Role of Transcription Factor-MicroRNA Feedback Circuits in the Canalization of Human Regulatory Networks

Daniel R. Chee

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

John Stamatoyannopoulos, Chair

Christine Queitsch

Michael Maccoss

Program Authorized to Offer Degree

Genome Sciences

University of Washington

Abstract

Role of Transcription Factor-MicroRNA Feedback Circuits in the Canalization of Human Regulatory Networks

Daniel R. Chee

Chair of the Supervisory Committee:
Professor John Stamatoyannopoulos
Department of Genome Sciences

Complex interactions between hundreds of transcription factors underlie the gene expression profiles that give rise to cellular form and function. However, it is still not entirely understood how organisms faithfully recapitulate complex phenotypes through differentiation and development. Here I discuss the role MicroRNAs play in the robustness of human gene regulatory networks. Specifically, I find that TF-miRNA negative feedback circuits are significantly enriched in networks across 156 diverse cell types and that the motifs occur preferentially with highly connected TFs that drive local network architecture. My work also implicates TF-miRNA circuits in the reinforcement of cell fate decisions, and expression changes observed during differentiation. Further to the aforementioned, a second aspect of my work has focused on approaches for targeted manipulation and perturbation of regulatory elements and networks. Although genome editing has now become commonplace, tools and approaches for precisely manipulating gene expression and regulatory networks are still lacking. Here I describe development of novel computational tools for the efficient design and synthesis of potent and specific Transcriptional Activator-Like Effectors (TALEs) and TALE nucleases (TALENs).

Acknowledgements

Over the past several years I have had the opportunity to work with many wonderful people in the UW Genome Science Department and at the Altius Institute. First and foremost, I would like to thank my mentor, John Stamatoyannopoulos. Graduate school has been quite an "experience", and it would have been much more difficult if not for his support, keen academic insights, and sheer enthusiasm for science.

I would also like to thank the many colleagues in my lab who have contributed to my work. Particularly Alister Funnell and Kyle Siebenthall have always been there for me to discuss science over coffee. John Lazar, as both a fellow graduate student and also a statistics mentor. Thanks to Ram Akilesh and Fyodor Urnov for being great mentors. Ericka Otterman and Eunice Choi for being an experimental resource while developing code for automated TALEN assembly/design. Essentially everyone else at Altius for providing such a great working environment.

I would also like to thank my thesis committee, Christine Queitsch, Mike Maccoss, Daniela Witten, and Elhanan Borenstein, for all of their guidance and support. I want to particularly acknowledge the Genome Science summer REU program, without which I would have never even thought to attend grad school in Seattle. Thanks, Brian Giebel, for being such a wonderful resource for the past 10 years.

Thanks to my siblings Christine Chee, Thomas Chee, and Eugene Chee for always having my back and special thanks to my mom, Patricia Chee, for always being my number one fan. Special thanks to my partner, Cassie Bryan, for all of your loving support and I can't wait to see what the future holds for us.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Background on Gene Regulation

Every cell in the human body contains the same genetic code, but through differentiation and development mature cells and tissues have drastically differing phenotypes and serve different functions. Even in a single-celled organism like bacteria, a seemingly static genome can still respond to environmental perturbations and stimuli (Covert et al. 2004; Martínez-Antonio, Janga, and Thieffry 2008). Complex organisms show robust development despite external and internal perturbations in a process Waddington referred to as Canalization (WADDINGTON 1942). The drivers behind development are, of course, Transcription Factors (TFs). TFs control when and where genes are expressed and how cells respond to signals from other cells and the environment. From the very beginning, growth of an organism from a fertilized egg is kicked off by TFs, usually received from the mother (Ptashne 2013). Needless to say, a great deal of effort has gone into understanding TFs and how they function.

On the most basic level, TFs consist of two core domains; a DNA binding domain (DBD) (PTASHNE 1967) and an activation domain (DOUGLAS and HAWTHORNE 1964). The DBD of a TF protein binds the promoter region of a gene and recruits RNA polymerase which activates transcription. This process, known as regulated recruitment, is one of the fundamental principles of gene expression and is one of the most common models of gene activation. This, however, is an oversimplification, as not all TFs activate gene expression. In fact, the very first models of regulated gene expression were that of repression (Ptashne 2005). In the early 1960s Jacob and Monad described the workings of the lac operon and the phage lambda genes, two

systems “switched off” until presence/detection of the right environmental stimuli, lactose in the case of the lac operon and UV radiation in the case of the phage lambda genes. However, it turns out that most genes, even in the absence of a repressor, are only expressed at basal levels and the presence of an activator can dramatically increase expression (Ptashne 2005).

The process of transcription is complicated in eukaryotes by the presence of nucleosomes, the DNA-histone octamer complex by which DNA is condensed into a higher structure known as chromatin (Luger et al. 1997; Kornberg 1977). Although chromatin is important for DNA packaging purposes, it presents a significant barrier for transcription that is overcome by regulated recruitment. In this eukaryotic context, gene activation primarily occurs through the cooperative action of multiple TFs, mediator, and chromatin modifying complexes (Myers and Kornberg 2000; Malik and Roeder 2000; Cosma, Tanaka, and Nasmyth 1999) to evict nucleosomes from the gene promoter so transcription can occur (Gross and Garrard 1988). Despite this added complexity, cell-type specific transcriptional programs are still initiated by a class of proteins known as pioneer factors, which can interact with nucleosome bound DNA and recruit other TFs and chromatin modifying factors (Cirillo et al. 1998; Bossard and Zaret 1998; Laganière et al. 2005). TFs are so instrumental to shaping a cell’s transcriptional landscape, that the addition of specific sets of TFs can reprogram the transcriptional landscape to that of another cell type (Graf and Enver 2009; Takahashi and Yamanaka 2006; Davis, Weintraub, and Lassar 1987).

1.2 Gene Regulatory Networks

Gene regulatory networks describe the connections between TFs and the genes that they regulate (Lee et al. 2002). Dissecting regulatory networks for key principles that dictate gene

expression and how cells respond to the environment is one of the fundamental questions in systems biology (Thompson, Regev, and Roy 2015). Traditionally, the construction of regulatory networks has been done by collecting information from individual experiments, focusing on a single gene of interest, and it often takes years to generate a network/subnetwork of moderate size (Davidson et al. 2002; Yun and Wold 1996). More recently, novel genome-wide techniques have been developed and applied, each with their own advantages and disadvantages. I classify these techniques into two broad classes; network discovery via direct binding assays and network inference from gene expression data. Here I describe a collection of these methodologies and consider the pros and cons associated with them.

