).

Step 2: Align trimmed reads to the reference genome (no mismatches).

Trimmed reads are aligned to the reference genome using the bowtie algorithm (Parameters: -q –a –m 20 –n 0 –e 70). This stage of the pipeline serves two purposes: (1) we acquire a best match for every read to the reference genome and (2) we interpret the regions surrounding these matches as transcriptionally active. The regions defined by these perfectly aligned reads are used to generate a set of genomic windows in which further alignments will be attempted.

Step 3: Generate library of windows surrounding exact matches.

To align short reads with several mismatches, we make the simplifying assumption that every expressed miRNA will have at least one read that will perfectly align to the genome (no edits, or NTAs). To capture miRNA related reads, we define regions surrounding perfectly aligned reads as a set of genomic windows. Windows separated by fewer than 65 nucleotides are merged as they may represent miRNA precursor loci. Merged windows are further extended by 5 nucleotides on either end to capture any non-templated additions. Finally these windows are converted to FASTA format.

Step 4: Align remaining reads to window library (mismatches allowed).

Our final alignment stage attempts to align any previously unaligned reads to the newly generated FASTA library of genomic windows (with mismatches allowed) using SHRiMP2. Specifically, the set of all possible "alignment seeds" containing one mismatch in the body and up to three mismatches at the 3'-end (depending on read length) is generated and used to align all reads to the genomic windows (Figure 2.2). The number of mismatches allowed at the 3'-end (M_1) for a read of length L is defined as: $M_1 = 0$ if $L < 16$, $M_1 = 1$ if $16 \leq L < 19$, $M_1 = 2$ if $19 \leq L \leq 23$, and $M_1 = 3$ if $L > 23$. Finally, alignments are scanned and only the best alignments are retained. All reads mapping equally well to multiple loci are proportionally assigned to those loci.

Step 5: Consolidate results and catalog isomiR diversity.

All aligned reads (from both stages) are grouped by 5'-start position (5'-isomiRs) and are annotated with respect to the 5'-start position of the reference (miRBase r19) miRNA at the same locus. For each 5'-isomiR, an overall expression level is reported, and all reads with mismatches at the 3'-end are reported as 3'-NTAs (by nucleotide(s) added). Alignments with internal mismatches are reported as potential RNA edits.

2.3.2 Comparing the miRNA and 5'-isomiR profiles of MIN6, human beta cell, and human islet

To characterize isomiR diversity (Figure 2.1) in mouse insulinoma cells (MIN6), we generated small RNA libraries (n=3) and performed deep sequencing on the Illumina HiSeq 2000 platform (Methods), which yielded ~18 million reads per replicate. After 3'-adaptor trimming, on average ~50% of the reads were within the expected size range (16-27 nt) for miRNAs. To analyze the small RNA-seq reads further, we developed and implemented an in-house bioinformatic analysis pipeline for highly sensitive detection and quantitation of isomiRs (miRquant; Figure 2.2). We successfully mapped ~92% of the trimmed MIN6 reads in each replicate and on average ~75% of the mapped reads corresponded to annotated miRNA loci.

We also used our pipeline to analyze published small RNA-seq datasets from primary human beta cells (n=2) and whole islet (n=1) [78]. These datasets had approximately 41, 33 and 79 million reads respectively, and in each case more than 80% of the 3'-adaptor trimmed reads were within the expected size range (16-27nt) for miRNAs. We successfully mapped ~97% of the trimmed reads and on average ~85% of the mapped reads corresponded to annotated miRNA loci.

In each of the datasets, >1,000 distinct mature miRNAs were represented by at least ten reads. However, 98% of the miRNA-related reads captured the top ~100-150 mature miRNAs depending on the dataset. We refer to these miRNAs as "highly expressed." To compare miRNA profiles across the MIN6 replicates and human samples, we first assembled a list of miRNAs that were "highly expressed" in at least one dataset, resulting in 209 distinct mature miRNAs produced from 187 unique pre-miRNAs . These 187 pre-miRNAs consisted of:

(1) 166 pre-miRNAs that generate at most one mature miRNA from each arm of the hairpin-like structure ("homogenous loci"), including one locus (pre-miR-5099) that produces only a 5'-shifted isomiR (mmu-miR-5099-2); and

(2) 21 pre-miRNAs that generate more than one mature miRNA from the same arm ("heterogeneous loci"), including one locus (pre-miR-375) that produces one 5'-reference miRNA and two 5'-shifted isomiRs.

Among the 209 mature miRNAs, 186 were 5'-reference miRNAs and 23 were 5'-shifted isomiRs. The miRNA profiles of each of the MIN6 replicates correlated extremely well with each other (r^2 ~ 0.99) and, strikingly, also with those of the human beta cells (average r^2 = 0.98) and whole islets (average r^2 = 0.97) (Figure 2.3A). As a control, we also sequenced libraries of small RNAs from the mouse liver prepared according to the same protocol and determined that as expected the correlation with the MIN6 profile was very poor (r^2 < 0.1; data not shown). The isomiRs from the heterogeneous loci were distributed across the spectrum of highly expressed miRNAs (Figure 2.3B), indicating that the heterogeneous status of a miRNA locus is not merely a function of the level of expression.

Although miRNA expression among these samples was highly correlated overall, such as in the case of miR-22-3p or miR-24-1-3p (Figure 2.4A), several miRNAs appeared to be specifically or preferentially expressed in either the MIN6 cells or human beta cells/islets (Figure 2.4A). For example, miR-143-3p and miR-204-5p were 791- and 265-fold more highly expressed, respectively, in the human beta cells than in the MIN6 cell line (Figure 2.4A). Likewise, miR-93-5p and miR-409-5p were 38- and 85-fold more highly expressed, respectively, in MIN6 cells than in human beta cells (Figure 2.4A).

2.3.3 Characterization of beta cell 5'-shifted isomiRs

Of the 23 highly expressed MIN6 5'-shifted isomiRs, only mmu-miR-5099-2 and mmu-miR-101b-3p-1 did not have a homologous miRNA in the human samples. Among the remaining 21, two were in the set of top 20 most highly expressed miRNAs in each of the MIN6 and human datasets: miR-375+1 and miR-375-1. Many of the 5'-shifted isomiRs, such as miR-375+1, miR-375-1, and miR-27b-3p-1, were expressed at similar levels in MIN6, human beta

cell, and human islet samples (Figure 2.4B). However, some 5'-shifted isomiRs were preferentially associated with either MIN6 or one of the human samples. For example, miR-192-5p+1 was 22-fold more highly expressed in human beta cells than in MIN6, miR-10a-5p+1 was 23-fold more highly expressed in human islets than in MIN6, and miR-183-5p+1 was nearly 3-fold more highly expressed in MIN6 than in human beta cells or islets (Figure 2.4B). These differences are likely in part a reflection of the disparities in cellular composition among MIN6 cells, beta cells, and whole islets. Nevertheless, the overall profile of 5'-shifted isomiRs was fairly highly correlated between MIN6 and human beta cells/islets ($r^2 \sim 0.7$).

To illustrate the potential regulatory impact of these 5'-shifted isomiRs we used TargetScan [79] to predict targets for miR-375 and its 5'-shifted isomiRs, miR-375+1 and miR-375-1. While miR-375 has 390 predicted targets conserved between human and mouse, miR-375-1 targets has more than twice that many, and strikingly, miR-375+1 has only 14 (Figure 2.4C). Only eight genes (*ELAVL4, HNF1B, NFIX, NPAS3, PAX2, SHOX2, SLC16A2,* and *TSC22D2*) have predicted conserved target sites for miR-375 and both of its 5'-shifted isomiRs.

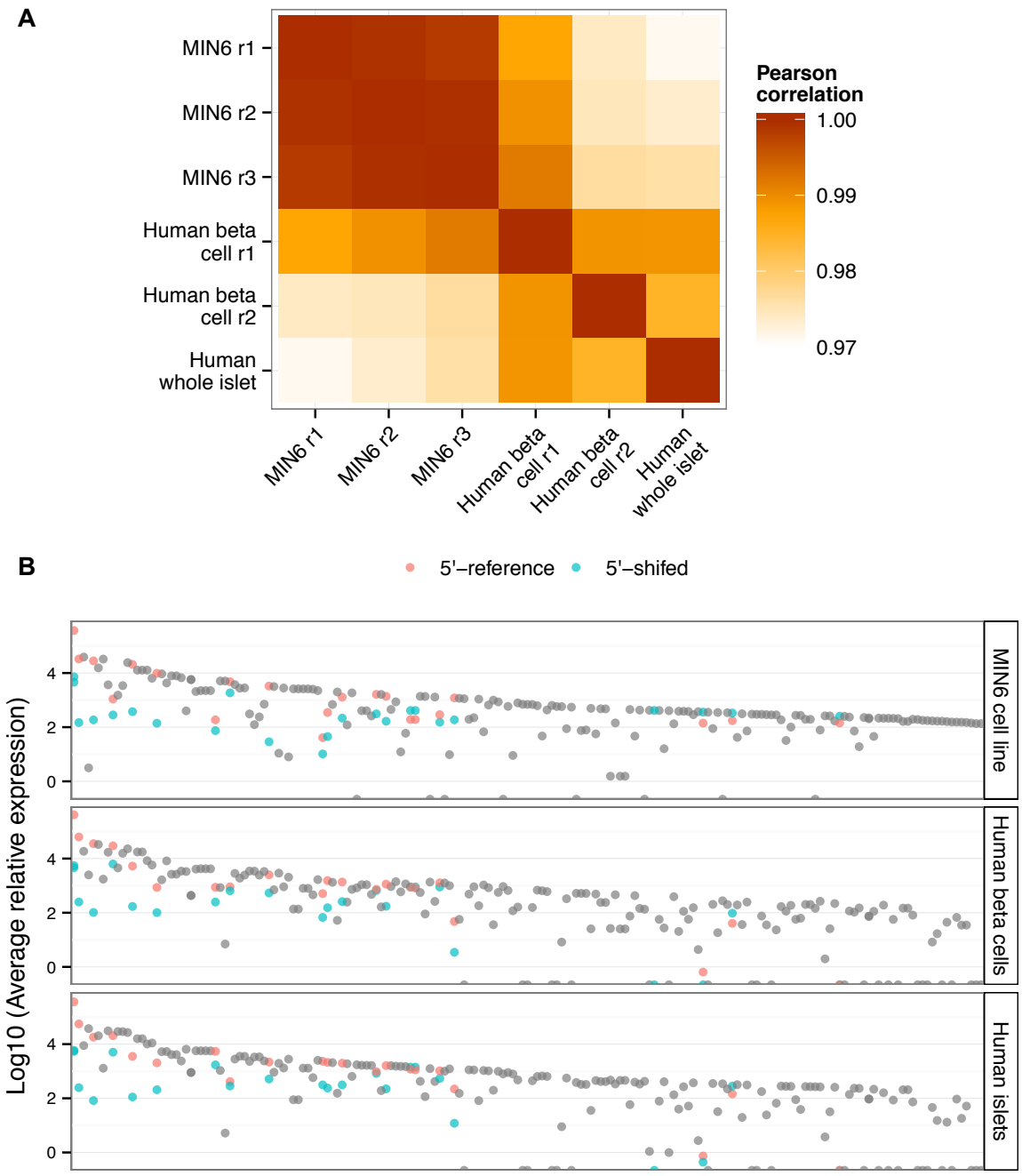


Figure 2.3: miRNA and isomiR profiles in MIN6 cells, primary human beta cells and human islet. (A) A heatmap is shown depicting the Pearson correlation coefficients of miRNA profiles between pairs of samples analyzed in this study. (B) The x-axis depicts highly expressed miRNAs ordered from left to right by decreasing maximal expression across all samples. The y-axis depicts the Log10 of the average read count per million. Each dot represents a miRNA. miRNAs from a homogenous locus (a pre-miRNA that produces only one mature miRNA per arm of the hairpin) are in gray. miRNAs from a heterogeneous locus (a pre-miRNA that produces more than one mature miRNA per arm of the hairpin) are either pink (5′-reference) or blue (5′-shifted).

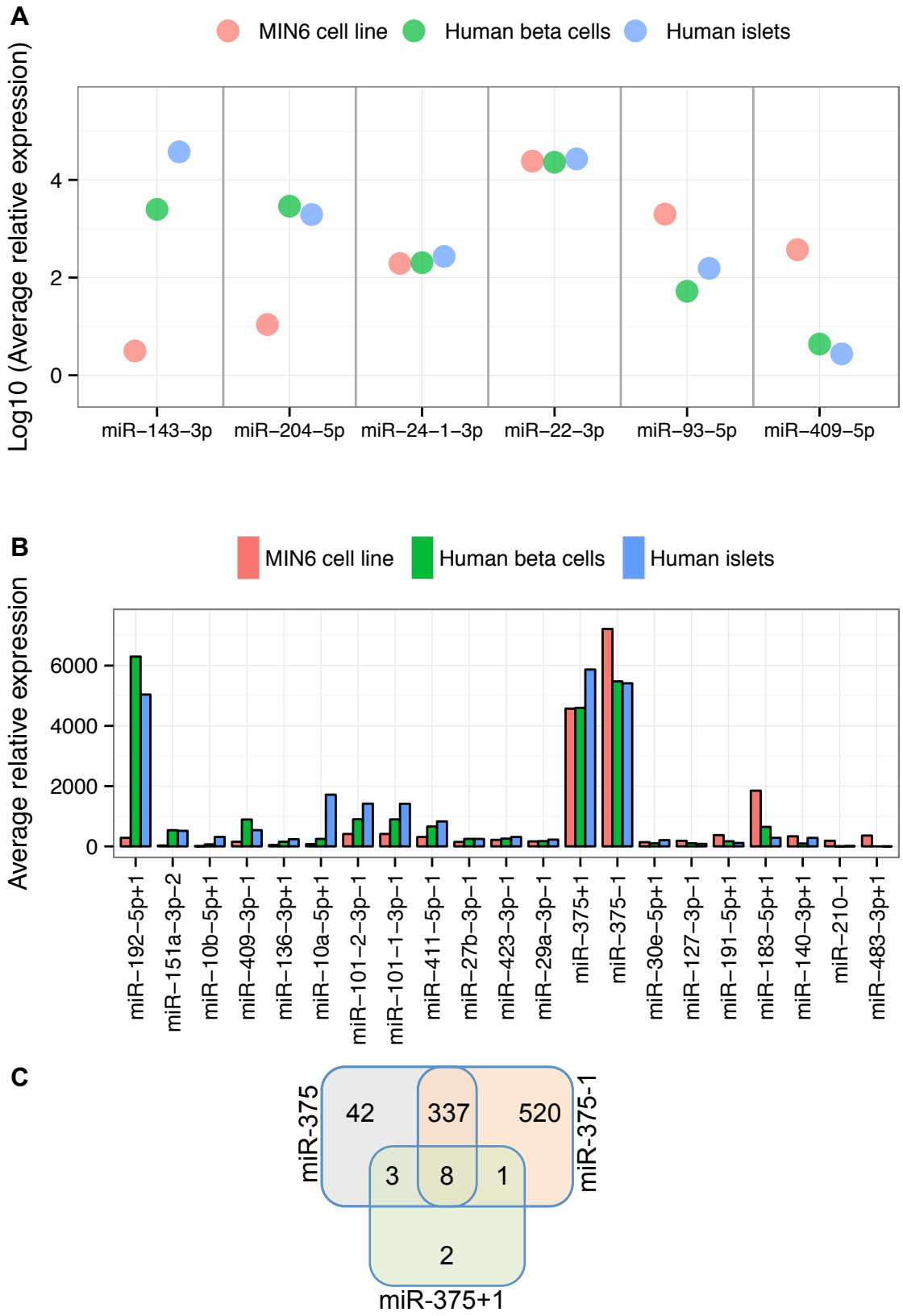

A
MIN6 cell line
Human beta cells
Human islets
Log10 (Average relative expression)
4
2
0
miR−143−3p
miR−204−5p
miR−24−1−3p
miR−22−3p
miR−93−5p
miR−409−5p
B
MIN6 cell line
Human beta cells
Human islets
Average relative expression
6000
4000
2000
0
miR−192−5p+1
miR−151a−3p−2
miR−10b−5p+1
miR−409−3p−1
miR−136−3p+1
miR−10a−5p+1
miR−101−2−3p−1
miR−101−1−3p−1
miR−411−5p−1
miR−27b−3p−1
miR−423−3p−1
miR−29a−3p−1
miR−375+1
miR−375−1
miR−30e−5p+1
miR−127−3p−1
miR−191−5p+1
miR−183−5p+1
miR−140−3p+1
miR−210−1
miR−483−3p+1
C
miR-375
miR-375-1
miR-375+1
42
337
520
3
8
1
2

Figure 2.4: Comparison of 5′-reference miRNA and 5′-shifted isomiR expression levels among MIN6 cells, human beta cells, and human islet. (A) The x-axis lists selected 5′-reference miRNAs in MIN6 (red), human beta cells (green), and human islets (blue). The y-axis depicts the Log10 of the average read count per million for each 5′-reference miRNA in each sample. (B) The x-axis shows the highly expressed 5′-shifted isomiRs ordered from left to right by decreasing fold-difference between primary human beta cells and MIN6 cells. The y-axis depicts the average read count per million for each 5′-shifted isomiR. (C) The number of genes with at least one conserved target site for miR-375 (gray), miR-375+1 (green), and miR-375-1 (orange) is shown. All sets are mutually exclusive: for example, a total of 390 genes have predicted conserved miR-375 target sites (42 unique to miR-375, 3 shared with miR-375+1 only, 337 shared with miR-375-1 only, and 8 common to all three).

2.4 Discussion

In this study we developed the miRquant pipeline to characterize isomiR diversity, and applied this method to study isomiR expression in the MIN6 cell line, primary human beta cells and islets. We found that (1) the miRNA expression profile in the MIN6 cell line is highly correlated with that of the primary human beta cells, (2) miRNA loci can be classified as either homogeneous (producing a single highly expressed 5'-isomiR) or heterogeneous (producing multiple highly expressed 5'-isomiRs), (3) 5'-shifted isomiRs can be as abundant as their 5'-reference counterparts, and (4) there are seven 5'-shifted isomiRs highly expressed in MIN6 cells that are also abundant in human beta cells and islets. Additionally, we identified 10 beta cell miRNAs, including three 5'-shifted isomiRs, as candidate regulatory hubs in type 2 diabetes. We evaluated several predicted gene targets of our top candidate regulatory hub, miR-29, and demonstrated the potential of the 5'-shifted isomiRs miR-375+1 and miR-375-1 to differentially regulate gene expression in MIN6 cells. The findings in this study promote the notion that 5'-shifted isomiRs are prevalent and potentially impact disease, thus widening the panoramic view of the functional miRNA-ome.

2.5 Materials and methods

Small RNA-seq datasets

MIN6 cells were cultured in high glucose (25mM) DMEM (Sigma) supplemented with 10% heat-inactivated fetal bovine serum. Cells were lysed and total RNA was extracted using the Norgen total RNA purification kit. RNA quality was assessed by Agilent 2100 Bioanalyzer, and only very high quality samples with RNA Integrity Number (RIN) above 9.0 were considered further. Small RNA libraries (three biological replicates) were generated using the Illumina TruSeq small RNA library preparation kit. These libraries were then sequenced on the Illumina HiSeq platform. Small RNA-seq data are available in the GEO database (submission in progress). Sequencing of small RNAs from mouse liver was conducted as well, in accordance with the protocol described above.

Human primary cell data: Primary beta cell and whole islet small RNA-seq datasets were obtained from GEO (GSE47720:[78]). This study included two libraries of beta cells (GSM1155397 and GSM1155398) and one whole islet sample (GSM1155395) that were prepared with the Illumina TruSeq protocol and sequenced on the Illumina HiSeq 2000 platform.

CHAPTER 3: ADDRESSING BIAS IN SMALL RNA LIBRARY PREPARATION FOR SEQUENCING: A NEW PROTOCOL RECOVERS MICRORNAS THAT EVADE CAPTURE BY CURRENT METHODS[2]

3.1 Overview

Recent advances in sequencing technology have helped unveil the unexpected complexity and diversity of small RNAs. A critical step in small RNA library preparation for sequencing is the ligation of adapter sequences to both the 5' and 3' ends of small RNAs. Studies have shown that adapter ligation introduces a significant but widely unappreciated bias in the results of high-throughput small RNA sequencing. We show that due to this bias the two widely used Illumina library preparation protocols produce strikingly different microRNA (miRNA) expression profiles in the same batch of cells. There are 102 highly expressed miRNAs that are >5-fold differentially detected and some miRNAs, such as miR-24-3p, are over 30-fold differentially detected. While some level of bias in library preparation is not surprising, the apparent massive differential bias between these two widely used adapter sets is not well appreciated. In an attempt to mitigate this bias, the new Bioo Scientific NEXTflex V2 protocol utilizes a pool of adapters with random nucleotides at the ligation boundary. We show that this protocol is able to detect robustly several miRNAs that evade capture by the Illumina-based methods. While these analyses do not indicate a definitive gold standard for small RNA library preparation, the results of the NEXTflex protocol do correlate best with RT-qPCR. As increasingly more laboratories seek to study small RNAs, researchers should be aware of the

[2] This chapter was previously published as an article in the journal Frontiers in Genetics. The original citation is as follows: Baran-Gale J, et al. Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. Front Gene. 2015;6: 352.

extent to which the results may differ with different protocols, and should make an informed decision about the protocol that best fits their study.

3.2 Introduction

Small RNAs, such as microRNAs (miRNAs), are important regulators of gene expression in a wide variety of normal biological and pathological processes[6,80]. Numerous technologies, including quantitative PCR (qPCR), microarray, and deep sequencing, are presently in use for high-throughput miRNA profiling[53,81,82]. Though each of these methods has both advantages and limitations, deep sequencing has emerged as the gold standard for discovery and quantification of miRNAs, particularly for those that are of low abundance. Numerous small RNA library preparation protocols are currently available, including kits from Illumina, Applied Biosystems (ABI) SOLiD, New England BioLabs (NEB), and TriLink Biotechnologies (Table 3.1). Recent studies have demonstrated that each of these technologies harbors different limitations that lead to variable biases[83,84].

Table 3.1: Current small RNA library preparation protocols and features. The protocols discussed in this study are in boldface font.

Company	Protocol	Adapters	Adapter dimer removal	RNA input recommendations
Illumina	**V1.5**	Fixed	None	1-10 µg total RNA
Illumina	**TruSeq**	Fixed	None	1 µg total RNA
Applied BioSystems	SOLiD small RNA expression kit	Degenerate hybridization	None	0.25-1 µg total RNA
Bioo Scientific	**NEXTflex V2**	Degenerate	Excess 3' adapter removal	1-10 µg total RNA
NEB	NEBNext	Fixed	Excess 3' adapter removal	0.1-1 µg total RNA
TriLink Biotechnologies	CleanTag	Fixed	Chemically modified adapters	1 ng -1 µg total RNA
SeqMatic	TailorMix miRNA V2	Fixed	Advanced gel extraction	> 10 ng total RNA

A critical step in the preparation of a small RNA library for deep sequencing is the ligation of adapter sequences to both ends of small RNAs. These adapters provide the template for primer-based reverse transcription, amplification, and sequencing. The efficiency of adapter ligation to small RNAs is thought to depend on the adapter sequence, the ligase, and the nucleotide composition and secondary structures of the small RNAs[12-15]. Differences in adapter ligation efficiency among available protocols can drastically alter the perceived abundance of individual miRNAs.

Currently, the most widely used library preparation kits are those provided by Illumina. Illumina introduced the v1.5 small RNA library preparation method in February 2009 and the TruSeq method in November 2010. Because one critical difference between these methods is the adapter sequences, some level of differential bias between these two methods is expected. However, the extent of the bias has not been evaluated previously and could be important for guiding accurate comparison of miRNA expression results between these two methods.

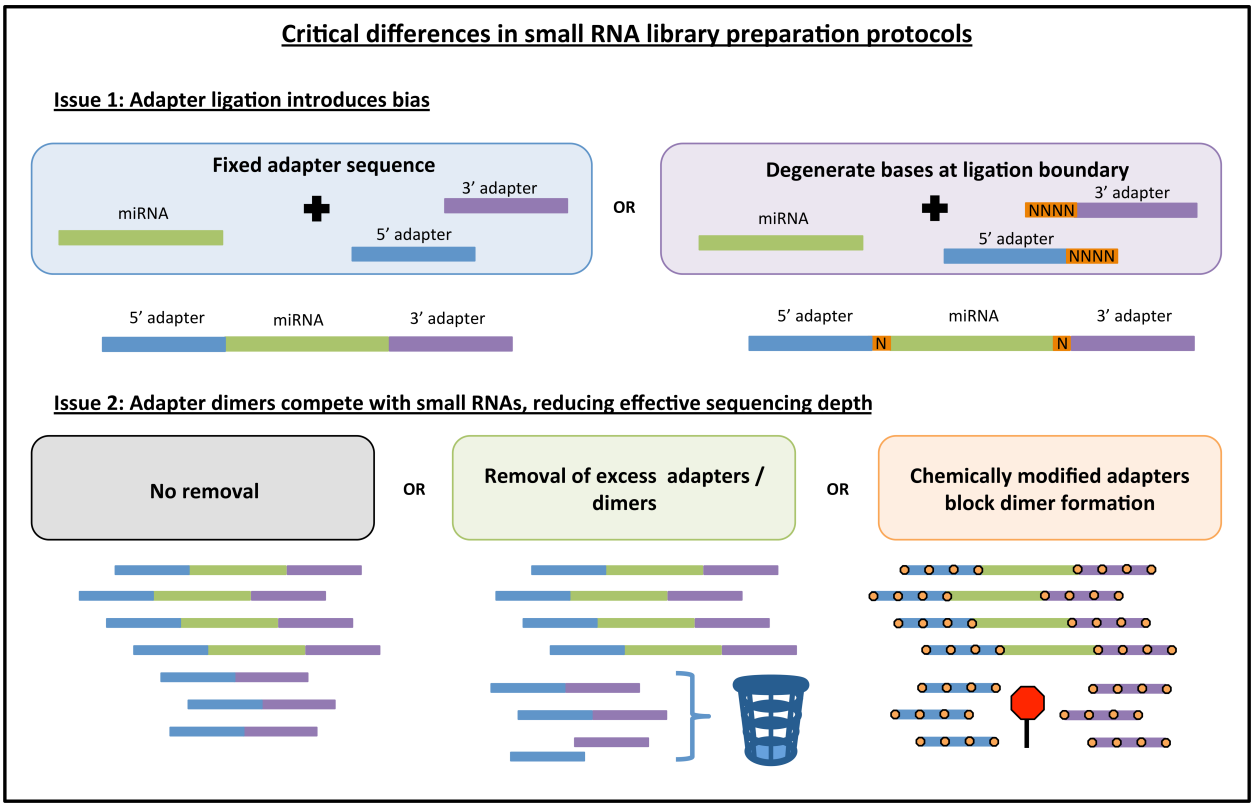


Figure 3.1: Key differences among different commercially available library preparation kits for small RNA sequencing. Some of the innovations in small RNA library preparation are highlighted here. First, current kits either used fixed adapter sequences or they introduce degenerate bases to both the 3' and 5' ligation boundary to improve adapter ligation efficiency. Second, adapter dimers can be generated causing a portion of sequenced reads to contain no insert. These dimers can be blocked or removed, thus increasing effective sequencing depth. Note: orange boxes indicating degenerate bases are not depicted in the adapter dimer graphic for the sake of simplicity.

To address this bias, a few new library preparation methods have been proposed. For example, one new protocol called NEXTflex V2 from Bioo Scientific uses a pool of adapters, each with random nucleotides (degenerate bases) at the ligation boundary (Figure 3.1). The idea behind this strategy is to increase the diversity of adapter sequences thereby increasing the chance that any given miRNA will be able to ligate efficiently, and thus mitigating the overall bias inherent to protocols that use only one set of adapters. As another example, a recent study[12] uses a 5' adapter with a short subsequence that is fully complementary to a region within the 3' adapter. The intent of this method is to encourage all ligated miRNAs to form the same circular RNA structure and thus mitigate structure-based bias across miRNAs.

Another important issue for small RNA library preparation is the formation of adapter dimers (Figure 3.1). An abundance of adapter dimers in small RNA libraries can lead to sequencing a substantial number of reads with no miRNA insert, thus effectively reducing the proportion of informative sequencing reads[85]. Currently available library preparation kits either (1) fail to address this issue or suggest more precise gel cutting to avoid the adapter dimer band, (2) use some method of eliminating excess 3' adapter prior to the 5' adapter ligation step, or (3) use chemically modified adapters that inhibit the formation of adapter dimers (Figure 3.1). Kits that address the issue of adapter dimers can typically produce high-quality results with a lower abundance of input RNA. As the field moves toward sequencing of sub-populations of cells or single cells, the adapter dimer issue will become increasingly important.

In this study, we directly compare the small RNA sequencing results between Illumina v1.5 and TruSeq. We also perform the sequencing on two different Illumina platforms (GAIIx and HiSeq) and at two different sequencing centers (UNC and NIH). While we expected some level of bias in the library preparation, the apparent extensive differential bias between these two widely used Illumina adapter sets is striking and not reported previously. For example, 50 highly expressed miRNA species are >10-fold differentially detected between v1.5 and TruSeq. This finding serves as an important caution, particularly to laboratories/facilities that used v1.5 but are now transitioning to the newer protocol. Finally, we compare these results to a library generated by a new protocol from Bioo Scientific that seeks to combat both adapter ligation bias and excessive adapter dimer formation. We show that this new protocol is able to detect miRNAs that evade capture by the more commonly used Illumina protocols, and also produces miRNA expression counts that are highly correlated with measurements acquired by RT-qPCR. The findings of this study add to the growing body of literature on bias in small RNA sequencing that merits continued investigation, particularly with regard to the development of strategies for bias remediation and improved miRNA quantification.

3.3 Results

We isolated RNA from a widely-used pancreatic beta-cell-like cell line (MIN6) and performed small RNA-seq using four different methods: (1) Illumina v1.5 library preparation sequenced on GAIIx platform (v1.5-GAIIx), (2) Illumina TruSeq library preparation sequenced on GAIIx platform (TS-GAIIx), (3) Illumina TruSeq library preparation sequenced on HiSeq platform (TS-HiSeq) and (4) Bioo Scientific NEXTflex V2 library preparation sequenced on the HiSeq platform (NF-HiSeq). TS-GAIIx and v1.5GAIIx were carried out at the NIH Intramural Sequencing Center (NISC) on June 25th, 2013; TS-HiSeq was performed at the UNC High throughout Sequencing Facility (HTSF) on June 6th, 2013; and NF-HiSeq was performed at the Genome Sequencing Facility (GSF) at Greehey Children's Cancer Research Institute (GCCRI) in University of Texas Health Science Center at San Antonio (UTHSCSA) on March 24th, 2015. Three replicate small RNA libraries were generated for each of the first three methods, and one replicate was generated for the fourth method, yielding a total of ten small RNA-seq datasets. The NEXTflex library was prepared from the same RNA that was used to prepare one of the TruSeq libraries.

We used our previously published bioinformatic pipeline[20] to analyze the small RNA-seq reads in each dataset. The total number of reads across the ten datasets range from ~17 million to ~29 million. In each of the datasets, >70% of the alignable reads map to annotated miRNAs and >1000 distinct mature miRNAs are represented by at least ten reads. Among these miRNAs, 358 have a relative expression of at least 100 reads per million mapped reads (RPMM) in at least one library. We refer to these miRNAs as "highly expressed." To compare miRNA expression profiles across datasets, we correlated the expression profiles of these abundant miRNAs across all ten datasets.

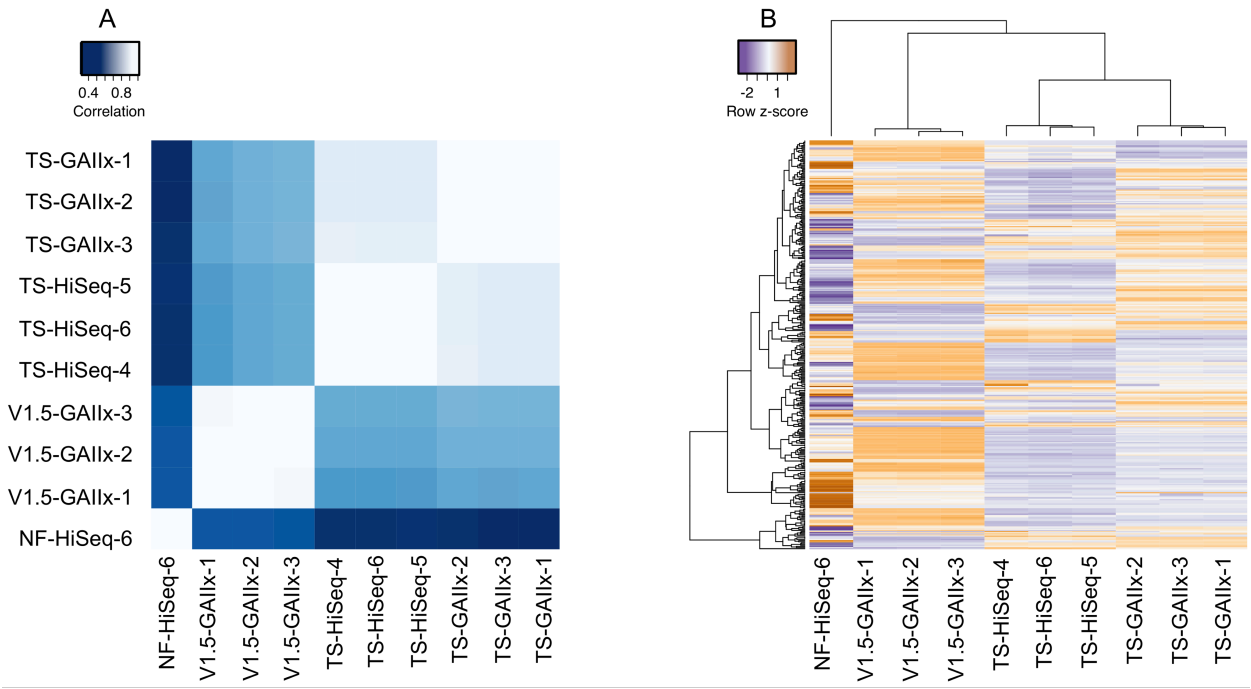


Figure 3.2: Comparison of miRNA expression profiles between two different Illumina library preparation protocols reveals massive differential bias. A comparison of the following four methods is shown: Illumina v1.5 library preparation sequenced on GAIIx platform (v1.5_GAIIx), Illumina TruSeq library preparation sequenced on GAIIx platform (TS_GAIIx), Illumina TruSeq library preparation sequenced on HiSeq platform (TS_HiSeq) and Bioo Scientific NEXTflexV2 library preparation sequenced on the HiSeq platform (NF-HiSeq). Three biological replicate small RNA libraries were generated for each of the first three methods and one replicate was generated for the NF-HiSeq method. (A) Correlation of miRNA profiles between each pair of datasets (correlation values were calculated by Pearson's metric). Similar results were obtained with Spearman's correlation coefficient, rho (data not shown). White and blue colors indicate strongest and weakest correlation, respectively. (B) miRNA expression profiles across all ten samples. Hierarchical clustering was used to identify samples with closely related expression profiles. Expression is represented as z-score, indicating the number of standard deviations below (purple) or above (orange) the mean across all ten libraries. Both (A) and (B) used only the set of miRNAs identified as "highly expressed" (n=358).

The miRNA expression profiles from biological replicates within each method are very highly correlated (average pairwise $r^2 > 0.99$), clearly demonstrating that both the method of library preparation and the sequencing platform yield exceptionally reproducible results (Figure 3.2A). Furthermore, we also observe a very strong correlation (average pairwise $r^2 > 0.86$) among TS-GAIIx and TS-HiSeq samples, but substantially lower correlation (average pairwise r^2 ~ 0.43) among TS-GAIIx (or TS-HiSeq) and v1.5-GAIIx samples (Figure 3.2A). These results indicate that neither sequencing platform (GAIIx vs. HiSeq) nor sequencing facility (UNC vs.

NIH) is a major contributor to technical variation, but that the method of library preparation (TS vs. v1.5) is a significant factor (Figure 3.2B).

Only 7 out of the 358 highly expressed miRNAs included in the correlation analysis are >10-fold differentially detected between TS-GAIIx and TS-HiSeq (Figure 3.3A, Supplementary Figure 3.1). Moreover, most of these differentially detected miRNAs are on the lower end of the expression spectrum. In stark contrast, when comparing TS-GAIIx with v1.5-GAIIx, 50 miRNAs are >10-fold differentially detected and 102 are >5-fold differentially detected (Figure 3.3B). Strikingly, ~80% (n=40/50) of the former and ~74% (n=75/102) of the latter set of miRNAs are present at greater abundance in the samples prepared by v1.5 compared to the samples prepared by TruSeq (Figure 3.3B). These miRNAs include several that are known regulators of beta cell development and function, including miR-24-3p[86], miR-29b-3p[87], and miR-200c-3p[88], which are ~36-fold, ~31-fold, and ~13-fold more highly detected in the samples prepared by v1.5, respectively. miR-24-3p is among the ten most highly expressed miRNAs in MIN6 cells according to v1.5, but is consistently not even in the top hundred according to TruSeq. It is worth noting that despite the overall bias toward higher miRNA expression levels in samples prepared by v1.5, a few miRNAs are more highly detected in samples prepared by TruSeq (Figure 3.3B). For example, miR-26a-5p, which is known to have functional relevance in the beta cell[89], is among the ten most highly expressed miRNAs in MIN6 cells according to TruSeq, but is scarcely in the top fifty according to v1.5.

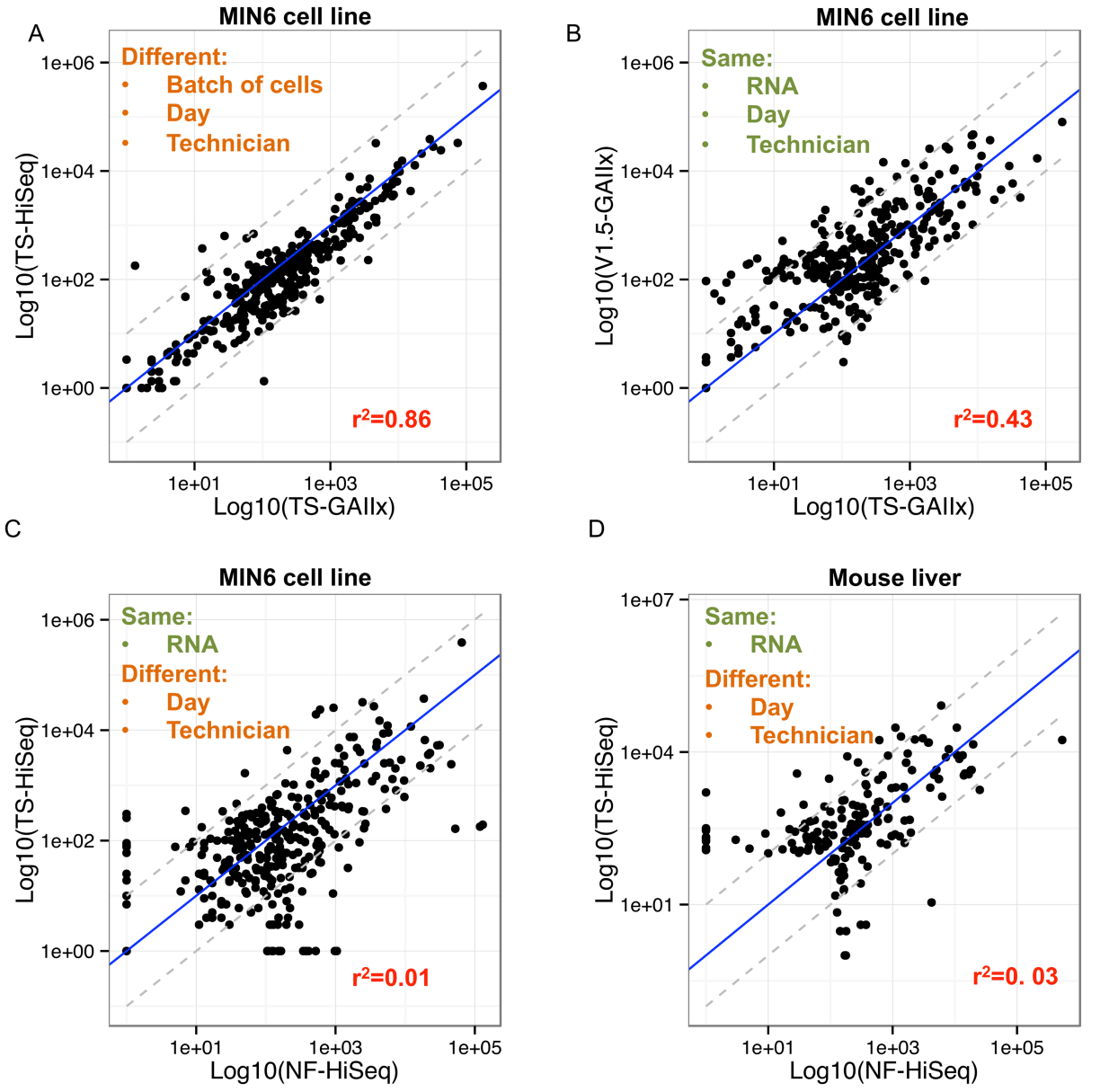


Figure 3.3: Fifty of the most abundant miRNAs are greater than ten-fold differentially detected between Illumina v1.5 and TruSeq. (A) Comparison of relative expression levels of miRNAs in MIN6 (n=358) between the GAIIx and HiSeq sequencing platforms with libraries prepared by TruSeq (TS) is shown. Each data point represents the average relative expression level for an individual miRNA across three biological replicates. (B) Comparison of relative expression levels of miRNAs in MIN6 (n=358) between the v1.5 and TruSeq (TS) library preparation methods is shown. Each data point represents the average relative expression level for an individual miRNA across three biological replicates. (C-D) Comparison of relative expression levels of miRNAs in MIN6 (n=358, (C)) and mouse liver (n=178, (D)) between the TruSeq (TS) and NEXTflex (NF) library preparation methods is shown. Each data point represents the average relative expression level for an individual miRNA across three biological replicates (A-B), or one biological replicate (C-D). Relative miRNA expression levels were calculated according to the following: $\log_{10}$(mean(miRNA RPMM)), where RPMM is reads per

million mapped reads. Pearson correlation values are displayed in red text within each panel, and grey dashed lines denote 10-fold differential expression.

We also used the new Bioo Scientific NEXTflex V2 protocol to prepare and sequence another small RNA library (NF-HiSeq-6) from the same MIN6 RNA that we had used previously for the preparation of a library by TruSeq (TS-HiSeq-6). The miRNA expression profiles produced by the two different library preparation methods are very poorly correlated (r^2 ~ 0.1; Figure 3.2A). The miRNA profile produced by NEXTflex V2 is completely different from that of Illumina v1.5 as well (Figure 3.2A-B). A total of 75 out of the 358 highly expressed miRNAs, including several with important functions in pancreatic beta cells, are >10-fold differentially detected between TS-HiSeq and NF-HiSeq (Figure 3.3C). For example, the miR-7 family of miRNAs, which regulates insulin secretion in beta cells[90], evades detection by the Illumina library preparation methods but is robustly detected by the NEXTflex V2 protocol. Strikingly, miR-7a-3p is ~670-fold more highly detected by NEXTflex V2 than by TruSeq, and ~50-fold more highly detected by NEXTflex V2 than by v1.5. Other miRNAs implicated in the control of beta cell function such as let-7b-5p[87,91,92] and miR-24-3p[86] are ~19-fold and ~15-fold more highly detected by NEXTflex V2 than by TruSeq, respectively. It is worth noting that not all miRNAs are more highly detected by the NEXTflex V2 method. For example, miR-375-3p (another miRNA critical to beta cell function [48,49]) is detected at levels ~6-fold lower by NEXTflex V2 than by TruSeq, although it is still identified as one of the most highly expressed miRNAs.

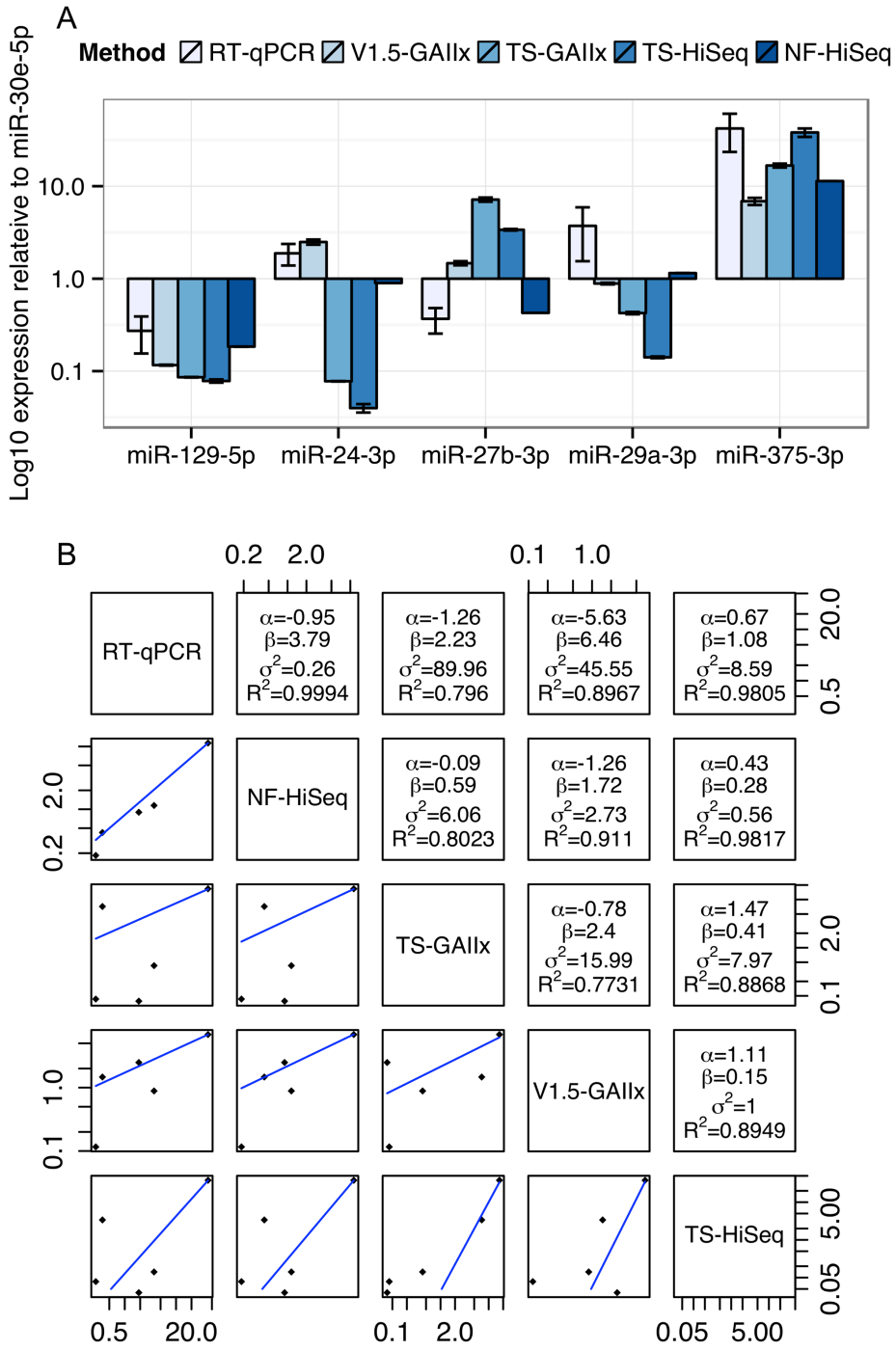


Figure 3.4: Measurements by quantitative PCR are best correlated with NEXTflex V2. (A) Comparison of relative expression levels of four miRNAs (miR-24-3p, miR-27b-3p, miR-29a-3p, and miR-375-3p) across five different methods of miRNA detection is shown. (B) Regression analysis of the relative expression of four miRNAs for each pair of detection methods is shown. The linear regression line is shown below the diagonal and the linear model parameters are

shown above the diagonal. miRNA expression levels were normalized to miR-30e-5p, which represents a housekeeping miRNA due to its invariance and robust expression across most datasets. Linear model parameters: α = intercept, β = coefficient, σ^2 = squared residual error, R^2 = fraction of variance explained by model.

To test whether the differences in miRNA expression profiles between TruSeq and NEXTflex V2 library preparation methods are unique to MIN6 or cell culture, we repeated the analysis with RNA from mouse liver tissue. Specifically, we prepared two separate small RNA libraries, using the TruSeq and NEXTflex V2 protocols, from the same mouse liver RNA and then performed sequencing on the HiSeq platform. Out of the 178 highly expressed miRNAs in the mouse liver, 40 were > 10-fold differentially detected between the TS-HiSeq-ML and NF-HiSeq-ML libraries (Figure 3.3D). Included in this list is miR-122, which has a critical role in liver biology and disease[93,94]. This miRNA is detected ~ 31-fold more highly with NEXTflex V2 protocol. In sum, ~20% of highly detected miRNAs are >10-fold differentially detected between the two protocols in both cell lines (MIN6) and primary tissue (mouse liver).