1.2.1 Inference Methods

One common form of network construction is to infer regulatory interactions from gene expression data, essentially leveraging the fact that interacting genes will have a higher expression correlation across a wide range of cellular conditions. This type of analysis has been successfully applied to reconstruct networks in various different organisms and conditions (Basso et al. 2005; Carro et al. 2010). To this effect, these methods are applied to datasets constructed from numerous, sometimes in the hundreds, samples or perturbations of a given system (Kemmeren et al. 2014; Amit et al. 2009; Walker et al. 2007). Different network inference algorithms apply various computational models and heuristics, but generally suffer from several major drawbacks. Firstly, these methods are computationally and statistically complex, needing to query hundreds to thousands of potential regulators for every gene, resulting in long computation time and multiple hypothesis testing (Dunn et al. 2014; Greenfield, Hafemeister, and Bonneau 2013; Friedman 2004). Secondly, these experiments lack direct evidence for interactions, resulting in false positives caused by indirect correlations. Finally,

reliable correlations require expression values from a large number of samples or perturbations per cell type which could take months or even years to collect.

The past ten years has seen the development of an abundance of methods for single cell sequencing and its application to RNAseq (Klein et al. 2015; Macosko et al. 2015; Ramsköld et al. 2012; Islam et al. 2011; Tang et al. 2009). Single cell RNAseq (scRNAseq) has great promise for the generation of gene regulatory networks. Instead of sequencing many samples over varying conditions, scRNAseq leverages the natural variation in gene expression to infer regulatory relationships, allowing expression to be measured in a more native context. However, one important consideration with scRNAseq is that transcription on a single cell level is stochastic, and "bursty" (Molin and Camillo 2018), and it is often difficult to distinguish between natural variation and noise, leading to spurious correlations. Many methodologies developed for bulk RNAseq don't work with scRNAseq data (Stegle, Teichmann, and Marioni 2015) and new methods need to be developed with single cell dynamics in mind.

1.2.2 Network Discovery via Direct Binding Assays

Chromatin Immunoprecipitation when combined with next generation sequencing (ChIP-seq) allows the discovery of transcription factor binding sites genome wide (Robertson et al. 2007; Harbison et al. 2004). This technique functions by cross linking DNA-binding proteins to DNA *in vivo* and then using an antibody to purify any protein complexes containing the TF of interest. The DNA is then separated from the complex and sequenced to obtain a direct readout of associated DNA sequences. Though this method allows for the simultaneous identification of every binding site for a given TF across the entire genome, it is limited by the availability of reliable antibodies, the detection of spurious binding sites from indirect binding , and by the fact that separate experiments must be done for every TF in every cell type (~ 1700 human TFs

(Vaquerizas et al. 2009)). ChIP-seq is useful for identifying high-quality binding sites for individual TFs and is often used as gold standard validation, however it is not currently suitable for the construction of genome wide regulatory networks across multiple cellular contexts.

The yeast one-hybrid assay has been implicated as a high-throughput method of regulatory network reconstruction (Reece-Hoyes et al. 2011; Walhout 2006). This method is conducted by creating a yeast plasmid containing a DNA sequence of interest, in this case the promoter region of human protein coding genes, upstream of a reporter gene and transfecting it into yeast. A fusion/hybrid protein is also introduced into the yeast consisting of the DNA-binding domain of the TFs of interest and a yeast activating domain. If the TF of interest binds the given DNA sequence, the reporter gene is expressed. Modifications of this protocol allow it to be easily applied to pairwise combinations of hundreds to thousands of DNA-binding proteins and promoters sequences. Though this method can be applied in a high-throughput fashion, it cannot determine whether a given TF is an activator or a repressor. Also, since these experiments are conducted in yeast, they lack native cellular context and do not capture interactions between multiple TFs in the same promoter region.

DNaseI is an enzyme that has been used to study regulatory DNA for over forty years. When a nucleus is treated with DNaseI, areas of accessible DNA are preferentially cleaved, producing so-called DNaseI hypersensitive sites (DHSs). Combining DNaseI cleavage with next generation sequencing yields a genome-wide accessibility profile. Cleavage within DHSs is not uniform, instead regions of DNA currently being bound by TFs and other DNA binding proteins are protected. This process, known as DNase footprinting (Galas and Schmitz 1978), has been used since the 70s to reliably identify regulatory elements and was implicated in the discovery of the first sequence specific transcription factor (Dynan and Tjian 1983). By searching within gene

promoters for footprints containing known TF binding motifs, a genome-wide regulatory network can be derived for a cell type from a single experiment. This method is ideal for analyzing binding dynamics across many cell types, but accurate calling of footprints requires expensive, high coverage sequencing data and successful identification of the binding TF requires accurate TF motif models.

For the context of this project, I use gene regulatory networks derived from DNaseI data because I am particularly interested in the binding dynamics of specific TFs across gene promoters of many different cell types and conditions. Using DNaseI sequencing data allows me to leverage the full diversity of the data collected for the ENCODE project (Consortium 2004) and this methodology has been shown to recapitulate know regulatory interactions (Neph et al. 2012).

Chapter 2: TF-miRNA Feedback Circuits and their potential role in the canalization of human regulatory networks

2.1 Introduction

The interactions of sequence-specific TFs with protein coding genes underlie the defining features of cellular identity, differentiation, and development (Iwafuchi-Doi and Zaret 2016). Of particular interest are the network interactions where a given TF regulates another TF. These core TF networks are believed to form the foundation of cell-type-specific gene expression, phenotype, and function. However, TFs do not operate in isolation, instead they work in tandem with non-coding RNAs, signaling molecules, and other regulatory machinery to regulate gene expression profiles. Specifically, microRNAs (miRNAs) inhibit gene expression of target

mRNAs in a sequence-specific fashion. miRNAs interact heavily with TFs and it is widely accepted that they are also implicated in cellular identity, differentiation, and development (Lin et al. 2015).

miRNAs are small (~22 nucleotide) non-coding RNAs that post-transcriptionally inhibit gene expression (T. Lee et al. 2002). miRNAs are transcribed as longer transcripts known as primary miRNAs (primiRNAs) that contain the sequence for one or miRNA stem loops. These primiRNAs are processed, first by Drosha into short hairpin molecules in the nucleus, and then subsequently by Dicer in the cytoplasm into mature miRNAs that incorporate into the RNA-Induced Silencing Complex (RISC) to silence mRNA translation (Bartel 2004). miRNAs target gene transcripts by forming by forming Watson-Crick base pairing in their seed region (nucleotides 2-8) with the complementary sites in the 3' UTRs of specific mRNAs (Bartel 2009). Since this targeting happens in such a predictable fashion, many algorithms have been developed to predict miRNA target sites (Betel et al. 2008; Lewis, Burge, and Bartel 2005). miRNA represent a major class of regulatory factors as it is predicted that roughly 60% of

protein coding genes are targeted by miRNAs (Dunn et al. 2014; Greenfield, Hafemeister, and Bonneau 2013; N. Friedman 2004).

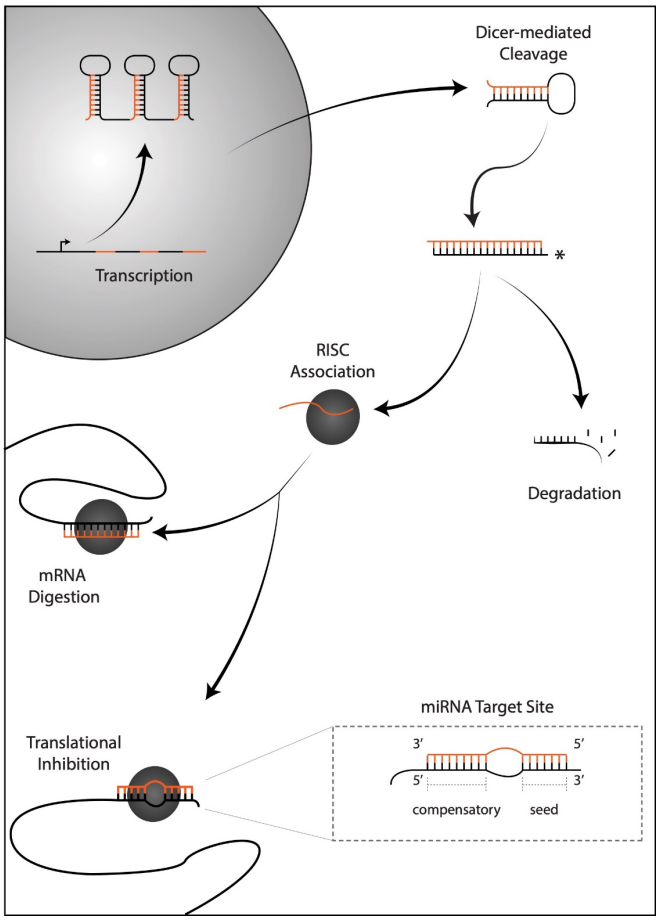

Figure 2.1 miRNA Biogenesis

miRNAs are transcribed either as introns of protein coding genes or as polycistron clusters. They are processed into hairpins by Drosha, and then exported into the cytosplasm where they are further processed by Dicer into two complementary mature miRNAs. One of the miRNAs is loaded into the RISC complex where it is used for targeted silencing, either by translational inhibition or mRNA degradation.

Due to their repressive nature, it has been hypothesized that miRNAs function primarily to reduce noise and enforce cell fate decisions (Hornstein and Shomron 2006). Like other sequence-specific regulatory factors, such as TFs, miRNAs have the potential to regulate many

different genes at once (Stark et al. 2005), allowing even a single miRNA to influence entire transcriptional programs. Deletion of DGCR8, a key protein in the miRNA biogenesis pathway, leaves Embryonic Stems cells in mice unable to properly downregulate pluripotency genes during differentiation (Wang et al. 2007), and aberration of dicer leads to many different developmental abnormalities (Graf and Enver 2009; Takahashi and Yamanaka 2006; R. L. Davis, Weintraub, and Lassar 1987).

Several studies have demonstrated that miRNAs canalize development by contributing to robust gene expression leading to faithful recapitulation of animal body plans (Sokol and Ambros 2005; Li et al. 2006; Cassidy et al. 2013). When a miRNA is expressed in the same system (and at similar levels) as its target genes, that miRNA serves as a tuning mechanism to reduce variations in the expression level of its target genes (Mukherji et al. 2011). Conversely, when a miRNA and its target genes are in different cell types/tissues then that miRNA serves a fail-safe mechanism to ensure that proteins are not expressed in the wrong cellular context (Moss, Lee, and Ambros 1997; Wightman, Ha, and Ruvkun 1993). In the former scenario, miRNAs ensure phenotypic stability in the face of transcriptional noise, and in the latter, they re-enforce cell fate decisions.

A prominent feature of complex networks is the recurrence of subnetworks known as network motifs (Milo et al. 2002). Our group has shown that Transcription Factor (TF) regulatory networks across a diverse set of tissues/cell types are enriched for a common set of these network motifs (Neph et al. 2012). Motifs involving the interactions between TFs and miRNAs, such as Feedforward loops (FFLs) and feedback circuits (FBCs), have been shown to be prominent features of regulatory networks in metazoans (Martinez et al. 2008; Tsang, Zhu,

and van Oudenaarden 2007). FBCs in particular have been implicated in many biological processes; the epithelial-mesenchymal transition (Bracken et al. 2008; Gregory et al. 2008), muscle development (Y. Wang et al. 2007), brain development (Johnston et al. 2005; Johnston and Hobert 2005), and heart development (Niu et al. 2007). Thus, understanding the interplay of TFs and miRNAs in the context of feedback circuits presents an interesting challenge.

2.2 Results

2.2.1 Identification of miRNA Promoters

I annotated a novel set of miRNA promoters by pairing transcripts from three different sources with a comprehensive set of non-overlapping DNase Hypersensitive Sites (DHS) derived in house using data from the ENCODE project. First, I took transcription start sites (TSSs) from the Integrated miRNA Expression Atlas (Rie et al. 2017) and merged them with TSSs derived from transcripts in GENCODE v26 and those derived from Drosha dominant-negative knockdown experiments (Chang et al. 2015) that overlapped miRNA genes in miRbase v22 on the same strand. Next, I identified the closest DHS from the ENCODE derived master list within 1000bp to each TSS. These DHS became the putative core promoters, and transcripts with the same core promoter were assumed to be under the same transcriptional control. Using this process, I identified 4,101 unique putative promoters for 2,200 of the 2,652 miRNA genes annotated in miRbase. Figure 2.2 describes this process in detail and as seen in panel **B** combining these three data sources gives us a much more comprehensive view of the total miRNA promoter landscape.