Finally, we selected five miRNAs highly expressed in MIN6 cells, miR-129-5p, miR-24-3p, miR-27b-3p, miR-29a-3p, and miR-375-3p for quantification by TaqMan-based real time reverse transcriptase quantitative PCR (RT-qPCR) (Figure 3.4A). To facilitate a comparison of the findings between RT-qPCR and the sequencing methods, we normalized the expression levels of each miRNA to that of miR-30e-5p, which is highly expressed and among the least variable across the ten small RNA-seq datasets. The sequencing method that the RT-qPCR results more closely resemble depends on the miRNA. For example, qPCR-based expression for miR-24-3p best matches that of v1.5, whereas for miR-27b-3p it best matches that of NEXTflex V2. We next generated a mathematical model to describe the linear relationship in miRNA expression between each pair of miRNA detection methods (Figure 3.4B). The residual error values (σ^2) are by far the lowest for the model relating RT-qPCR and NEXTflex V2, indicating that these two methods are the best correlated. As RT-qPCR experiments are not without their own biases, it is important to note that these data do not prove definitively that

NEXTflex V2 is the most accurate library preparation protocol. However, the data do suggest that the NEXTflex V2 protocol is indeed mitigating the adapter ligation bias inherent to the other protocols.

3.4 Discussion

The presence of bias in small RNA profiling is well established in the literature[13,95-97]. Differential bias across various expression platforms (e.g., microarray, qPCR, sequencing) and sequencing technologies (e.g., Illumina, ABI SOLiD, 454 Life Sciences) has also been demonstrated[83,84,98]. However, no study has focused on different library preparation methods within the same sequencing technology. Here we compare two of the most popular methods from Illumina (v1.5 and TruSeq). The results of our study point to a massive differential miRNA detection bias between these two library preparation methods. This finding was independent of the sequencing center (NIH, UNC) and sequencing platform (GAIIx, HiSeq). While some level of bias in library preparation is not surprising, the apparent extensive differential bias between these two widely used adapter sets is striking and not well appreciated (for example, miR-24-3p was detected very highly in the v1.5 libraries but was almost nonexistent in the TruSeq libraries).

Although we believe the extent of the bias remains poorly appreciated among many small RNA researchers, this bias has been investigated in a few studies, which together conclude that ligation efficiency is strongly affected by the co-fold structure of the target RNA and the adapter. In 2011, a study by Van Nieuwerburgh et al. demonstrated that sequencing of identical samples prepared with different barcodes at the 5' ligation boundary led to poor reproducibility, in contrast to methods in which the barcode is embedded within the adapter itself (such as TruSeq). This finding suggests that sequence diversity at the ligation boundaries could lead to variable efficiency of adapter ligation, which in turn would result in significant but artefactual effects on miRNA detection and quantification. A subsequent study by Jayaprakash

et al. provided further support for this finding, as they showed that certain miRNA species could be captured effectively only using a scheme that by introducing random bases at the ligation boundary. Specifically, this study concluded that introducing two random bases at both the 5' and 3' ligation boundaries could capture most miRNA species, but that at least one miRNA (miR-106b) required four random bases at the 5' ligation boundary in order to be captured efficiently. Other studies (Hafner et al., 2011; Sorefan et al., 2012; Zhuang et al., 2012; Fuchs et al., 2015) have investigated the contribution of RNA structure to the adapter ligation bias issue. Sorefan et al. found that the introduction of four degenerate bases to both the 3' and 5' ligation boundaries increased the diversity of structures produced by the adapter and target sequence, and thereby reduced adapter ligation bias. Zhuang et al. showed that certain RNA/adapter co-fold structures are preferred by a variety of T4 RNA ligases, but observed no sequence bias. Together these studies suggest that introducing degenerate bases to both ligation boundaries introduces both sequence and structural diversity that improves adapter ligation likely by introducing favored RNA/adapter co-fold structures.

Very recently, several new commercially available small RNA library preparation protocols have been introduced. Of these new methods, only the Bioo Scientific NEXTflex V2 protocol also addresses the important issue of adapter dimer formation. In our studies, we found that NEXTflex V2 is able to detect robustly several functionally important miRNAs that partially or completely evade detection by the widely used Illumina library preparation protocols. A prominent example of this in MIN6 cells is miR-7a-3p, which plays a critical role in beta cell function. Moreover, we show that miRNA expression levels according to NEXTflex V2 are very highly correlated with RT-qPCR measurements. While we cannot say that the results of the NEXTflex V2 method accurately represents the "absolute" expression levels of miRNAs, the results of our analysis lead us to suggest that this protocol provides the least biased measure of miRNA expression among the tested methods.

It is important to note that our study does not suggest that one method of library preparation is necessarily always more reliable or accurate for miRNA detection than the other. Because the ligation efficiencies of different adapter sequences may differ based on features that vary across miRNAs, such as nucleotide sequence, chemical modification, and secondary structure[13,84,99], care must be taken when using methods that utilize fixed adapter sequences. As the factors that control the differential biases between adapter sets continue to be investigated, we expect to see continued innovation in small RNA library preparation protocols. Researchers seeking to ameliorate the influence of adapter ligation biases on miRNA expression levels can consider using protocols that utilize degenerate bases at the adapter ligation boundaries (Table 3.1). No one protocol fits every experiment; for example, experiments with limited input RNA are better off selecting protocols optimized for such samples regardless of adapter bias considerations.

As increasingly more laboratories begin sequencing small RNAs, researchers should be aware of the extent to which the results may differ from previously published results (depending on the protocol used). We strongly caution researchers against merging together small RNA-seq data generated from different adapter sequences. Also, in any standard small RNA-seq study in which only one adapter set is used for library preparation, one should be aware of the potential pitfalls of applying arbitrary cutoffs based on expression (such as “top 100 detected”) to identify miRNAs for further functional analysis, because some miRNAs that appear lowly expressed could be inefficiently detected for purely technical reasons (such as miR-24-3p in the TruSeq datasets presented in this study). In general, we recommend against using small RNA-seq data to make calls on the “absolute” levels of miRNAs, unless additional precaution has been taken to substantially mitigate the biases discussed here. Despite these issues, deep sequencing is still an extremely valuable method for *de novo* discovery of isomiRs and novel small RNAs, as well as for studying relative miRNA expression changes across different conditions or time points.

3.5 Materials and Methods

Sequencing and bioinformatic analysis

Mouse insulinoma (MIN6) cells were cultured as previously described[20]. Cells were lysed and RNA was isolated using either the Norgen (Ontario, Canada) Total RNA Purification Kit (UNC) or TRIzol-mediated extraction (NIH). Only samples with an RNA Integrity Number (RIN) of 8 or higher, as measured by Agilent (Santa Clara, CA) Bioanalyzer 2100, were considered for further analysis. Small RNA libraries were generated using either the Illumina v1.5 protocol or the Illumina TruSeq protocol. Single-end sequencing was performed on either the Illumina GAIIx or Illumina HiSeq 2000 platforms. One library was also generated using the Bioo Scientific NEXTflex V2 protocol. For libraries generated with either of the Illumina protocols, small RNA-seq reads were trimmed using cutAdapt (-O 10 –e 0.1) to remove remnants of the 3'-adapter sequence. For the library generated with the NEXTflex protocol, the first 4 and last 4 nucleotides of small RNA-seq reads were trimmed to remove the degenerate nucleotides in the adapters. Subsequent mapping of trimmed reads to the mouse genome and miRNA/isomiR quantification were performed exactly as previously described[20].

A 9 week old C57BL/6J female mouse was purchased from Jackson Laboratories (Bar Harbor, ME) as part of a cohort of control mice for another study. This mouse was maintained on a 12 hr light/dark cycle with access to a standard chow diet and H_2O ad libitum. After a 10 day acclimation period, the control animal was weighed and injected via tail vein with RNase-free sterile saline (Bioo Scientific; Austin, TX). Seven days after dosing with saline, the animal was fasted (overnight), sacrificed by cervical dislocation without anesthesia and organs were collected. The liver was flash frozen in liquid nitrogen and stored at −80°C until RNA was extracted using the Norgen (Ontario, Canada) Total RNA Purification kit. Libraries were generated using either the Illumina TruSeq or Bioo Scientific NEXTflex V2 protocols. All animal work was performed in accordance with the Public Health Service Policy on Humane Care and

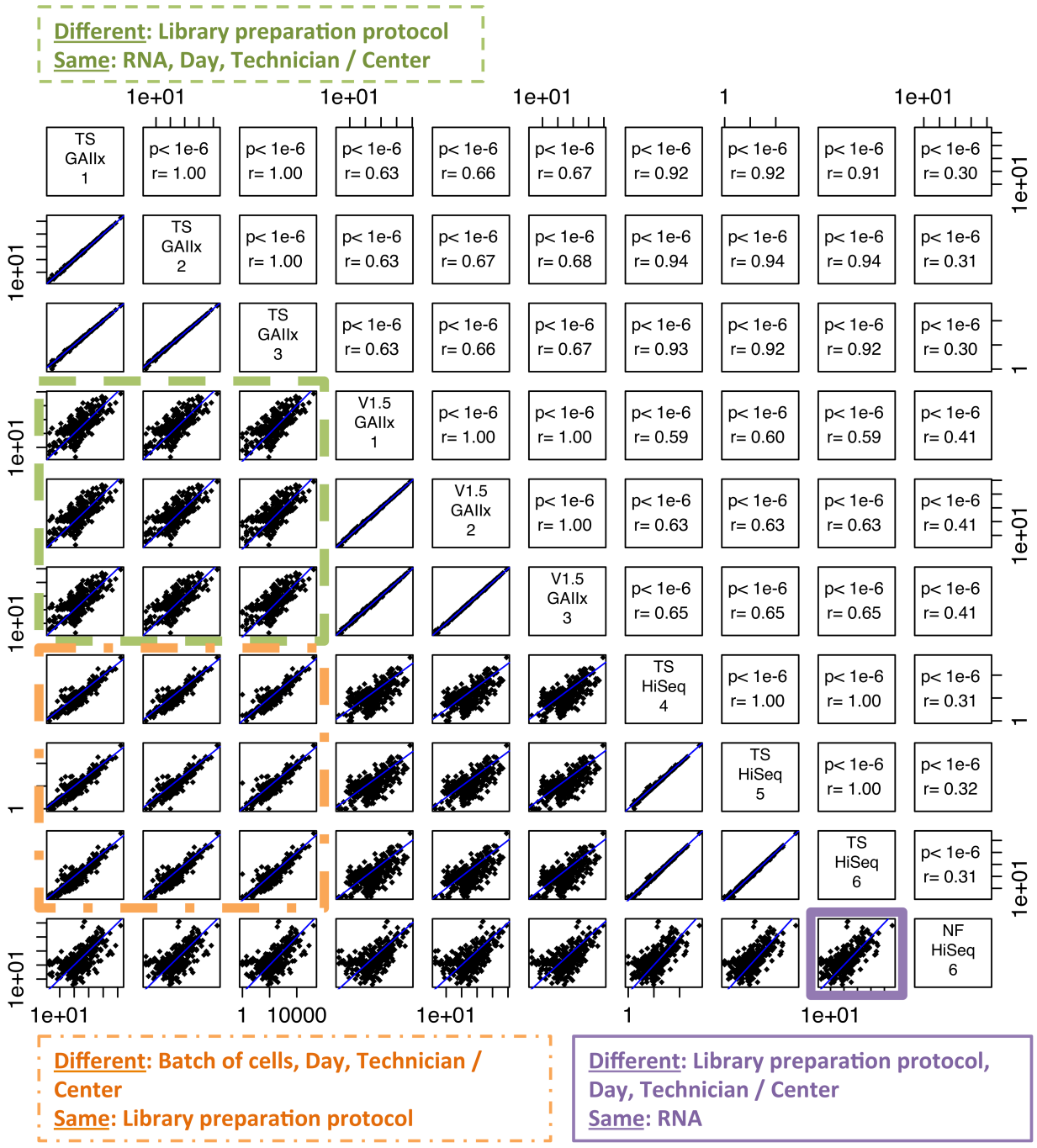
Use of Laboratory Animals, and all studies were approved by the Institutional Animal Care and Use Committee (IACUC) at the University of North Carolina at Chapel Hill.

Real time quantitative PCR analysis and linear regression

MIN6 cells were cultured and lysed as above and RNA was isolated using the Norgen Total RNA Purification Kit. Complementary DNA (cDNA) was synthesized using the TaqMan miRNA Reverse Transcription kit (Applied Biosystems; Grand Island, NY) according to the manufacturer's instructions. Real-time PCR amplification was performed using TaqMan miRNA assays in TaqMan Universal PCR Master Mix on a BioRad CFX96 Touch Real Time PCR Detection system (Bio-Rad Laboratories, Inc., Richmond, CA). Reactions were performed in triplicate using *U6* as the internal control. miRNA levels were expressed as relative quantitative values, which represent fold differences relative to miR-30e-5p (a miRNA we found to be among the least variable in expression across most library preparation protocols). All TaqMan assays used in this study where purchased from Applied Biosystems, Inc. (Grand Island, NY) and include: mmu-miR-24-3p (4427975-000402), mmu-miR-27b-3p (4427975-000409), mmu-miR-29a-3p (4427975-002112), mmu-miR-375-3p (4427975-000564), miR-30e-5p (4427975-002223), and *U6* (4427975-001973).

Linear regression was used to examine the relationship among different miRNA detection methods (RT-qPCR, Illumina V1.5, Illumina TruSeq and Bioo Scientific NEXTflex V2) in terms of the expression levels of five miRNAs (miR-129-5p, miR-24-3p, miR-27b-3p, miR-29a-3p, and miR-375-3p). For this analysis the expression level of each miRNA was normalized to that of miR-30e-5p (a miRNA we found to be among the least variable in expression across most library methods). A linear model was created in which the relative expression as measured by method Y (RE_Y) was modeled as a function of the relative expression as measured by method X (RE_X). In this model ($RE_Y = \alpha + \beta * RE_X + \varepsilon$), the term α represents the estimated expression level using method Y when the expression level is 0 using method X, β represents the weight applied to the expression as measured by method X, and ε represents the random

error in the model. To assess the model fit, two additional factors are computed and shown: R^2 (the fraction of variance that is explained by the model) and σ^2 (the estimated variance of the random error, ε).



Supplemental Figure 3.1: Pairwise differential detection of miRNAs in all 10 libraries. (Green) Pairwise comparison of relative expression levels of miRNAs between v1.5 and TruSeq libraries. (Orange) Pairwise comparison of relative expression levels of miRNAs between TruSeq libraries sequenced on either the GAIIx or HiSeq. (Purple) Pairwise comparison of relative expression levels of miRNAs between NEXTflex and TruSeq libraries. Relative miRNA expression levels were calculated according to the following: $\log_{10}$(miRNA RPMM), where RPMM is reads per million mapped reads. Pearson correlation and p-values are displayed in the upper triangle, and correlation plots on the lower triangle.

CHAPTER 4: IDENTIFYING MIRNA "MASTER REGULATORS" THROUGH TARGET SITE ENRICHMENT[3]

4.1 Overview

MicroRNAs (miRNAs) act as post-transcriptional repressors of gene expression by binding in a reverse complementary fashion to target sites typically within the 3'-UTR of mRNAs. The 5'-end of the miRNA sequence, called the "seed region" (nucleotides 2-8), plays a critical role in miRNA target recognition. Strong pairing between the "seed region" of the miRNA and target sites within the 3'-UTR of mRNAs is correlated with increased miRNA-mediated repression of a target mRNA [79]. Families of miRNAs with the same "seed" sequences share many of the same targets, and each miRNA family targets many mRNAs. To identify miRNA families that are candidates to regulate a set of gene, we developed a target site enrichment algorithm called miRhub. The miRhub algorithm identifies miRNA families that are predicted to target a gene list or network more than would be expected by chance.

Among the most highly expressed beta cell miRNAs (n=209; identified in Chapter 2), we identified 10 as candidate regulatory hubs in a type 2 diabetes (T2D) gene network. The most significant candidate hub was miR-29, which we demonstrated regulates the mRNA levels of several genes critical to beta cell function and implicated in T2D (*Slc16a1, Camk1d, Jazf1*, and *Glis3*). Further studies in the lab have confirmed that miR-29 is indeed an important regulator of pathways in diabetes in part by fine-tuning *Foxa2* activated lipid metabolism genes in the liver [16,21]. In the beta cell, three of the candidate miRNA hubs were novel 5'-shifted isomiRs: miR-

[3] Portions of this chapter were previously published as an article in the journal PLoS ONE. The original citation is as follows: Baran-Gale J, Fannin EE, Kurtz CL, Sethupathy P. Beta Cell 5′-Shifted isomiRs Are Candidate Regulatory Hubs in Type 2 Diabetes. PLoS ONE; 2013;8: e73240.

375+1, miR-375-1 and miR-183-5p+1. Although the canonical form of miR-375 is well-studied in the beta cell, the shifted forms have not been identified previously. We showed by *in silico* target prediction and *in vitro* transfection studies that both miR-375+1 and miR-375-1 are likely to target an overlapping, but distinct suite of beta cell genes compared to canonical miR-375. In summary, this study characterizes the isomiR profile in beta cells for the first time, and also highlights the potential functional relevance of 5'-shifted isomiRs to T2D.

4.2 Introduction

miRNAs bind to target sites typically within the 3'-UTR of mRNAs and tether the RNA Induced Silencing Complex (RISC) to the mRNA, leading to the translational inhibition of the mRNA or its degradation [6]. The "seed" region of the miRNA (nucleotides 2-8) is the most critical determinant of miRNA targeting potential. In fact, the level of post-transcriptional repression mediated by a miRNA on its target is correlated to the degree of complementarity between the miRNA "seed" region and the target mRNA [79].

Utilizing the same data from Chapter 2, we sought to identify highly expressed beta cell miRNAs and isomiRs that are predicted to target genes related to Type 2 Diabetes (T2D) more than is expected by chance. Accordingly, we developed a miRNA target site enrichment algorithm called miRhub, and applied it to genes implicated in T2D and related conditions. Finally, using the miRhub algorithm, we identified ten highly expressed beta cell miRNAs, including three 5'-shifted isomiRs, as significant candidate regulatory hubs in a T2D gene network.

4.3 Results

4.3.1 miRNA regulatory hubs enrichment algorithm

A miRNA acts as a "master regulator" of a gene network if it is predicted to target genes within that network more than is expected by chance. To identify miRNAs master regulators we

have developed a target enrichment algorithm titled miRhub. The miRhub algorithm computes the predicted impact for each miRNA in a set (typically those miRNAs that are highly expressed in a particular cell type) on a network of genes (typically those relevant to the study of a disease relevant in that cell type). The algorithm utilizes a Monte Carlo simulation to compare an impact score in the target gene network to scores from randomly generated networks with similar characteristics. miRNAs that are predicted to target the input gene network more strongly than the random networks are candidate master regulators of the input gene network.

First, the "seed"-based target prediction algorithm TargetScanS 5.2[79] is used to determine the number of predicted conserved targets among the human genes in the target network for each miRNA under test. Each predicted miRNA – gene interaction is assigned a score based on the strength of the "seed" match, the level of conservation of the target site, and the clustering of target sites within that gene's 3'-UTR. Additionally, the score for each gene is weighted according to the number of high-confidence protein-protein interactions reported in the STRING 9.0 database[100]. Finally, for each miRNA, the final an average targeting score is calculated for all genes in the network. In order to generate a background distribution of the predicted targeting scores for each miRNA, we repeated this procedure 30,000 times, with a new set of randomly selected human genes each time (genes and corresponding 3'-UTR sequences were downloaded from http://www.targetscan.org). These score distributions are then used to calculate an empirical p-value of the targeting score for each miRNA in the target gene set. Genes were selected at random from a pool with similar overall connectivity to the genes in the target gene set, and to account for differences in the average 3' UTR length between the genes of interest and the randomly selected genes in each simulation, the targeting score was normalized by 3' UTR length.

<u>miRhub algorithm:</u>

Inputs:

1. A list of predicted target sites for miRNA families in the multiple-sequence aligned 3'-UTRs of all genes for a set of species. This input can be acquired from TargetScan (http://www.targetscan.org).
2. A list containing the conservation (number of species) of each of the miRNA families included in the above TargetScan predictions. This list can be parsed out of the files available from the TargetScan website.
3. A list of high confidence protein-protein interactions. This list was derived from the STRING 9.0 database (http://string-db.org/) using only those interactions having an interaction score greater than 700 (high-confidence). All protein identifiers were mapped to their corresponding gene symbol. The gene symbol must match those used in the TargetScan predictions.
4. A list of miRNA families: This list contains the miRNAs the user wishes to include in the enrichment analysis. All miRNAs must be in the TargetScan miRNA family name format, and must match the family names in the Target Scan output file.
5. Gene list: A list of genes to use as central nodes in a gene network. Typically this list includes a set of genes relevant to the study of a particular disease or pathway. All gene names must match the gene symbol used in the TargetScan 3'-UTR sequence files.

Parameters:

1. ***C***: Conservation level. Requested minimum level of conservation of each miRNA and target site required for a miRNA - gene interaction to be scored in the simulation.
2. ***N***: Number of iterations. Requested number of random gene networks of similar design that are used to generate score distributions.
3. ***α***: Hub weighting. This parameter is used to weight the contribution of the number of high-confidence protein-protein interactions to the target scoring function.

A gene network is compiled using the input gene list and the connections from the protein-protein interaction database. The input gene network contains (1) all genes in the input

gene list that have a 3'-UTR listed in the target prediction files, (2) a weighted set of scores for each target site within each gene, and (3) the number of high confidence protein - protein interactions listed for that gene in the STRING 9.0 database. Each random gene network is generated by selecting a set of random genes having connectivity similar to each of those in the input gene list. A gene is said to have similar connectivity if the gene has a similar number of high confidence interactions in the STRING 9.0 database. To compute groups of genes with similar connectivity, we group each gene in the STRING 9.0 database by the number of high confidence protein-protein interactions that gene has. If any group contains fewer than 20 genes, the group is expanded to include neighboring groups (with both higher and lower number of interactions) until the new super-group contains at least 20 genes. Finally a score is computed that represents how strongly each miRNA family in the input list targets each gene network (input gene network and ***N*** random gene networks), and an empirical p-value is computed. The p-value is calculated as $\boldsymbol{p=(N_r+1)/(N+1)}$, where $\boldsymbol{N_r}$ is the number of random gene networks in which the targeting score for a particular miRNA was greater or equal to the score of that miRNA in the input gene network. The miRNA targeting score is calculated using the following procedure:

For a gene network ***G(L,D,U)***: where ***L*** is the list of genes in the network, ***D*** is the number of high confidence protein-protein interactions that each gene has, and ***U*** is the ratio of the average 3'-UTR length in the input gene network over the average 3'-UTR length in the current gene network (note: this value is one when scoring the input gene network).

miRhub pseudocode:

For each $\boldsymbol{miR_i}$ in ***miRlist*** having a conservation of at least ***C***:
    For each $\boldsymbol{gene_j}$ in ***L***:
        For each target site ***k*** of $\boldsymbol{miR_i}$ in $\boldsymbol{gene_j}$:

$$ScoreA_{ijk} = \begin{Bmatrix} 1.5 & if\ 8mer-1a \\ 1.25 & if\ 7mer-m8 \\ 1 & otherwise \end{Bmatrix}$$

$$ScoreB_{ijk} = \begin{Bmatrix} ScoreA_{ijk} & if\ site\ is\ consered\ in \geq \boldsymbol{C}\ \text{species} \\ 0 & otherwise \end{Bmatrix}$$

For each ***miR$_i$*** in ***miRlist*** having a conservation of at least ***C***:

$Score_i = 0$

For each ***gene$_j$*** in ***L***:

$ScoreC_{ij} = ScoreB_{ij0}$

For each additional target site ***k*** of ***miR$_i$*** in ***gene$_j$***:

$$ScoreC_{ij} \mathrel{+}= \begin{Bmatrix} 0.5 * ScoreB_{ijk} & if\ (Pos_k - Pos_{k-1}) \leq 8 \\ 1.5 * ScoreB_{ijk} & if\ 8 < (Pos_k - Pos_{k-1}) < 60 \\ ScoreB_{ijk} & if\ (Pos_k - Pos_{k-1}) \geq 60 \end{Bmatrix}$$

Where Pos_k is the position of target site ***k*** within the 3'-UTR of ***gene$_j$***.

$Score_i \mathrel{+}= \boldsymbol{U} * ScoreC_{ij} \left(1 + \boldsymbol{\alpha} * log10\left(\boldsymbol{D}_j\right)\right)$

$Score_i = Score_i / size(\boldsymbol{L})$

4.3.2 Candidate 5'-shifted isomiR regulatory hubs in type 2 diabetes

Genome-wide association studies for type 2 diabetes (T2D) have primarily (though not exclusively) implicated genes with critical function in the pancreatic beta cell[101,102]. Therefore, we sought to determine if any of the highly expressed human beta cell miRNAs, including 5'-shifted isomiRs (identified in Chapter 2), serve as regulatory hubs in T2D. We first assembled a list of genes (n=92) implicated in T2D and related conditions including maturing onset diabetes of the young (MODY) (Methods). We then applied the miRhub algorithm to determine for each miRNA whether the predicted regulatory impact on T2D genes is significantly (uncorrected P < 0.05) greater than expected by chance (such miRNAs are termed "candidate regulatory hubs"). We identified 10 candidate miRNA regulatory hubs (Figure 4.1A. The top two were the 5'-reference miRNAs miR-29 and let-7, both of which have been implicated in beta cell function and glucose homeostasis [87,91,92]. Though miR-29 has been shown to regulate glucose-stimulated insulin secretion, its target genes in the beta cell are largely unknown. To validate the *in silico* approach, we selected several predicted targets (*Camk1d*, *Glis3*, and *Jazf1*), and one previously validated target (*Slc16a1* [87]), of miR-29 from among the T2D gene list for evaluation in MIN6 cells. Specifically, we transiently transfected MIN6 cells with a miR-29 mimic or inhibitor (antagomiR) and measured the mRNA levels of each of the four genes by real-time quantitative PCR (RT-qPCR). Three of the four genes were

significantly ($p < 0.05$) down-regulated by the over-expression of miR-29 and three genes were significantly ($p < 0.05$) up-regulated by the antagomiR-mediated inhibition of miR-29 (Figure 4.1B). These findings are consistent with previous reports that miR-29 is involved in the regulation of beta cell function [87,103], and they serve as a validation of the *in silico* regulatory hub analysis.

Strikingly, three of the 10 candidate miRNA regulatory hubs in the T2D gene network were 5'-shifted isomiRs: miR-375+1, miR-375-1, and miR-183-5p+1 (Figure 4.1A). Moreover, all three of these were more significantly associated with T2D genes than their 5'-reference counterparts. This is particularly intriguing, given the already well-established role of 5'-reference miR-375 in beta cell formation and function.

4.3.3 5'-shifted isomiRs of the beta cell-enriched miRNA, miR-375

As depicted in Figure 2.3C, miR-375 and its 5'-isomiRs have overlapping, but distinct predicted target gene profiles. To further evaluate the putative differential targeting of the miR-375 5'-isomiRs, we selected the following three genes: *Mtpn*, which regulates insulin secretion, is a known target of the 5'-reference miR-375[48], but is not predicted to be targeted by the 5'-shifted isoforms; *Atp6v0c*, which mediates glucose-sensitive intracellular vesicular transport and is predicted to be preferentially targeted by the 5'-shifted isoform miR-375+1; and *Cdc42*, which is essential for the second phase of insulin secretion and is predicted to be preferentially targeted by the 5'-shifted isoform miR-375-1. We transfected MIN6 cells with (1) transfection reagent only (mock), (2) 10nM of miR-375 mimic, or (3) 10nM of a mimic for one of the 5'-shifted isomiRs of miR-375, and measured the mRNA levels of each of the three genes by RT-qPCR. *Mtpn* was repressed only by the 5'-reference miRNA (Figure 4.2). *Atp6v0c* and *Cdc42* were also modestly repressed by the 5'-reference miRNA, though slightly more so by miR-375+1 and miR-375-1, respectively (Figure 4.2). In each case, the strongest repression was conferred by the 5'-isomiR with the strongest predicted target site.

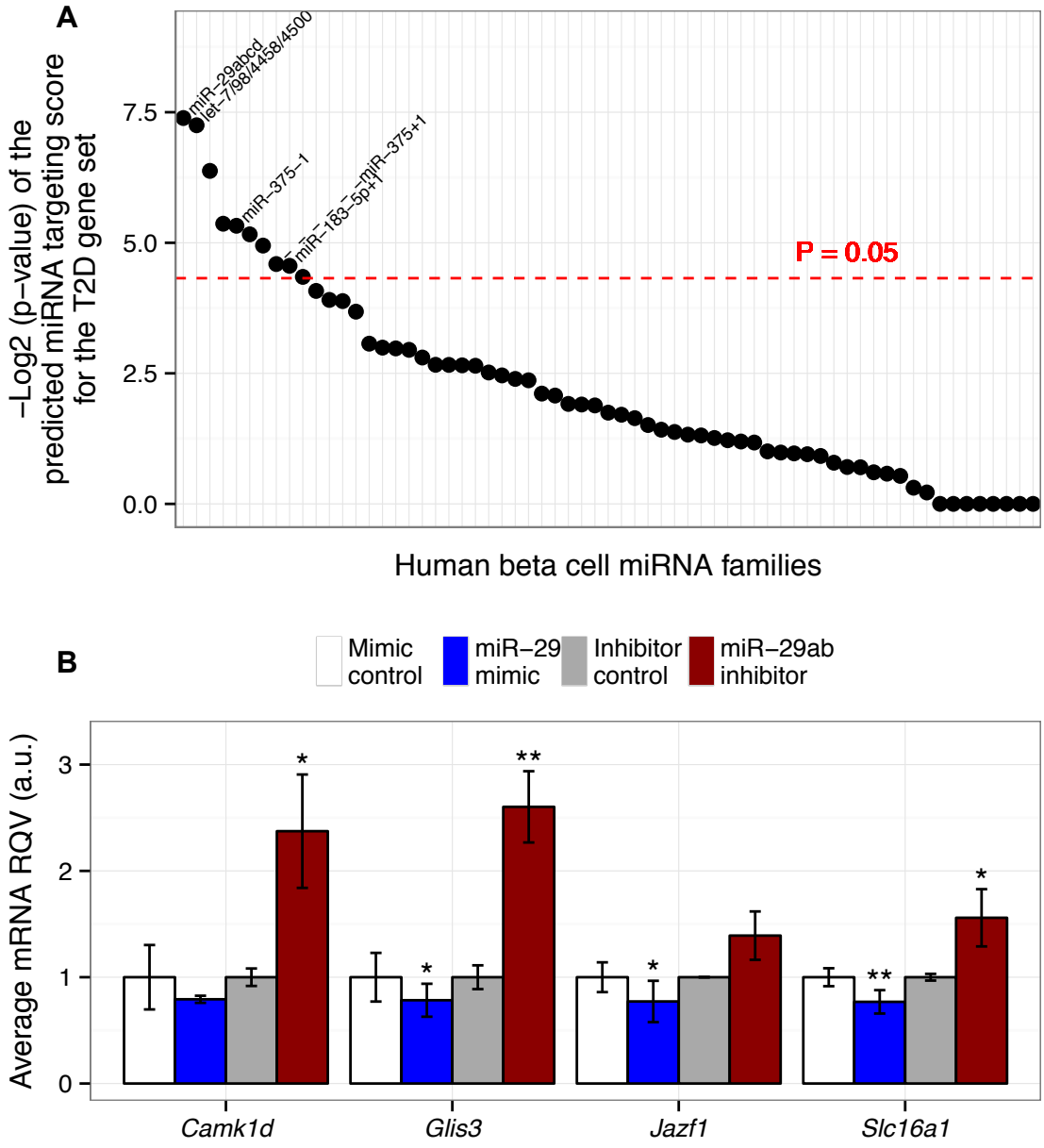


Figure 4.1: Candidate miRNA regulatory hubs in a type 2 diabetes gene network. (A) Each data point represents a 5′-reference miRNA or a 5′-shifted isomiR from primary human beta cells, and the y-axis shows the negative Log2 of the p-value of the predicted miRNA targeting score among genes in a type 2 diabetes (T2D) network. The dashed red line denotes the significance threshold (empirical P = 0.05). (B) Effects of miR-29 mimic and inhibitor in MIN6 cells on the mRNA levels of four T2D genes are shown. The x-axis lists the gene symbols for each of four predicted miR-29 target genes and the y-axis depicts the relative quantitative value (RQV; expression determined by RT-qPCR and normalized to *Rps9*) in response to the miR-29 mimic (blue) or the miR-29 inhibitor (red) relative to mock transfection. The data shown represent at least two independent experiments, each conducted in triplicate. P-values were calculated based on Student's t-tests. *, P<0.05; **, P<0.01.

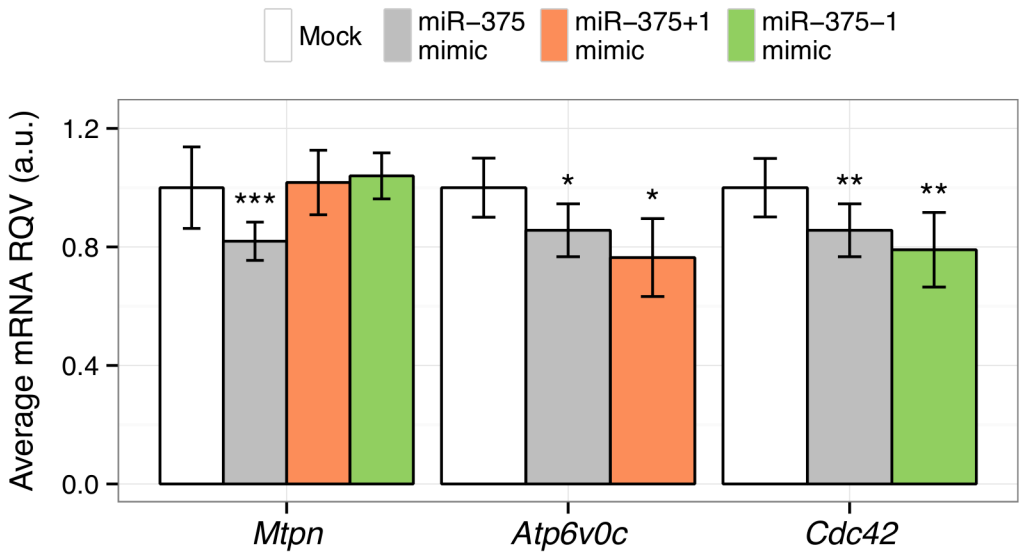


Figure 4.2: Evaluation of miR-375 and its 5′-shifted isomiRs in MIN6 cells. Effects of mimics for 5′-reference miR-375, 5′-shifted miR-375+1, and 5′shifted miR-375-1 in MIN6 cells on the mRNA levels of three genes are shown. *Mtpn* is a known target of 5′-reference miR-375 but not predicted as a target for either of the 5′-shifted miR-375 isomiRs; *Atp6v0c* is predicted to be preferentially targeted by miR-375+1; and *Cdc42* is predicted to be preferentially targeted by miR-375-1. The x-axis lists the gene symbols for each of three genes tested. The y-axis depicts the relative quantitative value (RQV; expression determined by RT-qPCR and normalized to *Rps9*) in response to the miR-375 mimic (gray), miR-375+1 mimic (orange), or miR-375-1 mimic (green) relative to mock transfection. The data shown represent at least two independent experiments, each conducted in triplicate. P-values were calculated based on Student's t-tests. *, P<0.05; **, P<0.01, ***, P<0.001.

4.4 Discussion

In this chapter we discussed an in-house miRNA target site enrichment algorithm (miRhub) that we developed to identify miRNAs and isomiRs that act as candidate master regulators of genes associated with T2D. Using this algorithm, we identified 10 beta cell miRNAs, including three 5'-shifted isomiRs, as candidate regulatory hubs in type 2 diabetes. We evaluated several predicted gene targets of our top candidate regulatory hub, miR-29, and demonstrated the potential of the 5'-shifted isomiRs miR-375+1 and miR-375-1 to differentially regulate gene expression in MIN6 cells.

While the unambiguous validation of the targeting activity of 5'-shifted isomiRs is important, it is hindered by inherent limitations of the currently available technologies. For

example, the gold standard experiment would be to specifically knock-down the 5'-shifted isomiR of interest. However, current strategies for knock-down (e.g. locked nucleic acids), and for testing the efficacy of the knock-down (e.q. TaqMan RT-qPCR), do not adequately distinguish between the 5'-reference and 5'-shifted isoforms. New approaches for studying miRNA function must be developed in order to tackle the technical challenges posed by 5'-shifted isomiRs, which are often identical in sequence to the 5'-reference form except for the addition/loss of a single nucleotide at the 5'-end. Though outside the scope of this study, further analyses are necessary to firmly establish the functional relevance of the 5'-shifted isomiRs. Nonetheless, to our knowledge, this is the first report of highly expressed 5'-shifted isomiRs in beta cells, several of which are candidate regulatory hubs in T2D.

The findings in this study promote the notion that 5'-shifted isomiRs are prevalent and potentially impact disease, thus widening the panoramic view of the functional miRNA-ome. In addition, we have identified for the first time three 5'-shifted isomiRs as significant candidate regulatory hubs in a disease network. The novel strategy employed in this study can be utilized for additional disease models to uncover potential roles for 5'-shifted isomiRs in the regulatory networks of complex diseases.

4.5 Materials and methods

Candidate miRNA regulatory hub identification in the T2D gene network

T2D gene list: We identified the nearest genes to each genetic variant significantly associated with T2D (p-values < 10^{-7}) from (1) the T2D genome-wide association studies (GWAS) listed in the NHGRI catalog (http://www.genome.gov/gwastudies) and (2) a GWAS reported by Morris et al. that was not included in the NHGRI catalog [104]. Additionally, we included twenty-three genes linked to maturity onset diabetes of the young (MODY), neonatal diabetes (NDM), and chronic hyperinsulinemia (CHI). The total number of genes was 92.

Validation of miRNA-mediated gene regulation

MIN6 cells were transiently transfected with (1) 10nM mmu-miR-29 mimic (Dharmacon); (2) 200nM mmu-miR-29 hairpin-inhibitor (Dharmacon); (3) 10nM mmu-miR-375 mimic (Dharmacon); (4) 10nM custom mmu-miR-375+1 mimic (Dharmacon: 5'-UUGUUCGUUCGGCUCGCGUGA-3') or (5) 10nM custom mmu-miR-375-1 mimic (Dharmacon: 5'UUUUGUUCGUUCGGCUCGCGUGA-3'). After 48 hours, RNA was extracted from cells using the Norgen Total RNA Purification Kit, and miRNA and mRNA levels were measured by real-time quantitative PCR (RT-qPCR) using TaqMan microRNA and gene assays (Applied Biosystems).

CHAPTER 5: AN INTEGRATIVE TRANSCRIPTOMICS APPROACH IDENTIFIES MIR-503 AS A CANDIDATE MASTER REGULATOR OF THE ESTROGEN RESPONSE

5.1 Overview

Estrogen receptor α (ERα) is an important biomarker of breast cancer severity and a common therapeutic target. In response to estrogen, ERα stimulates a transcriptional program including both coding and non-coding RNAs. However, although ERα is known to cyclically bind to estrogen response elements and initiate bursts of transcriptional activity, most studies assess only one or two time points in response to estrogen and thus fail to capture the fine-scale gene expression dynamics. With a map of this dynamic response, we can define the varying temporal patterns of expression regulated by ERα, which would greatly enhance the study of the regulatory network that underlies the estrogen response. In this chapter, I use the tools and techniques I developed and described in Chapter 2-4 to create and analyze the dynamic response of mRNAs and miRNAs to estrogen stimulation in breast cancer.

To determine how ERα signaling regulates gene expression dynamics, we performed temporal profiling of both messenger RNAs (mRNAs) and microRNAs (miRNAs) in MCF7 cells (an ER+ model cell line for breast cancer). Cells were cultured in estrogen free media, then exposed to estrogen and paired mRNA and miRNA sequencing libraries were generated at ten time points throughout the first 24 hours of the response to estrogen. We identified three primary expression trends—transient, induced, and repressed—that were each enriched for genes with distinct cellular functions. Integrative analysis of mRNA and miRNA temporal expression profiles identified miR-503 as the strongest candidate master regulator of the estrogen response, in part through suppression of *ZNF217*—an oncogene that is frequently

amplified in cancer. We confirmed experimentally that miR-503 directly targets *ZNF217* and that over-expression of miR-503 suppresses breast cancer cell proliferation. Overall, these data indicate that miR-503 acts as a potent estrogen-induced candidate tumor suppressor miRNA that opposes cellular proliferation and has promise as a novel therapeutic for breast cancer. More generally, our work provides a systems-level framework for identifying functional interactions that shape the temporal dynamics of gene expression.

5.2 Introduction

Breast cancer remains a prevalent cause of cancer-related death in women world-wide, and is categorized into at least five molecular subtypes that differ from each other in terms of biomarkers, etiology, and treatment modalities [25]. By far the most predominant forms of breast cancer are those that stain positive for the estrogen receptor (ER+). The ER, in particular ERα (encoded by the *ESR1* gene), has been widely studied in breast cancer [26-28]. ERα binds to estrogen (usually estradiol or E2), dimerizes, and translocates to the nucleus where it recruits co-activators or co-repressors to estrogen response elements (EREs) [28]. ERα is thought to be the primary receptor involved in the estrogen response of both normal and breast cancer cells [105].

In response to estrogen, ERα stimulates a transcriptional program involving both coding [26,29,106-109] and non-coding [41,110] RNAs. While numerous studies have investigated the coding transcriptional program, only a few studies have investigated the role played by microRNAs (miRNAs). Nevertheless, at least five miRNAs (miR-22, miR-222, miR-221, miR-18a and miR-206) have been identified that are both repressors of the *ESR1* mRNA and regulated by ERα [28]. Post-transcriptional regulation of *ESR1* is of significant interest in breast cancer because it is one proposed mechanism for loss of ERα expression in ER(-) tumors [32]. In another study, miR-375 was identified as an epigenetically deregulated miRNA that amplifies estrogen signaling in ER+ breast cancers [111]. Importantly, in that study, the inclusion of miR-

375 in a newer microarray probe set allowed the authors to identify a role for miR-375 in ER+ tumors. This highlights the importance of investigating miRNA expression via high-throughput sequencing, which is more sensitive and less biased than microarrays, and could potentially expand the set of miRNAs with potential relevance to the estrogen response.

Although high-throughput sequencing has increased in popularity and decreased in cost over the last decade, it has not been extensively applied toward the study of the gene expression response to estrogen. While microarray studies have identified a working set of estrogen responsive genes and miRNAs, these studies are limited by several factors. Firstly, microarrays can only identify targets for which probes exist and are subject to cross-hybridization errors [112]. Secondly, to date, most studies have assessed only one or two time points in response to estrogen, which fails to capture the full dynamic responses of ERα targets. ERα is known to cyclically bind to EREs and initiate bursts of transcriptional activity [39,113], and individual estrogen responsive mRNAs have been shown to exhibit diverse dynamical patterns of expression following estrogen stimulation [109]. Finally, to date no study has quantified both coding and non-coding RNAs in the same total RNA in response to estrogen. The estrogen response is highly dependent on the conditions of the study [114], and small changes in experimental design make it difficult to combine multiple studies together. In summary, the body of work on the estrogen response has demonstrated that ERα signaling enacts a dynamic and multilayered gene expression program, but we have very little understanding of how estrogen-stimulated regulatory networks change over time. The study of regulatory networks is greatly enhanced by the inclusion of temporal data, as it expands static interaction diagrams into dynamic models that can uncover complex behaviors, such as the generation of expression thresholds [42], or the existence of stable points that allow the cell to maintain expression in the absence of continued stimulation.

In this study, we investigate the global response to estrogen stimulation by analyzing paired messenger (mRNA) and miRNA measurements over time in MCF7 breast cancer cells.

We identify three major patterns of gene expression following estrogen stimulation and uncover miR-503 as an important estrogen-induced master regulator of the overall estrogen response. Based on these computational predictions, we confirm experimentally that miR-503 suppresses proliferation in breast cancer cells, and we identify a new target of miR-503, the oncogene *ZNF217*. These results provide a quantitative understanding of the temporal response of mRNAs and miRNAs to estrogen stimulation, and suggest that miR-503 is a candidate therapeutic target for treatment of breast cancer.

5.3 Results

To study the dynamics of gene expression in response to estrogen stimulation, we performed a parallel set of time-series measurements for mRNAs and microRNAs (Figure 5.1). We cultured MCF7 cells (a luminal A-type / ER+ cancer cell line) in stripped (estrogen-starved) media for 72 hours to synchronize cells in an estrogen-free state. At time zero, we supplemented the media with 10nM β estradiol (E2) and maintained the cells in this media for 1-24 hours. At each of ten time points (hourly from 0-6 hours after E2, and 8, 12 and 24 hours after E2), with three independent biological replicates for each, cells were harvested and used to prepare both small RNA and PolyA+ RNA libraries from the same total RNA sample for high-throughput sequencing. The PolyA+ RNA libraries had an average read depth of ~65 million reads (>90% of reads aligned uniquely), and the small RNA libraries had an average read depth of ~32 million reads (~90% of reads aligned). The expression levels of selected genes and miRNAs were confirmed by RT-qPCR (Supplementary Figure 5.1).

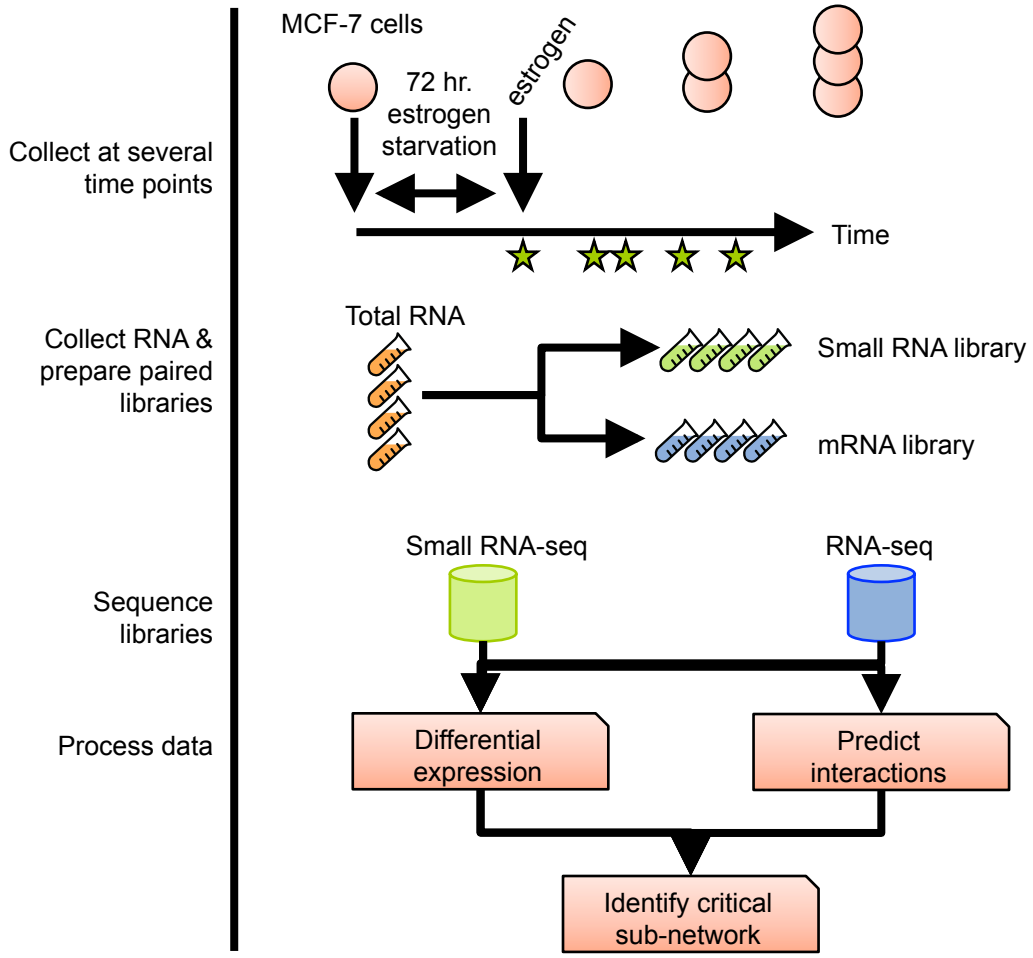


Figure 5.1: Experimental design. MCF-7 cells were cultured in stripped media for 72hrs, then 10nM E2 was added to the media. RNA was harvested at 0,1,2,3,4,5,6,8,12 and 24hrs post E2, and paired small RNA and RNA-seq libraries were generated. Each dataset was subject to differential expression analysis, and interactions were predicted between miRNAs and target mRNAs.