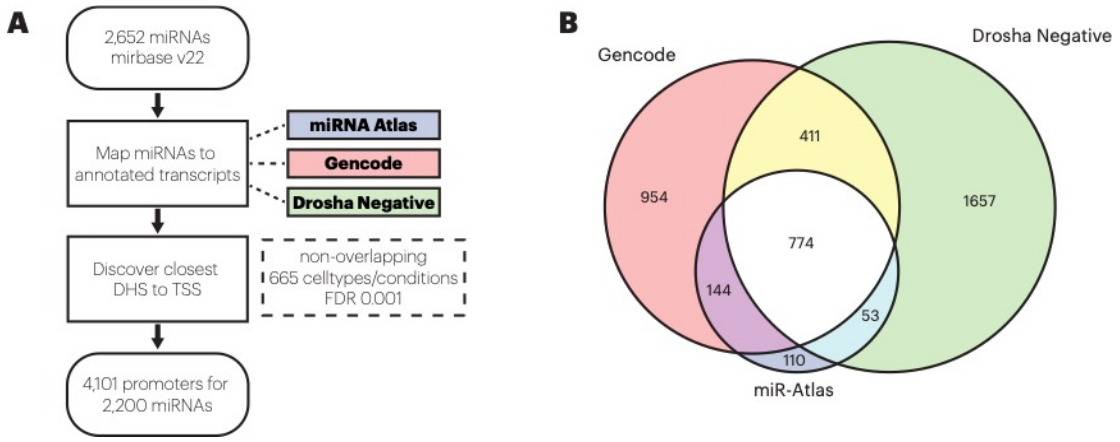

Figure 2.2 miRNA Promoters

A. The process of miRNA promoter annotation. Transcripts that overlapped with miRbase miRNA genes were integrated from three different sources; integrated miRNA atlas, gencode, and drosha dominant negative experiments. The TSS of these transcripts were mapped to the closest DHS in our ENCODE derived master list. This yielded promoters for a large number of miRNA genes in miRbase.
B. The overlap of promoters derived from the different source transcripts. Combining data from these three sources yields far more candidate promoters than any one of these sources alone.

2.2.2 MicroRNA Promoter activity reconstructs developmental lineages

miRNAs are known to act as key regulators in early embryonic development by facilitating robust gene expression programs during differentiation (Ivey and vastava 2010). Therefore, I expect the activity of miRNA promoters to reflect the taxonomy of human tissues during development. To assess the accuracy of our promoter annotations, I constructed an activity matrix for 32 fetal samples consisting of 15 distinct organs, collected and sequenced for the ROADMAP project, over the 4,101 candidate miRNA promoters. This binary matrix reflected the chromatin accessibility state, “accessible” or “not-accessible”, for each promoter across all samples. Hierarchical clustering performed on the samples yielded tight groups of physiologically similar tissues and clustering at the promoter level allow us to partition miRNA

promoters into four different classes (Figure 2.3). The first class consisted of promoters constitutively active across all samples. Another class are broadly active across many samples, but still have differential activity across tissues. A third class was active in only a select number of tissues, and the final class represents promoters active in a single tissue. Across all promoters, I observed tissue specific expression patterns of miRNAs, suggesting that the miRNA promoters at least reflect patterns of gene regulation specific to a given tissue.

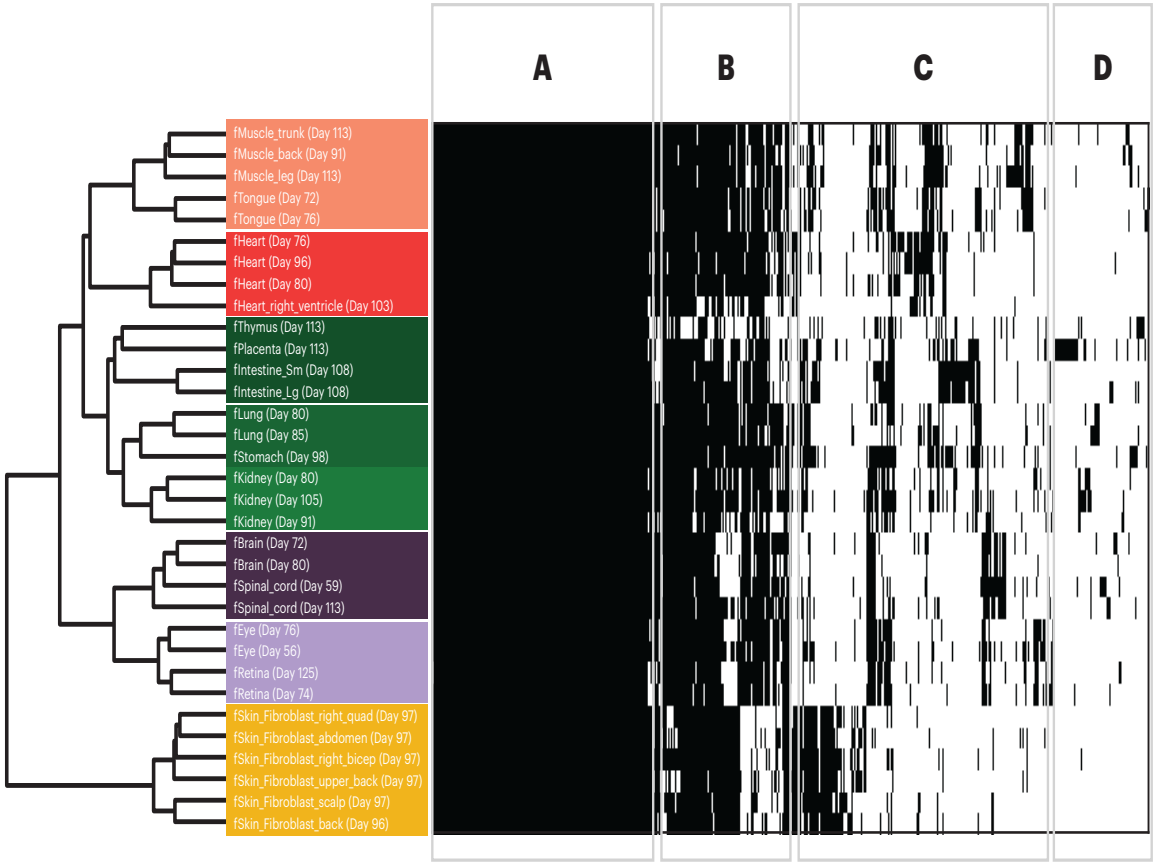

Figure 2.3 Promoter Clustering

Hierarchical clustering of binary promoter chromatin activity groups fetal samples into modules of similar taxonomy, suggesting miRNA promoter activity tracks with tissue through differentiation and development. Promoters have three classes:

A. *Constitutive* promoters are active across all samples
B. *Broadly Active* promoters are active across most samples, but have differential expression across tissue groups
C. *Tissue-Selective* are active in only a select group of tissues
D. *Tissue-Specific* specific to a single tissue

2.2.3 Comprehensive Mapping of TF-miRNA cross-regulatory networks

I constructed TF-miRNA cross regulatory networks by combining DNase I footprinting data from 156 samples across a diverse set of cell types and conditions with a curated set of

computationally predicted miRNA target from TargetScan. I used four well-annotated databases, Uniprobe, Selex, Transfac, and HOCOMOCO, to determine the identity of Transcription Factors occupying DNase Footprints within miRNA promoters. This process is similar to that done previously in our group to derive TF regulatory interactions and was shown to closely match ENCODE ChIP-seq for the same cognate factors (Stergachis et al. 2013). Repeating this procedure across our entire dataset, I mapped 847,108 unique regulatory interactions involving 741 TFs and 3,671 promoters for 2,164 miRNA genes. Networks for individual tissues/conditions contained an average of 158,749 edges.

To model miRNA-TF interactions I paired computationally predicted miRNA targets (Figure 2.4 Cross-regulatory Network Construction) with experimentally derived targets. I required experimentally derived targets to also have a computationally predicted canonical target site (conserved or non-conserved), which is aimed at overcoming the false positives associated with both types of miRNA targets. This yielded 13,037 miRNA-TF edges consisting of 1,933 mature miRNAs and 555 TFs. miRNAs target on average 7 TFs. These two types of edges are paired to form a bipartite representation of the interactions between TFs and miRNAs from which I draw inferences about network form and function. Figure 2.4 gives a visual schematic for network construction.

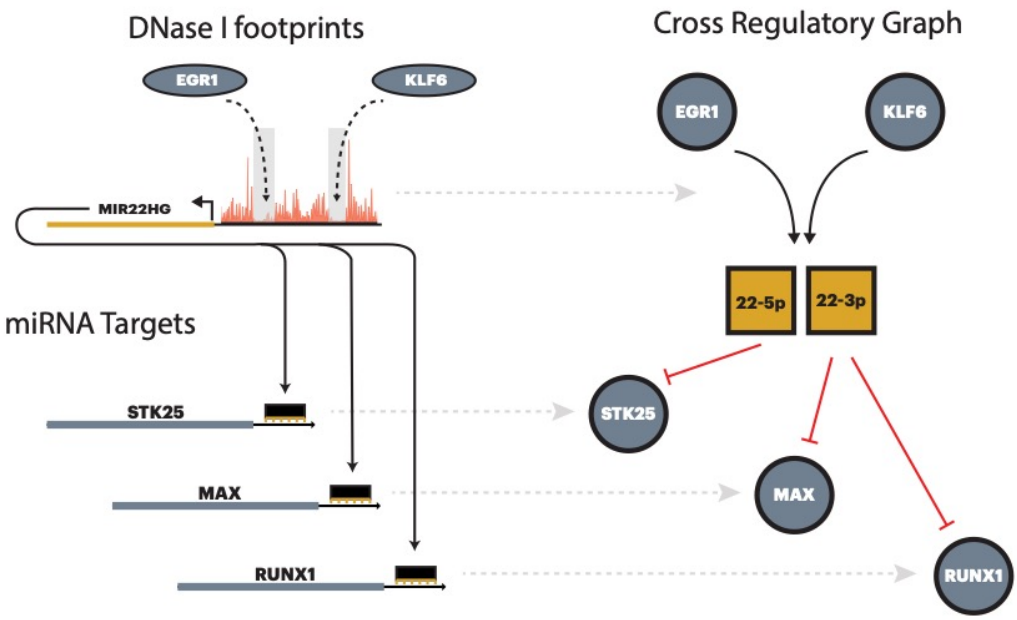

Figure 2.4 Cross-regulatory Network Construction

A. TF-miRNA edges are inferred by the presence of TF binding sites in footprints found in miRNA promoters.
B. miRNA-TF edges are derived from a TargetScan computationally predicted conserved and nonconserved miRNA targets that show degree of experimental validation in miRTarbase.

2.2.4 TF-miRNA feedback circuits are a significant motif in across all networks

Network motifs are subgraphs that repeat themselves in a specific network or across many networks of the same type and are a common way of looking at similarities or finding important substructures of biological networks (Milo et al. 2002). Previous work in *C. elegans* has identified negative feedback circuits (visualized in the context of our data in Figure 2.5), as being a prominent motif in cellular networks (Martinez et al. 2008), so I sought out to identify the abundance of this motif in our human derived networks. This motif is incredibly abundant, occurring an average of 3,125 times per tissue/conditions. However, even though this motif occurs thousands of times within a given condition, there is only a limited number of TFs that

tend to participate in circuits (Figure 2.6) suggesting that there may be some feature of certain TFs that underlie their participation in these feedback circuits. This notion is explored in more detail in subsequent sections.

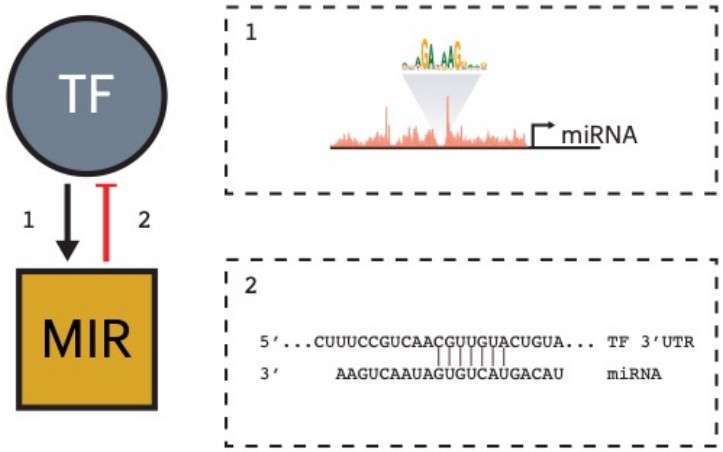

Figure 2.5 Negative Feedback Circuit

Schematic of a negative feedback circuit where:

1. miRNA promoter contains a footprinted binding site for a TF
2. TF mRNA contains a target site for that miRNA