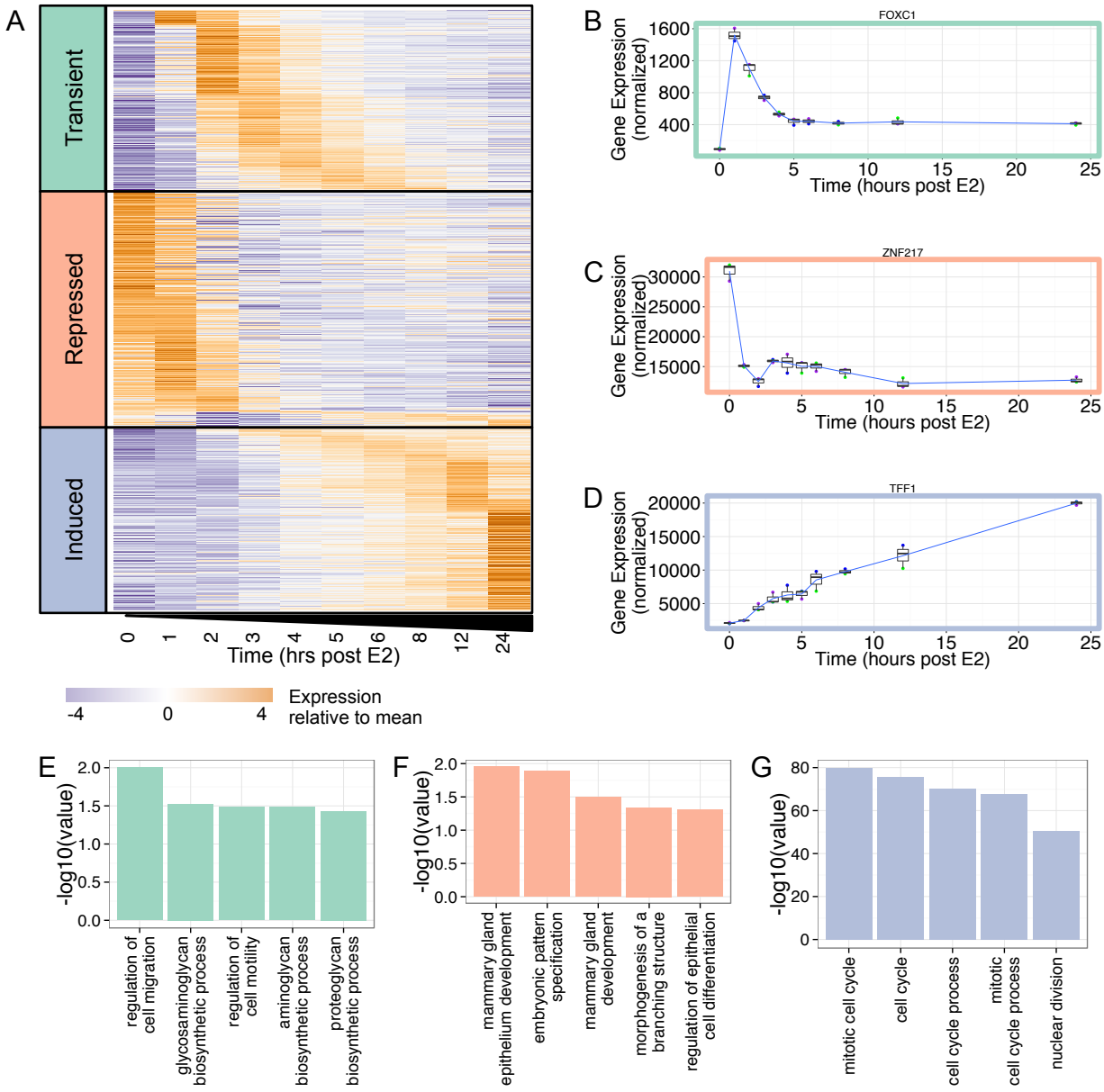


Figure 5.2: Gene expression response to estrogen stimulation. (A) The expression profile of 1546 mRNAs that have a significant (adjusted p-value <0.05) ≥2-fold change at one or more time points in response to estrogen stimulation, and a mean normalized expression of at least 500 across the time series. Genes clustered into three classes (Transient, Repressed, and Induced). (B-D) Expression of *FOXC1, ZNF217* and *TFF1,* examples of transient, repressed and induced gene classes. (E-F) Gene Ontology analysis of genes in the transient (E), repressed (F) and induced (G) gene classes.

5.3.1 Dynamics of estrogen-regulated mRNAs

Differential gene expression was calculated for each time point relative to time zero. Those genes with a mean normalized expression count across the time-series of at least 500, and a significant (adjusted p-value<0.05) ≥2-fold change relative to time zero at any time during the 24 hours, were considered to be estrogen-responsive. In total, 1546 genes met these criteria. Using this condensed data set, we then examined the dynamics of gene expression. By clustering the time series, we were able to stratify the 1546 genes into three temporal patterns of gene expression (Figure 5.2A). The first class includes 468 genes that show transient induction in response to estrogen. For example, the gene encoding the forkhead transcription factor, *FOXC1*, which is an important biomarker of basal-like breast cancer[115], is lowly expressed at time zero, exhibits a brief but significant increase in expression, and then stabilizes at a new equilibrium state around 5 hours after estrogen stimulation (Figure 5.2B). Both peak times and peak widths are variable within this class. Approximately 11% of the genes in this list have previously been identified (by meta analysis of microarray studies) to be up regulated at ~4hrs after estrogen stimulation [34]. The second class includes 608 genes that exhibit overall decreases in expression over time. These “repressed” genes include *ESR1* (gene encoding ERα)*, ERBB2, GATA3,* and *ZNF217* (Figure 5.2C)—all critical genes to the etiology of breast cancer [23,27,116,117]. *ZNF217*, a notable member of this class of genes, is a Krüppel-like finger (KLF) protein that acts as a transcriptional regulator that amplifies the estrogen response in breast cancer [117] and has been identified as a biomarker of poor survival in patients with Luminal A (ER+) breast tumors [118]. Approximately 40% of the genes in the repressed class were previously identified as down regulated at 24hrs after estrogen stimulation [34]. Finally, 470 genes are induced by estrogen stimulation. Included in this class are *TFF1* (Figure 5.2D; Supplementary Figure 5.1) and *CTSD* (Supplementary Figure 5.1)*,* for which there exists a detailed time course of ChIP data showing cyclic occupancy of ERα on their

promoters[39,113]. Approximately 56% of these genes have been previously identified as up regulated at 24hrs post-estrogen stimulation, including known breast cancer genes *BRCA1*, *BRCA2,* and *E2F1*[119]. Gene ontology (GO) analysis indicates that each of these three classes of genes is enriched for distinct functional categories (Figure 5.2E-G). The transient group of genes shows enrichment for GO terms dealing with cell migration and motility (Figure 5.2E), the repressed group for terms involving mammary development and differentiation (Figure 5.2F), and the induced category for functions relevant to cell cycle progression (Figure 5.2G). Taken together, this analysis reveals a complex and multilayered gene expression program with different temporal patterns of expression associated with distinct cellular functions.

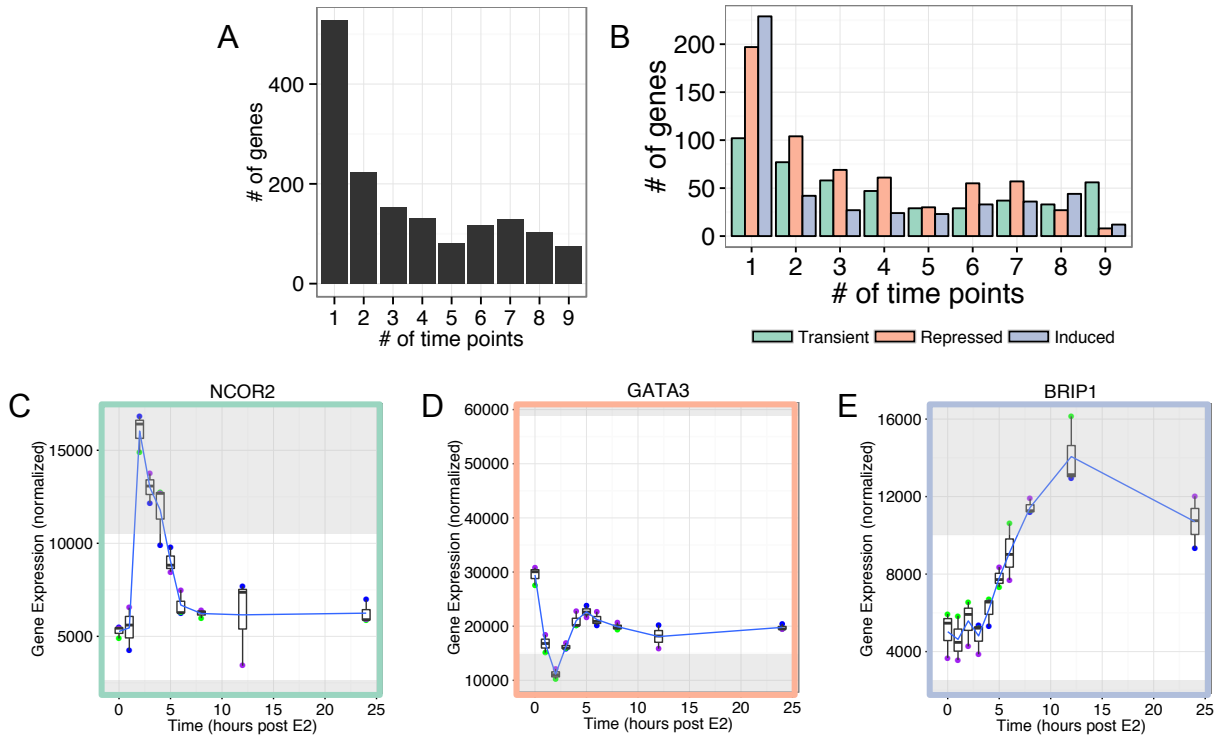


Figure 5.3: Most genes reach ≥ 2-fold change at few and disparate time points. (A) The number of time points during the 24hr collection for which each of the 1546 estrogen responsive genes reaches a ≥ 2-fold change from time zero. (B) The number of time points during the 24hr collection for which genes within the three classes reach a ≥ 2-fold change from time zero. (C) The expression profile of *NCOR2,* a representative gene from the transient class reaches ≥2-fold change from time zero at three of the time points. (D) The expression profile of *GATA3,* a representative gene from the repressed class reaches ≥2-fold change from time zero at one time point. (E) The expression profile of *BRIP1,* a representative gene from the induced class reaches ≥2-fold change from time zero at three of the time points. All plots show the mean of

three biological replicates as a blue line with box and whisker plots showing the variation in normalized expression among the replicates, and regions ≥2-fold different than time zero are shaded in gray.

It is important to note that our high-resolution temporal analysis identified many estrogen-responsive genes that would have been missed had we taken a more conservative approach and examined only one or few time points. We find that 59% (n=905) of the estrogen-responsive genes exhibit ≥2-fold change at three or fewer time points; furthermore, 34% (n=528) of the estrogen-responsive genes exhibit ≥2-fold change at only a single time point (Figure 5.3A). This observation that a large proportion of estrogen-responsive genes are significantly altered at only a few time points, and not necessarily at the same time points, is evident in all three classes of response patterns (Figure 5.3B). These data highlight the added value of a high-resolution temporal analysis. For example, consider nuclear co-repressor 2 (*NCOR2)*, which is a member of the same nuclear receptor super-family as ERα, and has been associated with early tumor recurrence in breast cancer [120]. *NCOR2* is only ≥2-fold up regulated at three out of the 10 time points analyzed (2-4 hours post estrogen stimulation) (Figure 5.3C, adjusted p-value = 4e-17 at 2hrs, 1e-11 at 3hrs, 2e-9 at 4hrs post E2). A study designed to assess estrogen-responsive genes at 12 or 24 hours post-stimulation would detect virtually no difference in gene expression levels of *NCOR2*. Additionally, within the repressed class, a group of genes drop significantly in expression in the first two hours after estrogen stimulation, but then recover at an expression level roughly 60% of the expression at time zero (Supplemental Figure 5.2A). Both *GATA3* (Figure 5.3D) and *ESR1* are members of this class*,* and these encode transcription factors that not only regulate each other [121], but also co-regulate many target genes [37]. *GATA3* is ≥2-fold down regulated only at two hours post estrogen. Nearly half of the genes in the induced class reach a ≥2-fold up-regulation (adjusted p-value < 0.05) at only a single time point (Figure 5.3B). Finally, *BRIP1* (BRCA1 interacting protein C-terminal helicase 1; Figure 5.3E) is an example of a gene in the induced class that is detected as differentially expressed at only three time points (all after eight hours post estrogen

stimulation). These observations demonstrate the dynamic nature of gene expression in response to estrogen and the importance of collecting and analyzing high-resolution time series data to fully capture these dynamics.

We also identified two groups of genes displaying opposite temporal responses to estrogen. The first group of 170 genes displays a temporal response to estrogen similar to that of *GATA3* (Supplementary Figure 5.2A). The importance of both GATA3 and ERα in regulating the estrogen response has been established [37]; therefore, these 170 genes may include other members of this regulatory network, either additional co-regulators of ERα or genes regulated by these transcription factors. The second group includes 118 genes that exhibit a temporal response to estrogen that is opposite of the *GATA3*-like genes (Supplementary Figure 5.2B). The opposite temporal responses of these groups suggest a possible inhibitory relationship between members of the two groups. To further explore this inverse relationship, we sought to identify potential negative regulators of *ESR1* or *GATA3* within the anti-*GATA3* group. To that end we assessed the correlation between both *ESR1* and *GATA3* and genes within the anti-*GATA3* group in The Cancer Genome Atlas (TCGA) breast cancer dataset (https://tcga-data.nci.nih.gov/tcga/). The transcription factor *FOXC1* (Figure 5.1B), an important biomarker of basal-like breast cancer[115], is a member of the anti-*GATA3* group and is anti correlated with *GATA3* expression in TCGA data (Supplementary Figure 5.2C; Pearson's r=-0.602). This analysis confirms previous findings that FOXC1 and GATA3 are involved in a switch (Supplementary Figure 5.2D) between basal-like and luminal-like expression programs in breast cancer [122], and indicates that the high-resolution time-series data may identify novel factors that underlie this switch.

5.3.2 Dynamics of estrogen-regulated miRNAs

To understand how the temporal gene expression patterns are regulated, we next sought to characterize the dynamics of miRNA expression in response to estrogen. Small RNA-

seq data were processed as previously described [20] to identify robustly expressed miRNAs and their isoforms (isomiRs). Resulting miRNA counts were normalized using a reads-per-million-mapped (RPMM) transformation. We detected 308 miRNAs with a mean expression of at least 50 RPMM across all samples. Consistent with previous studies of MCF7 cells [41], among the most highly expressed miRNAs in the dataset are miR-21-5p, miR-200c-3p, the let-7 family, and miR-93-5p.

To identify estrogen-responsive miRNAs, the expression of each miRNA was normalized to the mean of the three replicates at time zero. Of the 308 expressed miRNAs, 10 exhibited a fold change of at least 1.5 (uncorrected p-value <=0.05) at some time point during the 24 hours (Figure 5.4A), and 5 miRNAs had a fold change greater than 2 (uncorrected p-value <=0.05; miR-503, miR-424-3p, miR-1247-5p, miR-196a-1-5p and miR-196a-2-5p). The miRNA with the highest fold change following estrogen stimulation is miR-503 (Figure 5.4B), with a ~6-fold increase by 24 hours post estrogen stimulation. Interestingly, the second-most strongly increased miRNA is miR-424-3p (Figure 5.4C), which is encoded on the same primary transcript as miR-503. Although literature corresponding to miR-424 usually refers to miR-424-5p, in our data miR-424-3p is more consistently expressed across replicates, has a higher mean expression across the time series, and has a greater fold increase than miR-424-5p. miR-1247-5p exhibits a 3-fold increase in response to estrogen stimulation, and both paralogs of miR-196a-5p (Figure 5.4D) are ~2-fold increased by 4 hours after estrogen stimulation.

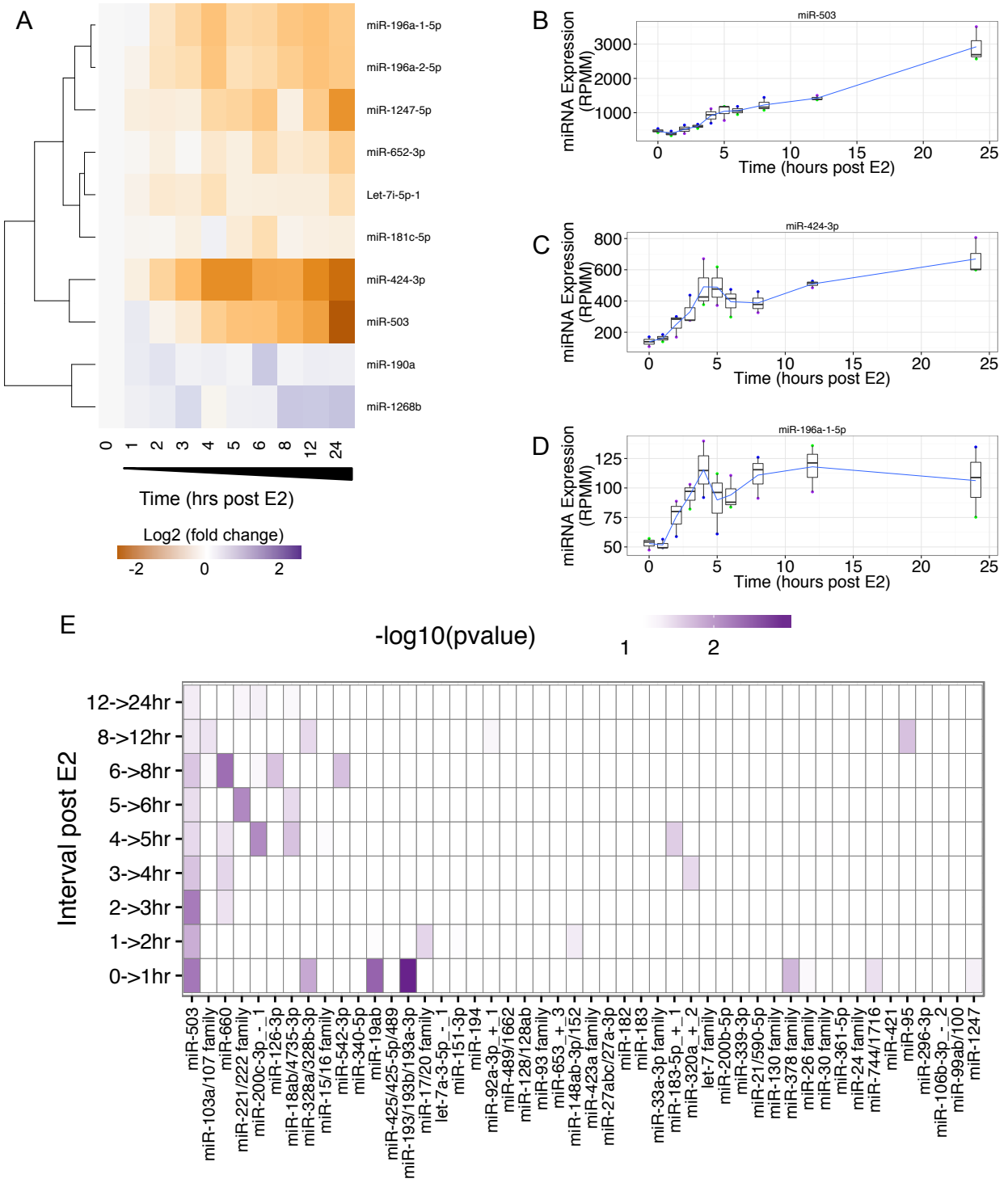


Figure 5.4: miRNA expression response to estrogen stimulation. (A) The expression profile of 10 miRNAs that have a ≥1.5 mean fold change at least one time point in response to estrogen stimulation, and a mean normalized expression of at least 50 RPMM. The expression profiles of miRNAs were clustered using a hierarchical clustering method. (B-D) Plots show the expression of the most estrogen responsive miRNA, miR-503 (B), miR-424-3p (C), and miR-196a-1-5p (D). All plots (B-D) show the mean of three biological replicates as a blue line with

box and whisker plots showing the variation in normalized expression among the replicates. (E) This plot shows the –Log10 (uncorrected p-value) of enrichment for each miRNA family among the genes that are characteristic of the change in expression between each time interval. miRNA families on the x-axis are sorted by decreasing significance (sum across all time intervals).

5.3.3 Computational prediction of miRNA-mRNA regulatory interactions

We next explored the potential regulatory interactions between miRNAs and mRNAs in the temporal response to estrogen using our previously published miRNA target site enrichment algorithm, miRhub [20]. miRhub identifies candidate master miRNA regulators by identifying those miRNAs that are predicted to target and regulate a gene set of interest significantly more than expected by chance. We sought to not only identify potential miRNA-mRNA regulatory interactions throughout the entire time course but to determine the specific time points at which these interactions were most significant. To do this, we used the Characteristic Directions method [123] to identify the sets of genes whose combined expression best distinguish the expression profiles between consecutive time points. We then assessed all expressed miRNA families to determine whether any are candidate “master regulators” of these sets of “characteristic genes”. Using this approach, miR-503 consistently emerged as the most significant candidate master miRNA regulator (Figure 5.4E). It was particularly prominent at time points 1hr, 2hr, 3hr, and 4hr post-treatment. Thus, miR-503 is both the most estrogen-responsive miRNA as well as the miRNA with the largest predicted impact on the dynamic gene expression response to estrogen.

Following our identification of miR-503 as a potential master regulator of the estrogen response, we next sought to identify potential targets of miR-503. Our miRNA target site enrichment analysis revealed 28 genes that are predicted targets of miR-503 (Figure 5.5A). One of the predicted targets, *CCND1* (Figure 5.5B), has already been validated as a target of miR-503 and the repression of *CCND1* by miR-503 has been reported to inhibit proliferation in breast cancer cell lines [124]. Another predicted target, *ZNF217* (Figure 5.5C), has not been reported

as a miR-503 target but was recently identified as both a biomarker and an oncogene in breast cancer [117]. As a final example, *RET* (Figure 5.5D) is a proto-oncogene that is reported to be transcriptionally up regulated in numerous human cancers [125]. Other genes within the list of potential miR-503 targets provide potentially interesting insights into the estrogen response regulatory network in breast cancer and warrant further study (Figure 5.5A).

Among these predicted miR-503 targets, *ZNF217* is of particular interest because of its known mechanistic role in the estrogen response. ZNF217 binds to many of the same promoters as the three transcription factors that coordinate the overall response to estrogen stimulation (ERα, GATA3 and FOXA1)[37,118]. Additionally, the c-terminus of ZNF217 physically binds to the hinge domain of ERα and enhances recruitment of ERα to EREs [126]. We carried out ChIP-X enrichment analysis (ChEA)[127] and found that ZNF217 binding sites are over-represented in estrogen responsive genes in our data set (Supplementary Figure 5.3). During the first four hours post-estrogen-treatment, the behavior of *ZNF217* bears strong resemblance to that of *GATA3*. However, unlike *GATA3*, *ZNF217* fails to recover from >2-fold repression and remains 2-fold down regulated for the rest of the time course. In fact, the first time point that *ZNF217*'s fold difference in expression deviates from that of *GATA3* is at four hours post estrogen stimulation. At that same time point (hour 4), miR-503 reaches a ~2-fold increase relative to time zero. Taking together these compelling observations with the etiological relevance of ZNF217 to breast cancer, we selected *ZNF217* as a potential target of miR-503 for further investigation.

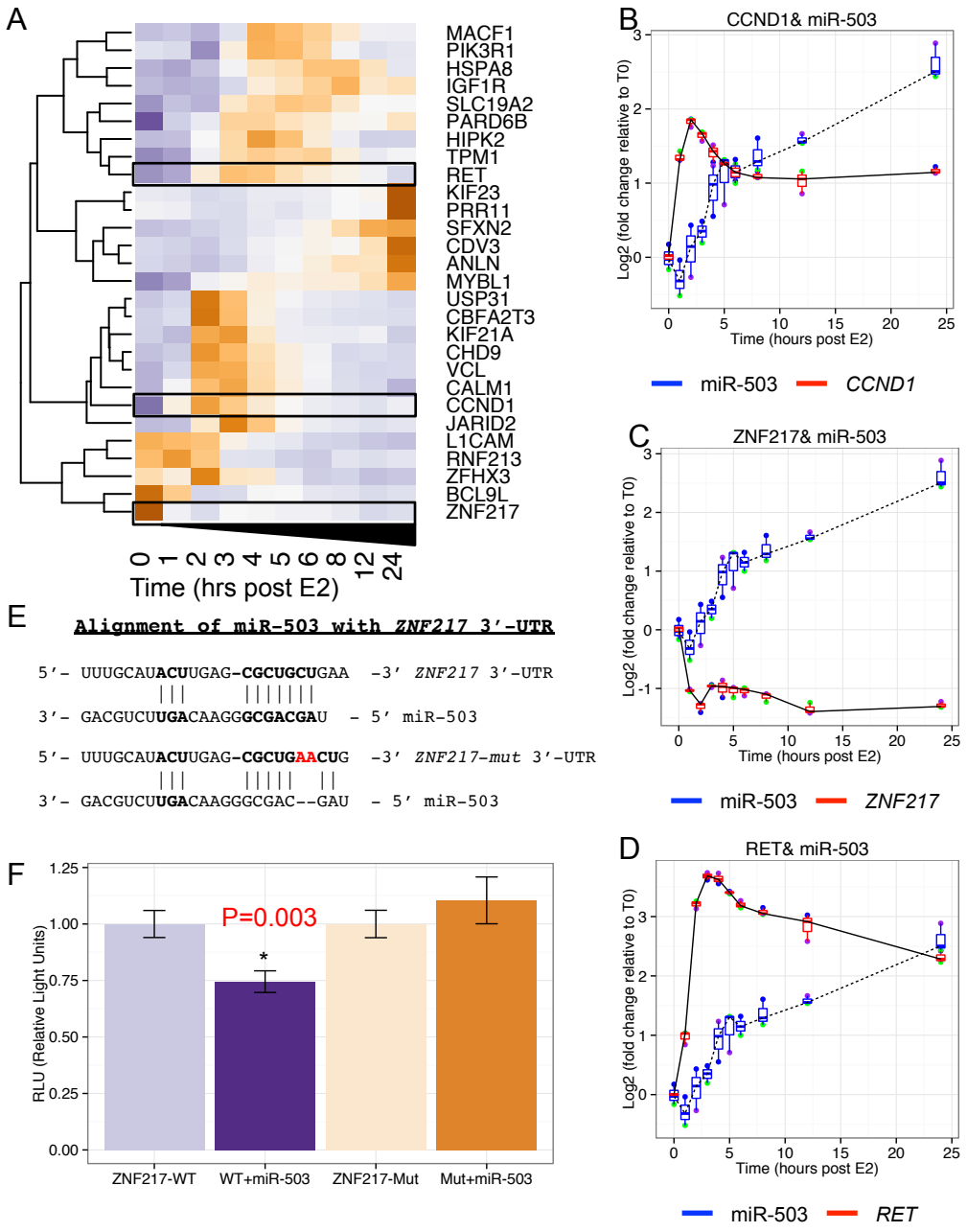


Figure 5.5: Potential miR-503 targets. (A) The expression profile of 28 miR-503 targets mRNAs that are characteristic of the difference in gene expression between consecutive time points. The expression profiles of miRNAs were clustered using a hierarchical clustering method. (B-D) Plots show the expression of miR-503 and its validated target *CCND1* (B), predicted targets *ZNF217* (C) and *RET* (D). All plots (B-D) show the mean of three biological replicates as a blue line with box and whisker plots showing the variation in log2 (fold change) between replicates. (E) miR-503 target site in the *ZNF217* 3'-UTR. A dual luciferase reporter was used to validate the response of *ZNF217* to miR-503. The reporter was mutated by inserting two As (red) to disrupt the "seed"-region binding of miR-503. (F) Response of the *ZNF217* reporter and mutant reporter with and without 10nM miR-503 mimic. Significance assessed using a student's t-test.

5.3.4 miR-503 targets ZNF217 and suppresses cellular proliferation

To validate the repression of *ZNF217* by miR-503 we carried out 3'-UTR-reporter gene assays. Specifically, we co-transfected miR-503 in MCF7 cells with dual-luciferase expression vectors containing a Renilla luciferase reporter gene (internal control) and a Firefly luciferase reporter gene linked to either the wild-type *ZNF217* 3'-UTR or a mutated version of the *ZNF217* 3'-UTR reporter with two adenosines inserted between the bases opposite of nucleotides 3 and 4 in the predicted miR-503 target site (Figure 5.5E). This mutation abolishes the perfect match to the miR-503 "seed" region, and therefore is expected to compromise the efficacy of miR-503 targeting. We found that miR-503 significantly (p=0.003) reduces the relative levels of the wild-type *ZNF217* 3'-UTR reporter, whereas the mutation in the miR-503 target site rescues this effect completely (Figure 5F). Together these results indicate that miR-503 represses *ZNF217* in MCF7 cells via direct targeting of its 3'-UTR. Because *ZNF217* is an oncogene, these findings strongly support miR-503 as a candidate tumor suppressor in breast cancer.

The role of miR-503 in opposing proliferation has been very recently investigated in breast cancer [124,128], prostate cancer [129] and osteosarcoma [130]. One previous study in MCF7 cells demonstrated that overexpression of miR-503 was able to inhibit cell cycle progression through repression of *CCND1* [124]. Based on our findings that miR-503 targets *ZNF217*, which promotes cell cycle progression [131], we hypothesized that the anti-proliferative effect of miR-503 resulted from a G1 arrest. To test this prediction, we transiently transfected MCF7 cells with miR-503 mimic (50nM). Cells were pulsed for 2 hours with EdU and fixed at 48 and 72 hours after transfection. For each cell, the total EdU signal was plotted against the total DNA signal, and cell cycle stage was assigned. We found that MCF7 cells transfected with miR-503 have a significantly reduced percentage of cells in S phase compared to mock transfection (Supplementary Figure 5.4A). This observation was observed at both 48 (p-value=0.01) and 72 (p-value=0.03) hours after transfection with miR-503 (Supplementary Figure S4B). We validated the anti-proliferative effect of miR-503 by showing reduced amount of Ki67 protein

(Supplementary Figure 5.5). Together these data confirm that miR-503 inhibits proliferation in MCF7 cells, and that the oncogene *ZNF217* is a novel target of miR-503. These findings motivate further mechanistic studies to determine whether the anti-proliferative role of miR-503 is mediated through suppression of ZNF217.

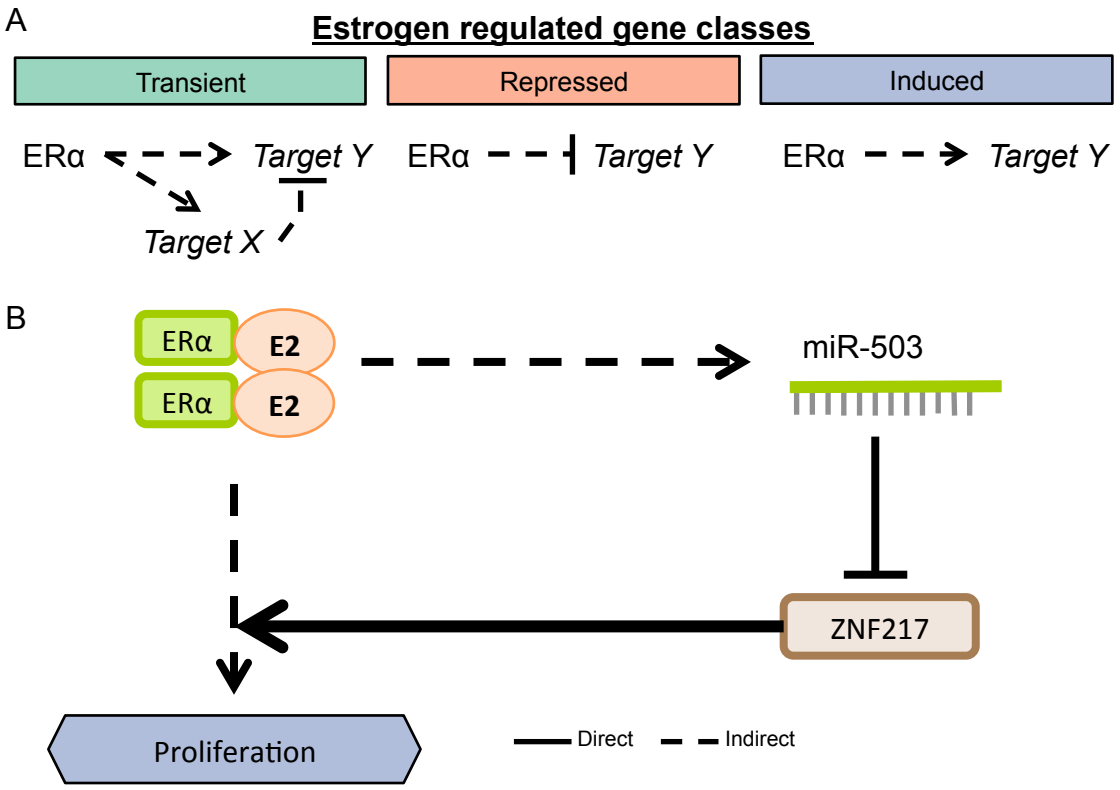


Figure 5.6: Summary. (A) Potential mechanism behind regulation of the three classes of estrogen responsive genes. (B) Summary mechanism showing the interaction of the estrogen responsive gene *ZNF217* and the estrogen responsive miRNA miR-503.

5.4 Discussion

An extensive body of literature exists detailing the importance of ERα in breast cancer, both as a biomarker of cancer severity and as a therapeutic target. Additional studies have highlighted the importance of various miRNAs in the etiology of breast cancer [41]. In this study, we provide the first detailed high throughput sequencing time course of paired mRNA and miRNA expression in response to estrogen stimulation in MCF7 cells. These data have provided

a wealth of insight into the dynamics of the estrogen response. We observe similar temporal responses for genes that encode the transcription factors previously reported to be involved in positive feedback loops (ERα and GATA3 [121]) and opposite temporal responses for genes encoding transcription factors reported to inhibit each other (GATA3 and FOXC1 [122]). These temporal relationships allow us to make inferences about the estrogen stimulated regulatory network and enhance our understanding of the timing of the cascade of signaling stimulated by estrogen in breast cancer.

This study is one of the few to investigate the temporal response of RNAs to stimuli in breast cancer, and is the first sequencing based study to investigate the matched temporal response of coding and non-coding RNAs to estrogen stimulation in breast cancer. A previous study used microarrays to investigate the response of genes and miRNAs to estrogen stimulation [40] but, interestingly, did not identify miR-503 as a significantly expressed miRNA. This discrepancy underscores the importance of using high throughput sequencing and fine-grained time resolution to study the dynamics of gene expression.

Estrogen stimulation induces a dynamic and varied response in 1546 mRNAs and 10 miRNAs. The transient class of estrogen-stimulated mRNAs contains genes that peak for various lengths of time and at different time points following estrogen stimulation (Figure 5.6A). Such behavior may be due to an incoherent feed-forward loop architecture, wherein both a target gene and its repressor are activated leading to a pulse in gene expression[4]. The repressed category also exhibits significant variation, with a large group of genes behaving similarly to *GATA3* (an initial drop in expression followed by a recovery at a new baseline expression level). Finally, the induced class of estrogen responsive mRNAs appears to either continually increase throughout the experiment as *TFF1* (Figure 5.1D) or level off at some new higher expression level. Among the estrogen responsive miRNAs, miR-503 emerges as the most strongly responsive miRNA, though others that are >2-fold changed (miR-424-3p, miR-1247-5p, miR-196a-1-5p and miR-196a-2-5p) warrant further investigation as well.

Interestingly, we also noted that a group of transient mRNAs has the exact opposite expression pattern to the *GATA3-like* group of repressed mRNAs (an initial increase followed by a reversion to a lower but still up-regulated expression level). A representative example of this group is *FOXC1* (Figure 5.1B), an important regulator of Basal-like breast cancer and a repressor of *GATA3* [122]. *FOXC1* and *GATA3* appear to be involved in a double negative-feedback loop (mutual inhibition) that may influence the transition between luminal-like and basal-like tumor phenotypes (Supplemental Figure 2D). Tkocz et. al. showed that knockdown of *GATA3* in luminal cells resulted in the expression of basal-like markers and cell morphology, while knockdown of *FOXC1* in basal-like cells resulted in the expression of luminal markers and cell morphology [122]. It is interesting to observe this pulse in a master regulator of the basal-like phenotype in response to estrogen. One explanation for this behavior is that the transition from estrogen-free media to estrogen rich-media may displace the ligand unbound estrogen receptor from its target sites and allow expression of genes that were repressed by the receptor in the absence of ligand. The introduction of estrogen would presumably then lead to the ligand-bound estrogen receptor repressing those same genes.

Among the targets (direct or indirect) of the ERα regulatory circuit are *ZNF217* and miR-503 (Figure 5.6B). ZNF217 is a transcriptional regulator that works together with ERα to amplify the estrogen response in breast cancer [117]. However, despite the association of ERα with proliferation, *ZNF217* is repressed in response to estrogen stimulation in these data indicating that there may be a built-in mechanism to avoid the *ZNF217*-induced attenuation of the estrogen response.

In this study, we identify three major patterns of gene expression following estrogen stimulation (1) transient, (2) induced, and (3) repressed. Among the genes repressed by estrogen stimulation is the oncogene *ZNF217,* which has been shown to enhance proliferation in ovarian [131] and breast cancer [132]. Additionally, we have shown that the estrogen-induced miRNA, miR-503, targets the 3'-UTR *ZNF217* and that while miR-503 is induced in response to

estrogen, the oncogene *ZNF217* is reduced. Together these observations point to the complexity of the estrogen-signaling network, and further highlight the beneficial aspects of the estrogen response. Several studies have recently shown that miR-503 may be a potent tumor suppressive miRNA. One such study showed that miR-503 targets *CCND1* and reduces proliferation in both MCF7 and MDA-MB-231 cells (an ER (-) Claudin-low model) [124]. Others have shown that miR-503 reduces proliferation and metastasis in prostate cancer [129], in osteosarcoma [130], and in hepatocellular carcinoma[133]. Furthermore, loss of miR-503 has been reported in several human cancers, and is associated with poor prognosis in cervical cancer [134]. The induction of miR-503 in response to estrogen stimulation has anti-proliferative effects, likely through its repression of both *CCND1* and *ZNF217*. Combined with the information that miR-503 is down regulated in human cancers, these results indicate that miR-503 presents a new candidate therapeutic option for the treatment of breast cancer.

5.5 Materials and methods

Cell culture and estrogen treatment

The Perou Lab at UNC Chapel Hill generously provided the MCF7 cells used in these experiments. Cells were cultured in Dulbecco's modified eagle's medium with nutrient mixture F-12 Ham, 15mM HEPES and sodium bicarbonate and without L-glutamine and phenol-red (Sigma; St. Louis, MO; #D6434) supplemented with 10% charcoal stripped serum (Sigma; St. Louis, MO; #F6765) and 5% GlutaMax (Gibco; #35050-061). For each biological replicate, a single plate of cells was split into 10 separate cell culture plates (one for each time point). Cells were maintained in the stripped serum media for 72 hours, then the time zero batch of cells was scraped from the plate, pelleted and flash frozen in ethanol dry ice slurry. For the remaining cells (other time points), media supplemented with 10nM β-estradiol (Sigma; St. Louis, MO; #E2758) was added at time zero, and cells were collected at 1, 2, 3, 4, 5, 6, 12 and 24 hours after addition of E2-media.

Sequencing and differential expression analysis

MCF7 cells were lysed and RNA was isolated using the Norgen (Ontario, Canada) Total RNA Purification Kit (#17200). Only samples with an RNA Integrity Number (RIN) of 9 or higher, as measured by Agilent (Santa Clara, CA) Bioanalyzer 2100, were considered for further analysis. Small RNA libraries were generated using the Bioo Scientific (Austin, TX) NEXTflex V2 kit (#5132-03) and sequenced on the Illumina HiSeq 2000 platforms. Small RNA-seq reads were trimmed using cutAdapt (-O 10 –e 0.1)[135] to remove remnants of the 3'-adapter sequence, then the first 4 and last 4 nucleotides of small RNA-seq reads were trimmed to remove the degenerate nucleotides in the adapters. Subsequent mapping of trimmed reads to the human genome and miRNA/isomiR quantification were performed exactly as previously described[20]. The threshold used to classify miRNAs as robustly expressed was set at a mean of 50RPMM across the time series.

RNA-seq libraries were generated from the same total RNA isolated above using the Illumina (San Diego, CA) TruSeq stranded mRNA library prep kit (#RS-122-2101) and were sequenced on the HiSeq2500 (2x50). Reads were aligned to the human genome (hg19) using MapSplice (2.1.4)[136], and transcript abundance was quantified using RSEM (v1.2.9)[137]. Finally, differentially expressed genes were identified using DEseq2 (1.4.5)[138]. The threshold used to classify mRNAs as robustly expressed was set at a mean normalized count of 500 across the time series to allow us to robustly detect changes in expression that may be well below the mean of the time series.

Clustering

To cluster the dynamic responses of estrogen regulated genes, 1D interpolation was preformed using a piecewise cubic Hermite interpolating polynomial [139] to estimate the expression at un-measured time points. Next the response of each gene was subjected to 1D wavelet decomposition [140] using a Daubechies 3 wavelet. Finally the vectors of wavelet coefficients were hierarchically clustered and split into three clusters.

Gene set and miRNA target site enrichment

Enrichment of the three classes of estrogen responsive genes within GO biological processes categories was assessed using the PANTHER overrepresentation test[141]. A selection of the most significant categories are depicted in Figure 2E-G.

miRNA target site enrichment was conducted by (1) identifying the list of miRNA families whose members have a mean expression of 50RPMM, (2) identifying lists of “characteristic genes” whose change in expression best describes the difference between consecutive time points[123], and (3) using our miRNA target site enrichment algorithm (miRhub) [20] to identify miRNA families that act as “master regulators” of the “characteristic gene sets”. miRNA target site predictions used in the miRhub enrichment algorithm are derived from TargetScan5.2 [79].

Real time quantitative PCR analysis

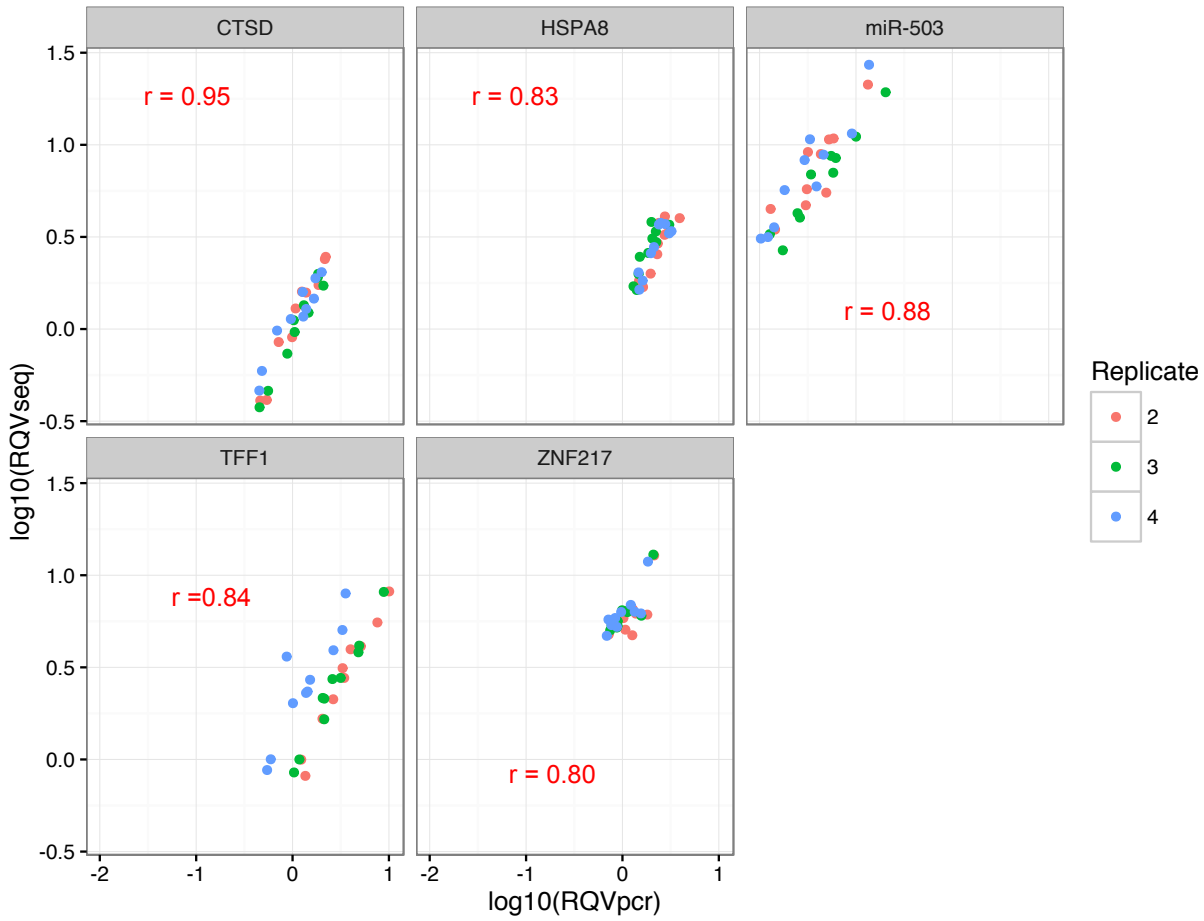
Using total RNA from above, complementary DNA (cDNA) was synthesized using the TaqMan miRNA Reverse Transcription kit (Applied Biosystems; Grand Island, NY; #4366596) according to the manufacturer’s instructions, or using the High-capacity RNA-to-cDNA kit (Applied Biossystems; Grand Island, NY; #4387406). Real-time PCR amplification of miRNAs was performed using TaqMan miRNA assays in TaqMan Universal PCR Master Mix (Applied Biosystems; Grand Island, NY; #4304437) on a BioRad CFX96 Touch Real Time PCR Detection system (Bio-Rad Laboratories, Inc., Richmond, CA). Reactions were performed in triplicate using RNU66 as the internal control. Real-time PCR amplification of mRNAs was performed using SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad Laboratories, Inc., Richmond, CA; #1725271) on a BioRad CFX96 Touch Real Time PCR Detection system (Bio-Rad Laboratories, Inc., Richmond, CA). Reactions were performed in triplicate using RPS9 as the internal control. All TaqMan assays used in this study where purchased from Applied Biosystems, Inc. (Grand Island, NY) and include: miR-503 and RNU66.