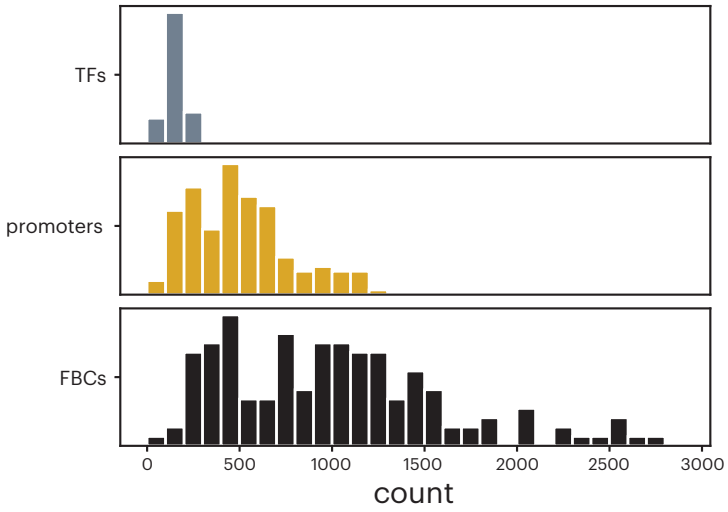

Figure 2.6 Circuit abundance across tissues/conditions

Feedback circuits are highly abundant across all cell types and conditions. Circuits occur with a limited set of TFs per cell type.

TFs and miRNAs tend to target many different genes. Therefore, the observed abundance of feedback circuits could have arisen simply by random chance. This can be address by assessing the significance of a given circuit count using the hypergeometric distribution. The number of possible circuits are counted given the factors in a specific network (*i.e.* testing all present pairs of TFs and miRNAs) and the expected number of circuits in all tissues/conditions is calculated. Figure 2.7 shows the expected distribution of random circuits in a network derived for a sample of CD34+ cells. In this case, the observed count is significantly higher than that expected by random chance, and indeed across all networks the observed counts are significantly higher than expected by random chance (Figure 2.8). I conclude that FBCs do not occur simply because TFs and miRNAs regulate many targets, but instead I suggest that they are a prominent feature of network architecture.

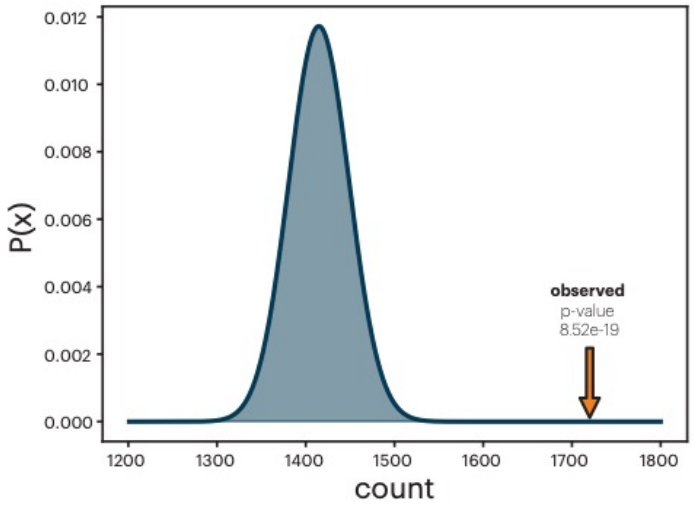

Figure 2.7 CD34+ FBC Significance

Plot of the probability mass function for the expected counts for CD34+ cells calculated with the TFs and miRNAs present in that network. The orange arrow indicates the observed counts, which is significantly higher than the expected count.

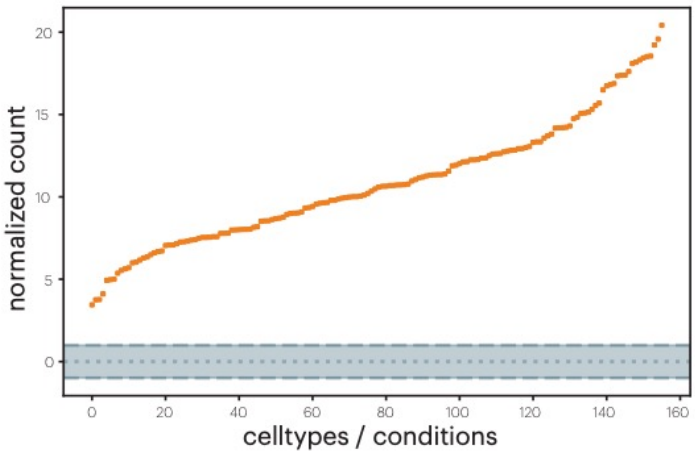

Figure 2.8 FBC Significance Across all Cell Types

Plot of the observed FBC count significance across all cell types and conditions. The light grey shaded region is the scaled PMF computed for every sample. The orange points are the observed counts in ascending order. The counts of every sample are highly significant to varying degrees.

2.2.5 Circuit TFs are highly connected and drive local network architecture

Our group has previously shown that the relative connectivity of master regulator genes is highest in their cognate cell type or lineage (Neph et al. 2012). It stands to reason that Normalized Network Degree (NND), a measure of relative connectivity, would be highest in TFs important to a particular cell type/condition. The NND is defined as $\frac{(\#\ in\ edges + \#\ out\ edges)}{(\#\ total\ edges)}$ and Figure 2.9 provides an example of the influence of network structure on the NND. Since only a limited number of TFs are found in FBCs, I wanted to compare the NNDs of circuit TFs to those of TFs that don't participate in circuits. I constructed paired TF-TF networks for each sample in our dataset and calculated the NND in the context of these networks. I found that across all samples that circuit TFs have a significantly higher NND in their specific tissue (figure 3b).

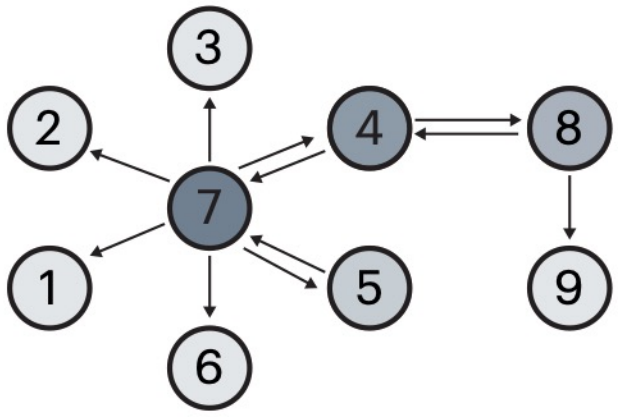

Figure 2.9 Normalized Network Degree

An example network arrangement demonstrating the normalized network degree. In this example. Node 7 has the highest NND and nodes are colored based on their corresponding NND.