Luciferase Reporter analysis

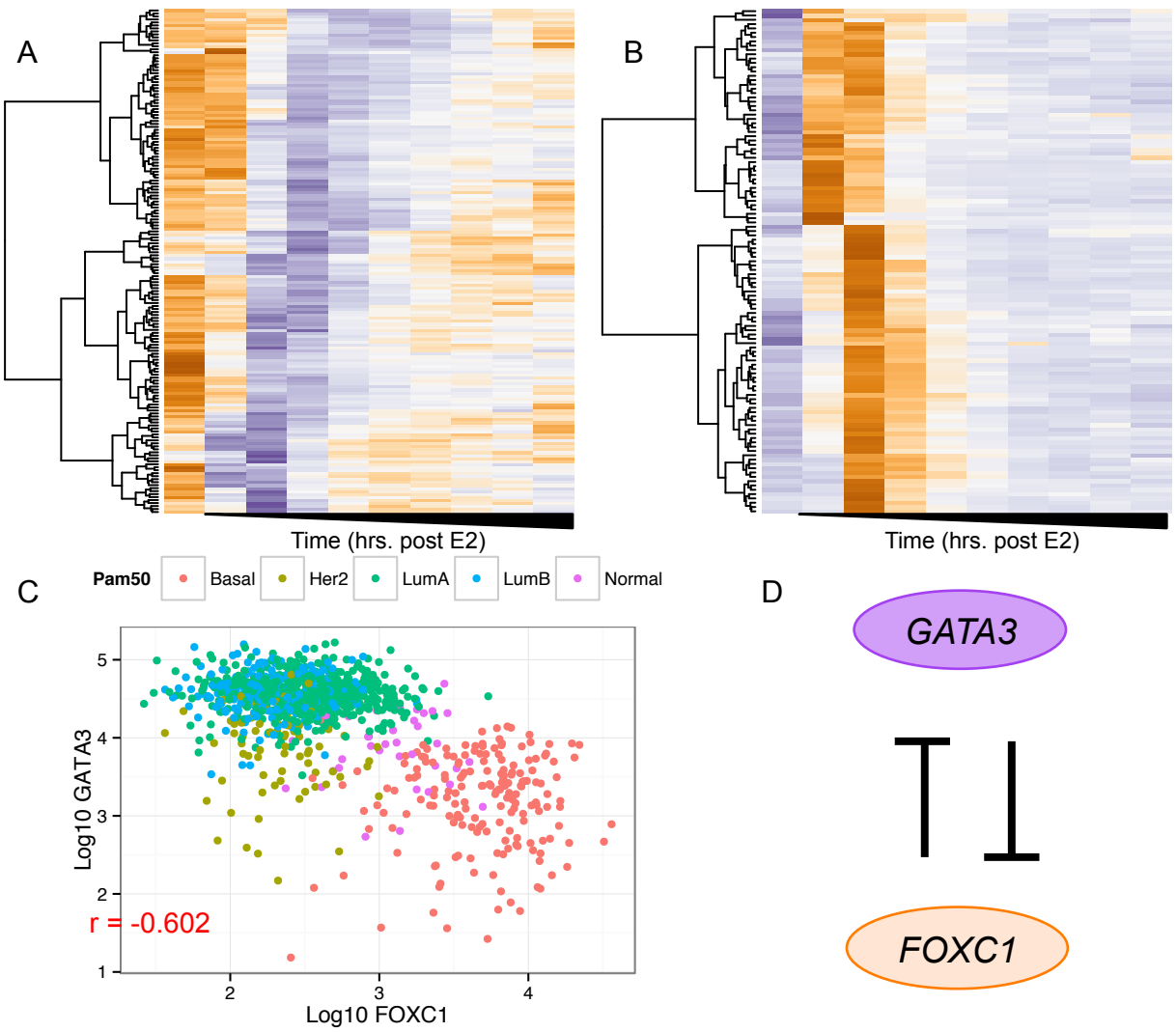
We utilized a dual luciferase reporter system (GeneCopoeia; Rockville, MD; # HmiT018728) in which the 3'-UTR of ZNF217 was fused to the end of Firefly luciferase. The construct also contains Renilla luciferase, which can be used as an internal control. Next, we mutated the vector to interfere with the binding of the miR-503 "seed" region (nucleotides 2-8) using a QuikChange II XL Site-Directed Mutagenesis Kit (Agilent; Santa Clara, CA; #200521). Two A's were added in the miR-503 target site of ZNF217 to induce a bulge in the target site opposite of nucleotides 3 and 4 of miR-503's "seed". MCF7 cells were transiently transfected with the wild-type or mutant reporter with or co-transfected with miRIDIAN microRNA Human hsa-miR-503-5p mimic (Dharmacon; Lafayette, CO #C-300841-05-0005) and with the vector. Luminescence was measured 48 hours after transfection using the Luc-Pair Duo-Luciferase Assay (GeneCopoeia; Rockville, MD; # LPFR-P030) on a Promega GloMax Multi+ Detection System luminometer.

Cell proliferation assay

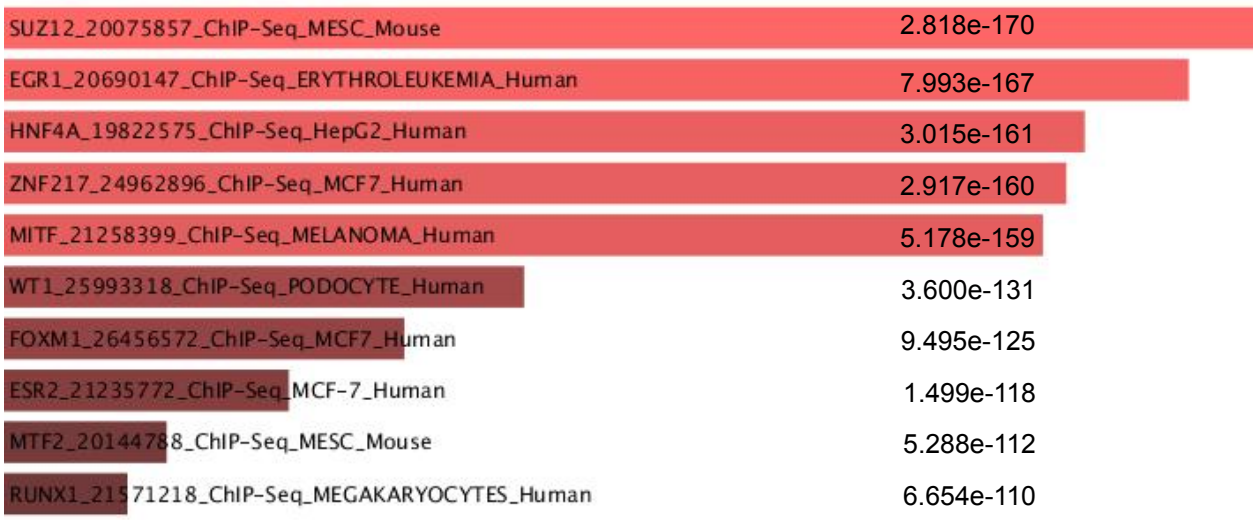
Cell proliferation was measured using the Click-iT® EdU Alexa Fluor® 488 Imaging Kit (Invitrogen; Carlsbad, CA; # C10337). MCF7 cells were transfected with 50nM miR-503 mimic or with transfection reagent only and were cultured for 46 or 70 hours. Cells were then pulsed with EdU for 2 hours before cells were fixed according to the Click-iT protocol. Finally EdU was labeled with Alexa Fluor 488 dye, DNA was stained with Hoechst and were imaged at 20x on an inverted fluorescence microscope with a Nikon TI Eclipse camera. Image segmentation analysis was performed using the Nikon Elements software package. Ki67 levels were measured using the Anti-Ki67 antibody (Abcam; Cambridge, MA; # ab15580).



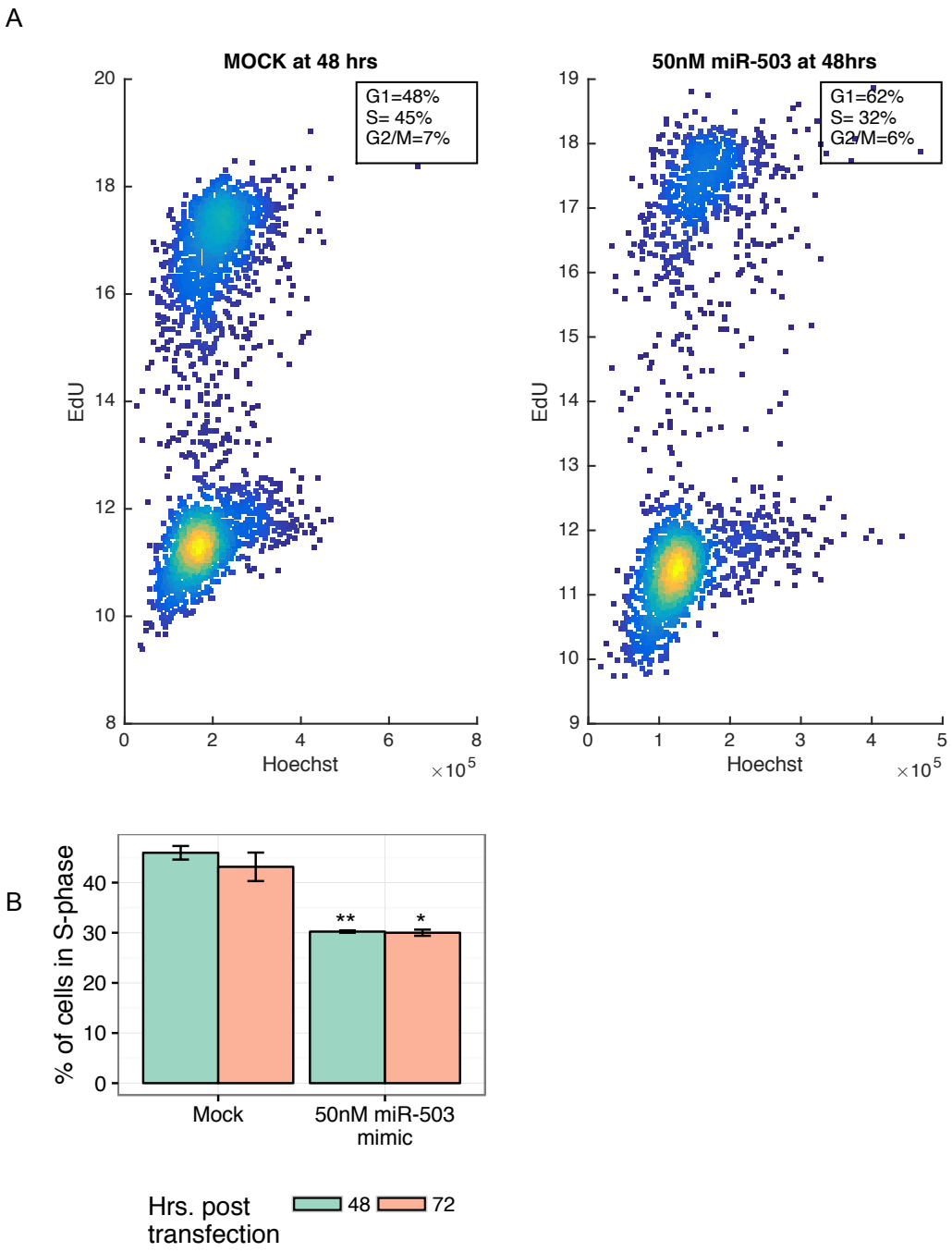
Supplemental Figure 5.1: PCR validation of selected RNAs. Comparison between the relative quantitative value (RQV) of selected RNAs as measured by high throughput sequencing (y-axis) or RT-qPCR (x-axis). Pearson's correlation value (r) is shown in red on each plot.



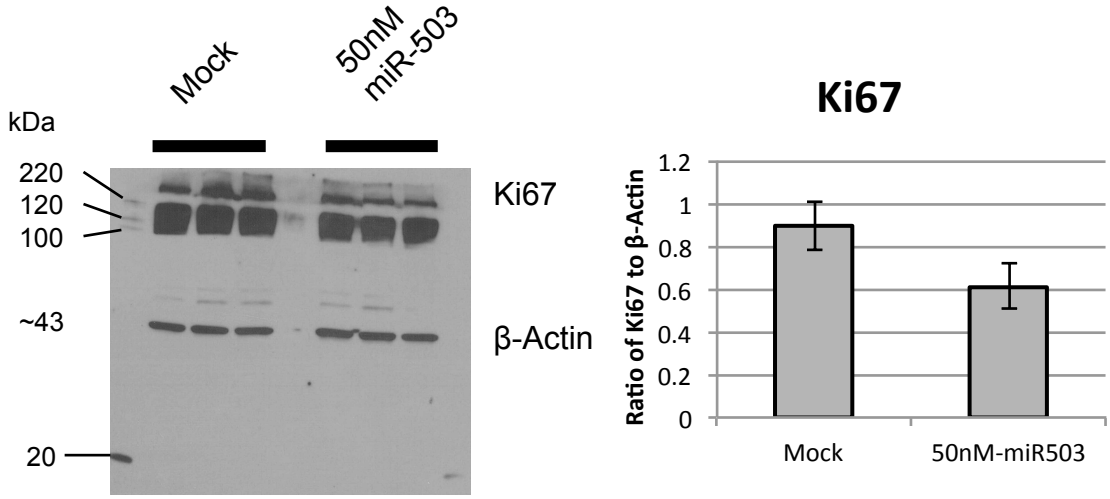
Supplemental Figure 5.2: *GATA3*-like and anti-*GATA3*-like genes. (A) The expression profile of 170 mRNAs that exhibit an expression pattern similar to *GATA3*. (B) The expression profile of 118 mRNAs that exhibit a pattern of expression opposite of that of *GATA3.* (C) Normalized expression of *GATA3* and *FOXC1* in 1046 samples from The Cancer Genome Atlas breast cancer dataset. Pearson's correlation (r) is shown in red. (D) Opposite expression pattern of *GATA3* and *FOXC1* is consistent with a mutual inhibition relationship.



Supplemental Figure 5.3: Top ChEA binding hits for estrogen responsive mRNAs. The ten most enriched transcription factor binding sites for estrogen responsive mRNAs in our dataset using ChEA. Transcription factors are sorted by p-value ranking.



Supplemental Figure 5.4: miR-503 inhibits proliferation. (A) Total EdU and DNA (Hoechst) fluorescence in cells from Mock or miR-503 (50nM) transfected MCF7 cells. (B) The percentage of MCF7 cells in S-phase during the EdU pulse experiment form Mock or miR-503 (50nM) transfected MCF7 cells at 48 or 72 hours post transfection. Significance assessed using a student's t-test (* P< 0.05, **P< 0.01).



Supplemental Figure 5.5: Ki67 protein levels in Mock and miR-503 transfected MCF7 cells. Western blot showing Ki67 protein levels in Mock or miR-503 (50nM) transfected MCF7 cells at 72 hours post transfection.

CONCLUSIONS AND FUTURE DIRECTIONS

Estrogen stimulation of ERα promotes proliferation and survival, but also opposes transformation. Utilizing the tools and techniques I developed and described in Chapters 2-4, we were able to investigate the temporal response of both mRNAs and miRNAs in MCF7 cells to estrogen stimulation. We identify three major patterns of gene expression following estrogen stimulation (1) transient, (2) induced, and (3) repressed. Among the genes repressed by estrogen stimulation is the oncogene *ZNF217,* which has been shown to enhance proliferation in ovarian [131] and breast cancer [132]. Additionally, we have shown that the estrogen-induced miRNA, miR-503, targets the 3'-UTR *ZNF217* and that while miR-503 is induced in response to estrogen, the oncogene *ZNF217* is reduced. Together these observations reveal some of the complexity of the estrogen-signaling network, and further highlight the beneficial aspects of the estrogen response. Several studies have recently shown that miR-503 may be a potent tumor suppressive miRNA. One such study showed that miR-503 targets *CCND1* and reduces proliferation in both MCF7 and MDA-MB-231 cells (an ER (-) Claudin-low model) [124]. Others have shown that miR-503 reduces proliferation and metastasis in prostate cancer [129], in osteosarcoma [130], and in hepatocellular carcinoma[133]. Furthermore, loss of miR-503 has been reported in several human cancers, and is associated with poor prognosis in cervical cancer [134]. The induction of miR-503 in response to estrogen stimulation has anti-proliferative effects, likely through its repression of both *CCND1* and *ZNF217*. Combined with the information that miR-503 is down regulated in human cancers, these results indicate that miR-503 presents a new candidate therapeutic option for the treatment of breast cancer.

The dynamic map of the response of mRNAs and miRNAs to estrogen stimulation generated in this study is a resource that can continue to be mined to identify additional interactions relevant to the estrogen response. For example, we identified groups of genes that exhibit opposite temporal patterns of expression (the *GATA3*-like and anti-*GATA3*-like groups). Investigation of the expression of these groups of genes within the TCGA breast cancer dataset led us to hypothesize that the genes *GATA3* and *FOXC1* might be involved in a double negative feedback loop. A literature search for these two genes uncovered a study by TKocz et. al. that showed knockdown of *GATA3* in luminal cells resulted in the expression of basal-like markers and cell morphology, while knockdown of *FOXC1* in basal-like cells resulted in the expression of luminal markers and cell morphology. The *GATA3* - *FOXC1* network warrants further investigation and this mechanism is one example of a system that would benefit from the type of study described in Chapter 5. The transition between luminal and basal morphologies in response to knockdown of *GATA3* in luminal-like cells and *FOXC1* in basal-like cells provides us with a system that can be studied to identify the steps involved the reprogramming of breast cancer cells from a more mild luminal subtype to a more aggressive basal subtype and back. Investigation of this system may uncover additional insights into the progression of breast cancer. More generally, as most biological processes are dynamic, the type of study described in Chapter 5 can be used to investigate any such process.

One natural extension of the work described in this thesis is to incorporate additional types of data. For example, one such data type is the expression of long non-coding RNAs (lncRNAs). HTS studies have revealed that as much as 85% of the human genome is transcribed, and much of this transcription produces non-coding RNAs[142]. miRNAs are one class of short non-coding RNAs, and years of research has revealed some of the features and functions of miRNAs. By comparison, the study of long non-coding RNAs is in its infancy. Nevertheless, functionality of individual lncRNAs has been demonstrated. For example, the X inactive specific transcript (*XIST*) is involved in X chromosome inactivation [143]. The RNA-seq

data described in Chapter 5 can also reveal lncRNAs that respond to estrogen stimulation. A preliminary lncRNAs analysis of the data shows that the lncRNA *GATA3-AS1* exhibits a similar temporal pattern of expression to *GATA3*. However, *GATA3-AS1* fails to recover following the dip in express at 2 hours post estrogen stimulation. An examination of estrogen responsive lncRNAs may reveal other non-coding RNAs that may play a role in breast cancer.

Alternative polyadenylation (APA) of mRNAs is prevalent during proliferation and development [144,145]. In cancer, widespread shorting of 3'-UTRs has been observed [146,147], and the expression of shorter 3'-UTR isoforms is correlated with poor prognosis in breast and lung cancer[148]. Although, not discussed in this thesis, we have also collected matched 3'-end sequencing data for several time points following estrogen stimulation. These data will enable us to identify not only which 3'-UTR isoforms are expressed, but also whether 3'-UTR isoforms are differentially expressed in response to estrogen. Such information would provide another layer of data that can help to uncover the mechanisms underlying the regulatory architecture that underlies the estrogen response. As miRNAs bind to their targets primarily within the 3'-UTR region, accurate annotation of expressed 3'-UTRs is critical to identifying potential miRNA targets.

In Chapter 5 we identify miR-503 as an estrogen responsive miRNA that inhibits proliferation and represses *ZNF217.* Others have demonstrated that miR-503 also represses *CCND1*[124], and that miR-503 inhibits proliferation in prostate cancer[129], osteosarcoma [130] and hepatocellular carcinoma[133]. Another extension of this work would be to investigate the estrogen response of miR-503 knockout MCF7 cells generated by deleting the miR-503 precursor using a CRISPR-Cas9 system[149]. The MCF7-ΔmiR-503 cells could be studied to identify any potential phenotypic differences, and the same estrogen response time-course could be repeated in the knockout cells. Such a study could lead to further insights into the importance of miR-503 in the estrogen response.

Finally, the inclusion of additional cell types would greatly improve these data. Other ER+ cell lines such as T47D or ZR-75-1 could be exposed to the same estrogen response time-course to determine if the observations from Chapter 5 are consistent across other ER+ models of breast cancer. Such data would more strongly indicate that the estrogen-signaling network in breast cancer produces the responses we observed in mRNAs and miRNAs, and that those responses are not specific to the MCF7 cell line.

Overall, the tools and techniques I developed and described in Chapters 2-4 have enabled the Sethupathy and Purvis labs to investigate miRNA regulation in numerous cell and tissue types. Chapter 5 describes the application of these methods to the field of breast cancer and the estrogen response, but these methods have also been applied to the study of Type 2 diabetes[16,20,21], Crohn's disease [17], Hepatitis B and C [18] and the response of MCF7 cells to different dynamic expression patterns of the tumor suppressor p53 (unpublished). Moreover, the study design described in Chapter 5 can be applied to study the dynamics of any biological process.

REFERENCES

1. Batchelor E, Loewer A, Mock C, Lahav G. Stimulus-dependent dynamics of p53 in single cells. Molecular Systems Biology. 2011;7: 488–488. doi:10.1038/msb.2011.20

2. Purvis JE, Karhohs KW, Mock C, Batchelor E, Loewer A, Lahav G. p53 dynamics control cell fate. Science. 2012;336: 1440–1444. doi:10.1126/science.1218351

3. Santos SDM, Verveer PJ, Bastiaens PIH. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. Nat Cell Biol. 2007;9: 324–330. doi:10.1038/ncb1543

4. Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet. 2007;8: 450–461. doi:10.1038/nrg2102

5. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Molecular Cell. 2007.

6. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136: 215–233. doi:10.1016/j.cell.2009.01.002

7. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5: 522–531. doi:10.1038/nrg1379

8. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004;23: 4051–4060. doi:10.1038/sj.emboj.7600385

9. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014. doi:10.1038/nrm3838

10. Xiao M, Li J, Li W, Wang Y, Wu F, Xi Y, et al. MicroRNAs Activate Gene Transcription Epigenetically as an Enhancer Trigger. RNA Biol. 2016;: 0. doi:10.1080/15476286.2015.1112487

11. Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes & Development. 2010;24: 992–1009. doi:10.1101/gad.1884710

12. Fuchs RT, Sun Z, Zhuang F, Robb GB. Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. Zhang B, editor. PLoS ONE. 2015;10: e0126049. doi:10.1371/journal.pone.0126049

13. Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011;17: 1697–1712. doi:10.1261/rna.2799511

14. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and

remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Research. 2011;39: e141–e141. doi:10.1093/nar/gkr693

15. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. Nucleic Acids Research. 2012;40: e54–e54. doi:10.1093/nar/gkr1263

16. Kurtz CL, Peck BCE, Fannin EE, Beysen C, Miao J, Landstreet SR, et al. microRNA-29 fine-tunes the expression of key FOXA2-activated lipid metabolism genes and is dysregulated in animal models of insulin resistance and diabetes. Diabetes. 2014;63: 3141–3148. doi:10.2337/db13-1015

17. Peck BCE, Weiser M, Lee SE, Gipson GR, Iyer VB, Sartor RB, et al. MicroRNAs Classify Different Disease Behavior Phenotypes of Crohn's Disease and May Have Prognostic Utility. Inflamm Bowel Dis. 2015;21: 2178–2187.

18. Selitsky SR, Dinh TA, Toth CL, Kurtz CL, Honda M, Struck BR, et al. Transcriptomic Analysis of Chronic Hepatitis B and C and Liver Cancer Reveals MicroRNA-Mediated Control of Cholesterol Synthesis Programs. MBio. 2015;6: e01500–15. doi:10.1128/mBio.01500-15

19. Baran-Gale J, Kurtz CL, Erdos MR, Sison C, Young A, Fannin EE, et al. Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. Front Gene. 2015;6: 352. doi:10.3389/fgene.2015.00352

20. Baran-Gale J, Fannin EE, Kurtz CL, Sethupathy P. Beta Cell 5′-Shifted isomiRs Are Candidate Regulatory Hubs in Type 2 Diabetes. PLoS ONE. Public Library of Science; 2013;8: e73240. doi:10.1371/journal.pone.0073240

21. Kurtz CL, Fannin EE, Toth CL, Pearson DS, Vickers KC, Sethupathy P. Inhibition of miR-29 has a significant lipid-lowering benefit through suppression of lipogenic programs in liver. Sci Rep. 2015;5: 12911. doi:10.1038/srep12911

22. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66: 7–30. doi:10.3322/caac.21332

23. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA. 2001;98: 10869–10874. doi:10.1073/pnas.191367098

24. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. 2013. pp. 2206–2223. doi:10.1093/annonc/mdt303

25. Sabatier R, Gonçalves A, Bertucci F. Personalized medicine: Present and future of breast cancer management. Crit Rev Oncol Hematol. 2014. doi:10.1016/j.critrevonc.2014.03.002

26. Creighton CJ, Cordero KE, Larios JM, Miller RS, Johnson MD, Chinnaiyan AM, et al. Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. Genome Biol. 2006;7: R28. doi:10.1186/gb-2006-7-4-r28

27. Oh DS, Troester MA, Usary J, Hu Z, He X, Fan C, et al. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J Clin Oncol. 2006;24: 1656–1664. doi:10.1200/JCO.2005.03.2755

28. Zhou W, Slingerland JM. Links between oestrogen receptor activation and proteolysis: relevance to hormone-regulated cancer therapy. Nat Rev Cancer. 2014;14: 26–38.

29. Chang EC, Charn TH, Park S-H, Helferich WG, Komm B, Katzenellenbogen JA, et al. Estrogen Receptors alpha and beta as determinants of gene expression: influence of ligand, dose, and chromatin binding. 2008;22: 1032–1043. doi:10.1210/me.2007-0356

30. Guttilla IK, Adams BD, White BA. ERα, microRNAs, and the epithelial–mesenchymal transition in breast cancer. Trends in Endocrinology & Metabolism. 2012;23: 73–82. doi:10.1016/j.tem.2011.12.001

31. Kaboli PJ, Rahmat A, Ismail P, Ling K-H. MicroRNA-based therapy and breast cancer: A comprehensive review of novel therapeutic strategies from diagnosis to treatment. Pharmacological Research. 2015. doi:10.1016/j.phrs.2015.04.015

32. Lapidus RG, Ferguson AT, Ottaviano YL, Parl FF, Smith HS, Weitzman SA, et al. Methylation of estrogen and progesterone receptor gene 5' CpG islands correlates with lack of estrogen and progesterone receptor gene expression in breast tumors. Clin Cancer Res. 1996;2: 805–810.