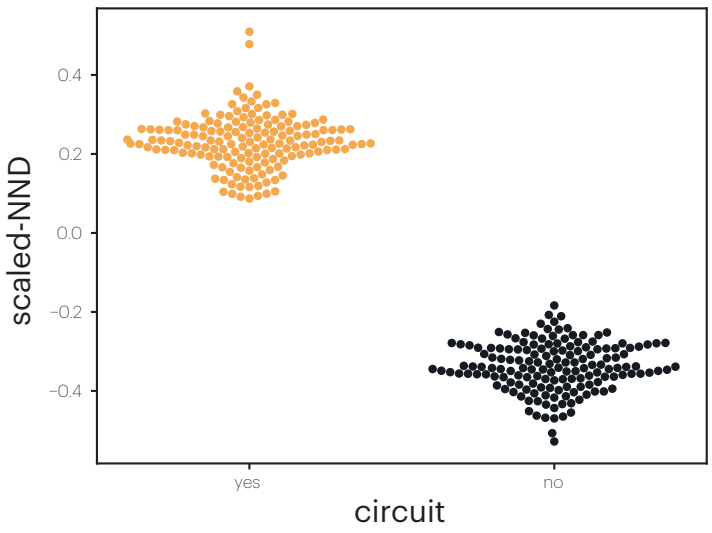

Figure 2.10 Circuit TF tissue specific NND

Swarmplot of the average NNDs of TFs across all cell types and conditions. Orange points are TFs that are in FBCs and black points are the NNDs of TFs not found in FBCs. FBCs preferentially occur with TFs that are highly connected.

Another way of assessing a TF's contribution to a given network is to calculate the frequency with which a given TF resides in the driver position in a localized network architecture. To do this analysis, I took the 13 distinct 3-node motifs described in the literature (Milo et al. 2002) and identified the ones with a clear driver (Figure 2.11 shows 2 different motifs, one with a clear driver and one without). I then calculated the ratio for which all TFs reside in the driver position within this set of informative motifs. I compared the average ratios of circuit TFs vs non-circuit TFs and found that the average ratios of circuit TFs was significantly higher than TFs not found in circuits (Figure 2.12). I conclude that circuit TFs are frequently found as the drivers in local network architecture.

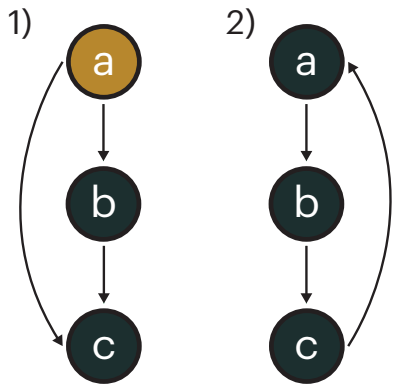

Figure 2.11 Driver Motifs

Example 3-node motifs. 1) has a clear driver seen in yellow, 2) is cyclical and therefore has no clear driver.

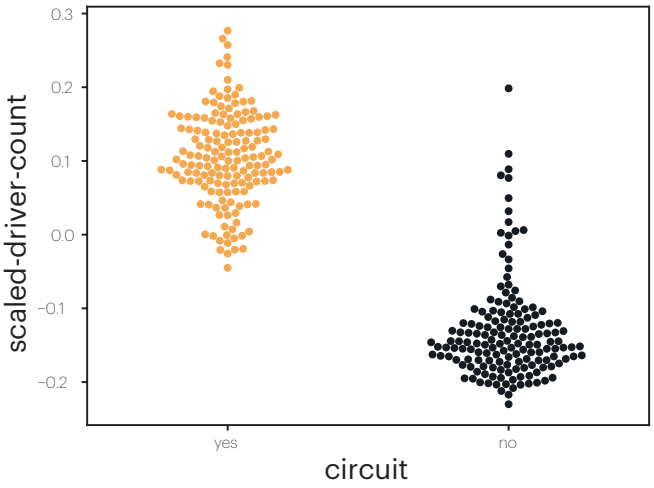

Figure 2.12 TF Driver Counts

Swarmplot of the average driver counts for TFs within a network. The orange points are the averages of TFs found in FBCs and the black points are the averages of those TFs not for in FBCs in a particular network. FBC TFs generally have a higher driver counts than those not found in FBCs.

2.2.6 FBC occurrence is consistent with a "weak-buffer" model of miRNA function

One interesting consideration of this type of network analysis is that the underlying process I use to determine where a transcription factor is binding does not take expression into consideration. For example, Figure 2.13 portrays a promoter primarily active in heart tissue samples. In our model, every member of the GATA family of transcription factors has a predicted binding site that is consistent with binding within the observed footprint. However, only two of these factors, GATA4 and GATA6, are expressed in heart. Therefore, only two of the three predicted FBCs (miR1-3p only targets three of the GATA factors) are consistent with gene expression.

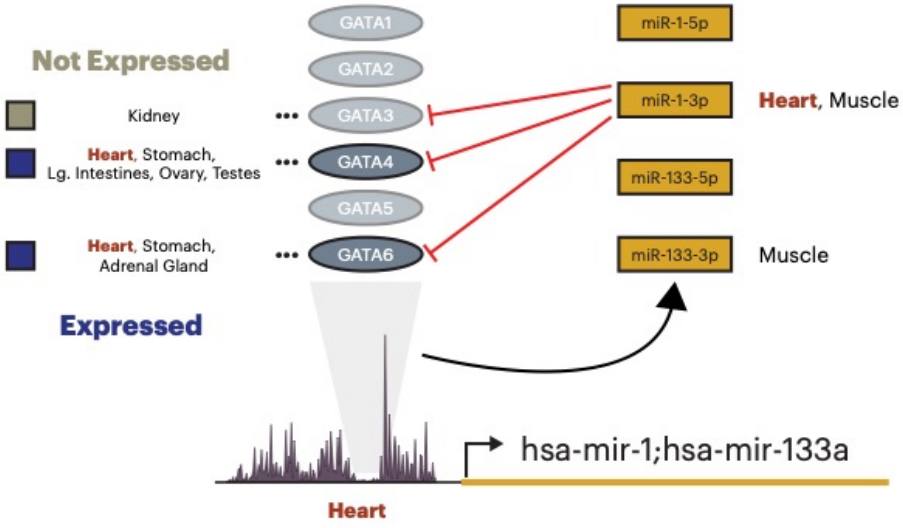

Figure 2.13 Context Specific FBCs

Example of context specific binding. The example promoter is a heart-specific miRNA promoter. Though all six of the GATA factors are predicted to bind via our network construction method, only three of those factors are targeted by the contained miRNAs and only two of those factors are expressed in heart cells. GATA3 suggests a mechanism where this miRNA prevents expression of GATA3 in heart and the expression of GATA4 and GATA6 is "fine-tuned".