33. Yoshimoto N, Toyama T, Takahashi S, Sugiura H, Endo Y, Iwasa M, et al. Distinct expressions of microRNAs that directly target estrogen receptor α in human breast cancer. Breast Cancer Res Treat. 2011;130: 331–339. doi:10.1007/s10549-011-1672-2

34. Jagannathan V, Robinson-Rechavi M. Meta-analysis of estrogen response in MCF-7 distinguishes early target genes involved in signaling and cell proliferation from later target genes involved in cell cycle and DNA repair. BMC Syst Biol. 2011;5: 138. doi:10.1186/1752-0509-5-138

35. Stanculescu A, Bembinster LA, Borgen K, Bergamaschi A, Wiley E, Frasor J. Estrogen promotes breast cancer cell survival in an inhibitor of apoptosis (IAP)-dependent manner. Horm Cancer. 2010;1: 127–135. doi:10.1007/s12672-010-0018-6

36. Song RX-D, McPherson RA, Adam L, Bao Y, Shupnik M, Kumar R, et al. Linkage of rapid estrogen action to MAPK activation by ERalpha-Shc association and Shc pathway activation. Mol Endocrinol. 2002;16: 116–127. doi:10.1210/mend.16.1.0748

37. Kong SL, Li G, Loh SL, Sung W-K, Liu ET. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. Molecular Systems Biology. 2011;7: 526–526. doi:10.1038/msb.2011.59

38. Guttilla IK, Phoenix KN, Hong X, Tirnauer JS, Claffey KP, White BA. Prolonged mammosphere culture of MCF-7 cells induces an EMT and repression of the estrogen receptor by microRNAs. Breast Cancer Res Treat. 2012;132: 75–85. doi:10.1007/s10549-011-1534-y

39. Reid G, Hübner MR, Métivier R, Brand H, Denger S, Manu D, et al. Cyclic, proteasome-mediated turnover of unliganded and liganded ERalpha on responsive promoters is an integral feature of estrogen signaling. Molecular Cell. 2003;11: 695–707.

40. Cicatiello L, Mutarelli M, Grober OMV, Paris O, Ferraro L, Ravo M, et al. Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. Am J Pathol. 2010;176: 2113–2130. doi:10.2353/ajpath.2010.090837

41. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, et al. Estradiol-regulated microRNAs control estradiol response in breast cancer cells. Nucleic Acids Research. 2009;37: 4850–4861. doi:10.1093/nar/gkp500

42. Mukherji S, Ebert MS, Zheng GXY, Tsang JS, Sharp PA, van Oudenaarden A. MicroRNAs can generate thresholds in target gene expression. Nat Genet. Nature Publishing Group; 2011;43: 854–859. doi:10.1038/ng.905

43. Stuwe E, Tóth KF, Aravin AA. Small but sturdy: small RNAs in cellular memory and epigenetics. Genes & Development. 2014;28: 423–431. doi:10.1101/gad.236414.113

44. Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol. 2005;6: 376–385. doi:10.1038/nrm1644

45. Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol. 2009;10: 126–139. doi:10.1038/nrm2632

46. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116: 281–297.

47. Hudder A, Novak RF. miRNAs: effectors of environmental influences on gene expression and disease. Toxicol Sci. 2008;103: 228–240. doi:10.1093/toxsci/kfn033

48. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, MacDonald PE, et al. A pancreatic islet-specific microRNA regulates insulin secretion. Nature. 2004;432: 226–230. doi:10.1038/nature03076

49. Poy MN, Hausser J, Trajkovski M, Braun M, Collins S, Rorsman P, et al. miR-375 maintains normal pancreatic alpha- and beta-cell mass. Proc Natl Acad Sci USA. 2009;106: 5813–5818. doi:10.1073/pnas.0810550106

50. Landgraf P, Rusu M, Sheridan R, Alain Sewer, Nicola Iovino, Alexei Aravin, et al. A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. Trends in Biochemical Sciences. Elsevier; 2007;129: 1401–1414. doi:10.1016/j.cell.2007.04.040

51. Neilsen CT, Goodall GJ, Bracken CP. IsomiRs--the overlooked repertoire in the dynamic microRNAome. Trends Genet. 2012;28: 544–549. doi:10.1016/j.tig.2012.07.005

52. Kim Y-K, Heo I, Kim VN. Modifications of small RNAs and their associated proteins. Cell. Elsevier Inc; 2010;143: 703–709. doi:10.1016/j.cell.2010.11.018

53. Gunaratne PH, Coarfa C, Soibam B, Tandon A. miRNA data analysis: next-gen sequencing. In: Fan J-B, editor. Methods in molecular biology; 2012. pp. 273–288. doi:10.1007/978-1-61779-427-8_19

54. Burroughs AM, Kawano M, Ando Y, Daub CO, Hayashizaki Y. pre-miRNA profiles obtained through application of locked nucleic acids and deep sequencing reveals complex 5“/3” arm variation including concomitant cleavage and polyuridylation patterns. Nucleic Acids Research. 2012;40: 1424–1437. doi:10.1093/nar/gkr903

55. Liu N, Abe M, Sabin LR, Hendriks G-J, Naqvi AS, Yu Z, et al. The Exoribonuclease Nibbler Controls 3′ End Processing of MicroRNAs in Drosophila. Current Biology. 2011;21: 1888–1893. doi:10.1016/j.cub.2011.10.006

56. Han BW, Hung J-H, Weng Z, Zamore PD, Ameres SL. The 3′-to-5′ Exoribonuclease Nibbler Shapes the 3′ Ends of MicroRNAs Bound to Drosophila Argonaute1. Current Biology. 2011;21: 1878–1887. doi:10.1016/j.cub.2011.09.034

57. Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, et al. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. Genome Research. 2011;21: 1450–1461. doi:10.1101/gr.118059.110

58. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79: 321–349. doi:10.1146/annurev-biochem-060208-105251

59. Reese TA, Xia J, Johnson LS, Zhou X, Zhang W, Virgin HW. Identification of novel microRNA-like molecules generated from herpesvirus and host tRNA transcripts. J Virol. 2010;84: 10344–10353. doi:10.1128/JVI.00707-10

60. Sdassi N, Silveri L, Laubier J, Tilly G, Costa J, Layani S, et al. Identification and characterization of new miRNAs cloned from normal mouse mammary gland. BMC Genomics. 2009;10: 149. doi:10.1186/1471-2164-10-149

61. Lee LW, Zhang S, Etheridge A, Ma L, Martin D, Galas D, et al. Complexity of the microRNA repertoire revealed by next-generation sequencing. RNA. 2010;16: 2170–2180. doi:10.1261/rna.2225110

62. Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, et al. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome

Biol. 2011;12: R126. doi:10.1186/gb-2011-12-12-r126

63. Burroughs AM, Ando Y, de Hoon MJL, Tomaru Y, Suzuki H, Hayashizaki Y, et al. Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. RNA Biol. 2011;8: 158–177.

64. Zhou H, Arcila ML, Li Z, Lee EJ, Henzler C, Liu J, et al. Deep annotation of mouse iso-miR and iso-moR variation. Nucleic Acids Research. 2012;40: 5864–5875. doi:10.1093/nar/gks247

65. Burroughs AM, Ando Y, de Hoon MJL, Tomaru Y, Nishibu T, Ukekawa R, et al. A comprehensive survey of 3“ animal miRNA modification events and a possible role for 3” adenylation in modulating miRNA targeting effectiveness. Genome Research. 2010;20: 1398–1410. doi:10.1101/gr.106054.110

66. Fehniger TA, Wylie T, Germino E, Leong JW, Magrini VJ, Koul S, et al. Next-generation sequencing identifies the natural killer cell microRNA transcriptome. Genome Research. 2010;20: 1590–1604. doi:10.1101/gr.107995.110

67. Newman MA, Mani V, Hammond SM. Deep sequencing of microRNA precursors reveals extensive 3' end modification. RNA. 2011;17: 1795–1803. doi:10.1261/rna.2713611

68. Westholm JO, Ladewig E, Okamura K, Robine N, Lai EC. Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs. RNA. 2012;18: 177–192. doi:10.1261/rna.030627.111

69. Fernandez-Valverde SL, Taft RJ, Mattick JS. Dynamic isomiR regulation in Drosophila development. RNA. 2010;16: 1881–1888. doi:10.1261/rna.2379610

70. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science. 2007;315: 1137–1140. doi:10.1126/science.1138050

71. Alon S, Mor E, Vigneault F, Church GM, Locatelli F, Galeano F, et al. Systematic identification of edited microRNAs in the human brain. Genome Research. 2012;22: 1533–1540. doi:10.1101/gr.131573.111

72. Ekdahl Y, Farahani HS, Behm M, Lagergren J, Öhman M. A-to-I editing of microRNAs in the mammalian brain increases during development. Genome Research. 2012;22: 1477–1487. doi:10.1101/gr.131912.111

73. Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. Nature Publishing Group; 2012;30: 253–260. doi:doi:10.1038/nbt.2122

74. Martí E, Pantano L, Bañez-Coronel M, Llorens F, Miñones-Moyano E, Porta S, et al. A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. Nucleic Acids Research. 2010;38: 7219–7235. doi:10.1093/nar/gkq575

75. Azuma-Mukai A, Oguri H, Mituyama T, Qian ZR, Asai K, Siomi H, et al. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. Proc Natl Acad Sci USA. 2008;105: 7964–7969. doi:10.1073/pnas.0800334105

76. Voellenkle C, Rooij JV, Guffanti A, Brini E, Fasanaro P, Isaia E, et al. Deep-sequencing of endothelial cells exposed to hypoxia reveals the complexity of known and novel microRNAs. RNA. 2012;18: 472–484. doi:10.1261/rna.027615.111

77. Ishihara H, Asano T, Tsukuda K, Katagiri H, Inukai K, Anai M, et al. Pancreatic beta cell line MIN6 exhibits characteristics of glucose metabolism and glucose-stimulated insulin secretion similar to those of normal islets. Diabetologia. 1993;36: 1139–1145. doi:10.1007/BF00401058

78. van de Bunt M, Gaulton KJ, Parts L, Moran I, Johnson PR, Lindgren CM, et al. The miRNA Profile of Human Pancreatic Islets and Beta-Cells and Relationship to Type 2 Diabetes Pathogenesis. PLoS ONE. Public Library of Science; 2013;8: e55272. doi:10.1371/journal.pone.0055272

79. Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular Cell. 2007;27: 91–105. doi:10.1016/j.molcel.2007.06.017

80. Couzin J. MicroRNAs make big impression in disease after disease. Science. 2008.: 1782–1784. doi:10.1126/science.319.5871.1782

81. Baker M. MicroRNA profiling: separating signal from noise. Nature methods. 2010;7: 687–692. doi:10.1038/nmeth0910-687

82. Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. Nat Rev Genet. 2012;13: 358–369. doi:doi:10.1038/nrg3198

83. Tian G, Yin X, Luo H, Xu X, Bolund L, Zhang X, et al. Sequencing bias: comparison of different protocols of microRNA library construction. BMC Biotechnol. 2010;10: 64. doi:10.1186/1472-6750-10-64

84. Linsen SEV, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, et al. Limitations and possibilities of small RNA digital gene expression profiling. Nature methods. 2009;6: 474–476. doi:10.1038/nmeth0709-474

85. Kawano M, Kawazu C, Lizio M, Kawaji H, Carninci P, Suzuki H, et al. Reduction of non-insert sequence reads by dimer eliminator LNA oligonucleotide for small RNA deep sequencing. BioTechniques. 2010;49: 751–755. doi:10.2144/000113516

86. Zhu Y, You W, Wang H, Li Y, Qiao N, Shi Y, et al. MicroRNA-24/MODY gene regulatory pathway mediates pancreatic β-cell dysfunction. Diabetes. 2013;62: 3194–3206. doi:10.2337/db13-0151

87. Pullen TJ, da Silva Xavier G, Kelsey G, Rutter GA. miR-29a and miR-29b contribute to pancreatic beta-cell-specific silencing of monocarboxylate transporter 1 (Mct1). Molecular and Cellular Biology. 2011;31: 3182–3194. doi:10.1128/MCB.01433-10

88. Liao X, Xue H, Wang Y-C, Nazor KL, Guo S, Trivedi N, et al. Matched miRNA and mRNA signatures from an hESC-based in vitro model of pancreatic differentiation reveal novel regulatory interactions. J Cell Sci. 2013;126: 3848–3861. doi:10.1242/jcs.123570

89. Melkman-Zehavi T, Oren R, Kredo-Russo S, Shapira T, Mandelbaum AD, Rivkin N, et al. miRNAs control insulin content in pancreatic β-cells via downregulation of transcriptional repressors. EMBO J. 2011;30: 835–845. doi:10.1038/emboj.2010.361

90. Latreille M, Hausser J, Stützer I, Zhang Q, Hastoy B, Gargani S, et al. MicroRNA-7a regulates pancreatic β cell function. J Clin Invest. 2014;124: 2722–2735. doi:10.1172/JCI73066

91. Roggli E, Gattesco S, Caille D, Briet C, Boitard C, Meda P, et al. Changes in microRNA expression contribute to pancreatic β-cell dysfunction in prediabetic NOD mice. Diabetes. 2012;61: 1742–1751. doi:10.2337/db11-1086

92. Frost RJA, Olson EN. Control of glucose homeostasis and insulin sensitivity by the Let-7 family of microRNAs. Proc Natl Acad Sci USA. 2011;108: 21075–21080. doi:10.1073/pnas.1118922109

93. Tsai W-C, Hsu S-D, Hsu C-S, Lai T-C, Chen S-J, Shen R, et al. MicroRNA-122 plays a critical role in liver homeostasis and hepatocarcinogenesis. J Clin Invest. 2012;122: 2884–2897. doi:10.1172/JCI63455

94. Bandiera S, Pfeffer S, Baumert TF, Zeisel MB. miR-122--a key factor and therapeutic target in liver disease. J Hepatol. 2015;62: 448–457. doi:10.1016/j.jhep.2014.10.004

95. Alon S, Vigneault F, Eminaga S, Christodoulou DC, Seidman JG, Church GM, et al. Barcoding bias in high-throughput multiplex sequencing of miRNA. Genome Research. 2011;21: 1506–1511. doi:10.1101/gr.121715.111

96. Van Nieuwerburgh F, Soetaert S, Podshivalova K, Ay-Lin Wang E, Schaffer L, Deforce D, et al. Quantitative bias in Illumina TruSeq and a novel post amplification barcoding strategy for multiplexed DNA and small RNA deep sequencing. PLoS ONE. 2011;6: e26969. doi:10.1371/journal.pone.0026969

97. Linsen SEV, Cuppen E. Methods for small RNA preparation for digital gene expression profiling by next-generation sequencing. Methods Mol Biol. 2012;822: 205–217. doi:10.1007/978-1-61779-427-8_14

98. Willenbrock H, Salomon J, Sokilde R, Barken KB, Hansen TN, Nielsen FC, et al. Quantitative miRNA expression analysis: Comparing microarrays with next-generation sequencing. RNA. 2009;15: 2028–2034. doi:10.1261/rna.1699809

99. Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, et al. Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence. 2012;3: 4. doi:10.1186/1758-907X-3-4

100. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.

101. Prokopenko I, McCarthy MI, Lindgren CM. Type 2 diabetes: new genes, new understanding. Trends Genet. 2008;24: 613–621. doi:10.1016/j.tig.2008.09.004

102. McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. Curr Diab Rep. 2009;9: 164–171.

103. Bagge A, Clausen TR, Larsen S, Ladefoged M, Rosenstierne MW, Larsen L, et al. MicroRNA-29a is up-regulated in beta-cells by glucose and decreases glucose-stimulated insulin secretion. Biochem Biophys Res Commun. 2012. doi:10.1016/j.bbrc.2012.08.082

104. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012;44: 981–990. doi:10.1038/ng.2383

105. Higa GM, Fell RG. Sex hormone receptor repertoire in breast cancer. Int J Breast Cancer. 2013;2013: 284036–14. doi:10.1155/2013/284036

106. Gaube F, Wolfl S, Pusch L, Kroll TC, Hamburger M. Gene expression profiling reveals effects of Cimicifuga racemosa (L.) NUTT. (black cohosh) on the estrogen receptor positive human breast cancer cell line MCF-7. BMC Pharmacol. 2007;7: 11. doi:10.1186/1471-2210-7-11

107. Rae JM, Johnson MD, Scheys JO, Cordero KE, Larios JM, Lippman ME. GREB 1 is a critical regulator of hormone dependent breast cancer growth. Breast Cancer Res Treat. 2005;92: 141–149. doi:10.1007/s10549-005-1483-4

108. Frasor J, Danes JM, Komm B, Chang KCN, Lyttle CR, Katzenellenbogen BS. Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. Endocrinology. 2003;144: 4562–4574. doi:10.1210/en.2003-0567

109. Bourdeau V, Deschênes J, Laperrière D, Aid M, White JH, Mader S. Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells. Nucleic Acids Research. 2008;36: 76–93. doi:10.1093/nar/gkm945

110. Castellano L, Giamas G, Jacob J, Coombes RC, Lucchesi W, Thiruchelvam P, et al. The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response. Proc Natl Acad Sci USA. 2009;106: 15732–15737. doi:10.1073/pnas.0906947106

111. Simonini P, Breiling A, Gupta N, Malekpour M. Epigenetically deregulated microRNA-375 is involved in a positive feedback loop with estrogen receptor α in breast cancer cells. Cancer Research. 2010;70: 9175–9184. doi:10.1158/0008-5472.CAN-10-1318

112. Casneuf T, Van de Peer Y, Huber W. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. BMC Bioinformatics. 2007;8: 461. doi:10.1186/1471-2105-8-461

113. Shang Y, Hu X, DiRenzo J, Lazar MA, Brown M. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. Cell. 2000;103: 843–852.

114. Wiese TE, Kral LG, Dennis KE, Butler WB, Brooks SC. Optimization of estrogen growth response in MCF-7 cells. In Vitro Cell Dev Biol. 1992;28A: 595–602.

115. Jensen TW, Ray T, Wang J, Li X, Naritoku WY, Han B, et al. Diagnosis of Basal-Like Breast Cancer Using a FOXC1-Based Assay. J Natl Cancer Inst. 2015;107. doi:10.1093/jnci/djv148

116. Usary J, Llaca V, Karaca G, Presswala S, Karaca M, He X, et al. Mutation of GATA3 in human breast tumors. Oncogene. 2004;23: 7669–7678. doi:10.1038/sj.onc.1207966

117. Cohen PA, Donini CF, Nguyen NT, Lincet H, Vendrell JA. The dark side of ZNF217, a key regulator of tumorigenesis with powerful biomarker value. Oncotarget. 2015;6: 41566–41581. doi:10.18632/oncotarget.5893

118. Frietze S, O'Geen H, Littlepage LE, Simion C, Sweeney CA, Farnham PJ, et al. Global analysis of ZNF217 chromatin occupancy in the breast cancer cell genome reveals an association with ERalpha. BMC Genomics. BioMed Central; 2014;15: 520. doi:10.1186/1471-2164-15-520

119. Suter R, Marcum JA. The molecular genetics of breast cancer and targeted therapy. Biologics. 2007;1: 241–258.

120. Smith CL, Migliaccio I, Chaubal V, Wu M-F, Pace MC, Hartmaier R, et al. Elevated nuclear expression of the SMRT corepressor in breast cancer is associated with earlier tumor recurrence. Breast Cancer Res Treat. 2012;136: 253–265. doi:10.1007/s10549-012-2262-7

121. Eeckhoute J, Keeton EK, Lupien M, Krum SA, Carroll JS, Brown M. Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. Cancer Research. 2007;67: 6477–6483. doi:10.1158/0008-5472.CAN-07-0746

122. Tkocz D, Crawford NT, Buckley NE, Berry FB, Kennedy RD, Gorski JJ, et al. BRCA1 and GATA3 corepress FOXC1 to inhibit the pathogenesis of basal-like breast cancers. Oncogene. 2011;31: 3667–3678. doi:10.1038/onc.2011.531

123. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics. 2014;15: 79. doi:10.1186/1471-2105-15-79

124. Long J, Ou C, Xia H, Zhu Y, Liu D. MiR-503 inhibited cell proliferation of human breast cancer cells by suppressing CCND1 expression. Tumour Biol. 2015;36: 8697–8702. doi:10.1007/s13277-015-3623-8

125. Plaza-Menacho I, Mologni L, McDonald NQ. Mechanisms of RET signaling in cancer: Current and future implications for targeted therapy. Cellular Signalling. 2014;26: 1743–1752. doi:10.1016/j.cellsig.2014.03.032

126. Nguyen NT, Vendrell JA, Poulard C, Győrffy B, Goddard-Léon S, Bièche I, et al. A functional interplay between ZNF217 and Estrogen Receptor alpha exists in luminal breast cancers. Mol Oncol. 2014;8: 1441–1457. doi:10.1016/j.molonc.2014.05.013

127. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14: 128. doi:10.1186/1471-2105-14-128

128. Polioudakis D, Abell NS, Iyer VR. miR-503 represses human cell proliferation and directly targets the oncogene DDHD2 by non-canonical target pairing. BMC Genomics. 2015;16: 40. doi:10.1186/s12864-015-1279-9

129. Guo J, Liu X, Wang M. miR-503 suppresses tumor cell proliferation and metastasis by directly targeting RNF31 in prostate cancer. Biochem Biophys Res Commun. 2015;464: 1302–1308. doi:10.1016/j.bbrc.2015.07.127

130. Chong Y, Zhang J, Guo X, Li G, Zhang S, Li C, et al. MicroRNA-503 acts as a tumor suppressor in osteosarcoma by targeting L1CAM. PLoS ONE. 2014;9: e114585. doi:10.1371/journal.pone.0114585

131. Li J, Song L, Qiu Y, Yin A, Zhong M. ZNF217 is associated with poor prognosis and enhances proliferation and metastasis in ovarian cancer. International Journal of Clinical and Experimental Pathology. 2014;7: 3038–3047.

132. Thollet A, Vendrell JA, Payen L, Ghayad SE, Ben Larbi S, Grisard E, et al. ZNF217 confers resistance to the pro-apoptotic signals of paclitaxel and aberrant expression of Aurora-A in breast cancer cells. Mol Cancer. 2010;9: 291. doi:10.1186/1476-4598-9-291

133. Li B, Liu L, Li X, Wu L. miR-503 suppresses metastasis of hepatocellular carcinoma cell by targeting PRMT1. Biochem Biophys Res Commun. 2015;464: 982–987. doi:10.1016/j.bbrc.2015.06.169

134. Yin Z-L, Wang Y-L, Ge S-F, Guo T-T, Wang L, Zheng X-M, et al. Reduced expression of miR-503 is associated with poor prognosis in cervical cancer. Eur Rev Med Pharmacol Sci. 2015;19: 4081–4085.

135. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17: pp. 10–12.

136. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research. 2010;38: e178. doi:10.1093/nar/gkq622

137. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12: 323. doi:10.1186/1471-2105-12-323

138. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

139. Fritsch FN, Carlson RE. Monotone Piecewise Cubic Interpolation. SIAM J Numer Anal. 1980;17: 238–246. doi:10.1137/0717021

140. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. IEEE; 1989;11: 674–693. doi:10.1109/34.192463

141. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8: 1551–1566. doi:10.1038/nprot.2013.092

142. Fang Y, Fullwood MJ. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. Genomics Proteomics Bioinformatics. 2016;14: 42–54. doi:10.1016/j.gpb.2015.09.006

143. Galupa R, Heard E. X-chromosome inactivation: new insights into cis and trans regulation. Curr Opin Genet Dev. 2015;31: 57–66. doi:10.1016/j.gde.2015.04.002

144. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. Science. 2008;320: 1643–1647. doi:10.1126/science.1155390

145. Ulitsky I, Shkumatava A, Jan C, Subtelny AO, Koppstein D, Bell G, et al. Extensive alternative polyadenylation during zebrafish development. Genome Research. 2012. doi:10.1101/gr.139733.112

146. Mayr C, Bartel DP. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009. doi:10.1016/j.cell.2009.06.016

147. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Comms. Nature Publishing Group; 2014;5: 5274. doi:10.1038/ncomms6274

148. Lembo A, Di Cunto F, Provero P. Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. Li J, editor. PLoS ONE. 2012;7: e31129. doi:10.1371/journal.pone.0031129

149. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014;157: 1262–1278. doi:10.1016/j.cell.2014.05.010