To explore this further, I limit our analysis to a set of ten fetal tissues from our DNaseI dataset for which I have RNA-seq expression data in a similar tissue from the ROADMAP project. I wanted to look at the expression state of TFs in FBCs in these tissues, and thus I defined three states of expression; not expressed, expressed, and enriched. For a gene to be considered enriched, it had to have an expression level significantly higher in that tissue when compared to other samples (z-score >= 1.96). To be considered merely expressed it just has to have an average FPKM value of 1.0 for all ROADMAP samples for that tissue, otherwise it was categorized as not expressed. I found that across all fetal tissues most of the TFs involved in FBCs are classified as "expressed" (Figure 2.14). It has been argued that one fundamental function of miRNAs is to act as a weak buffering system, both preventing leaky expression of factors from alternative lineages and to adjust the expression of a given gene to physiological levels (Mukherji et al. 2011), and FBCs in these tissues are consistent with these functions. The fact that FBCs still occur regularly with TFs not expressed in a tissue suggests that FBCs may serve as a mechanism for a transcription factor to contextually modulate its own expression, *i.e.* if the TF is detected in a tissue in which it is not supposed to be expressed, it activates its own repressor.

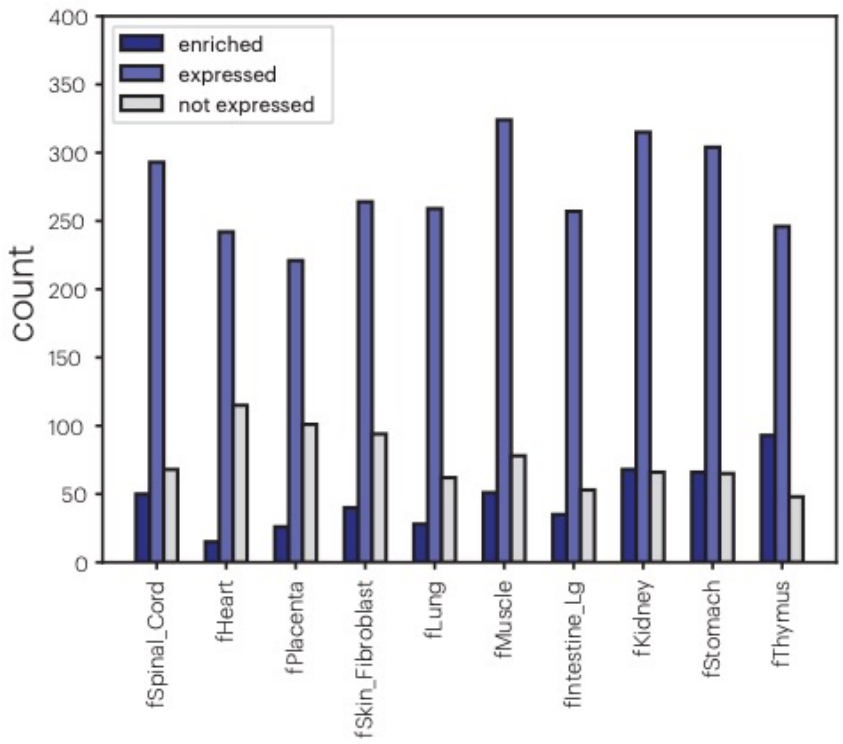

Figure 2.14 Tissue-wide Expression Context

The tissue-wide expression context of circuits in the fetal tissues from the ROADMAP expression atlas. Across all tissues, the lion's share of TFs is expressed and not enriched within a given tissue.

Because of the tissue-specific nature of many miRNA promoters, I hypothesize that FBCs involving TFs with expression that is enriched in alternative tissue lineages might occur primarily in tissue specific promoters. I tested this hypothesis by compiling a set of 644 promoters only active in samples from a single tissue within the fetal samples. Figure 2.15 juxtaposes the occurrence of a TF in an FBC with whether or not a TF is has enriched expression across tissue types. The enrichment of a TF's expression indeed occurs in tissues where the TF is not in an FBC, as seen in first panel of this figure. This suggests that promoters that become active during the development of a specific tissue (that is, promoters that are only active in that tissue) target TFs that are expressed at higher levels in alternative tissues. This fundamentally

provides a potential mechanism by which FBCs reinforce cell-fate decisions during differentiation by dampening TF expression only in specific tissues.

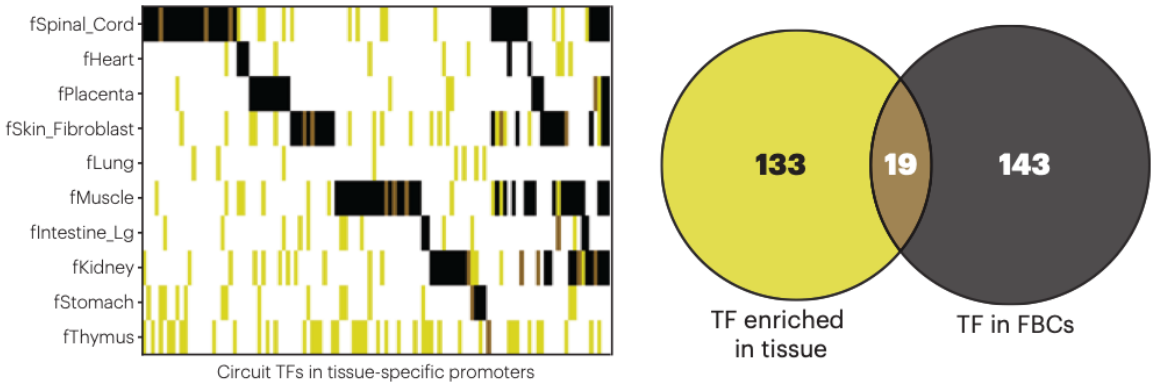

Figure 2.15 TFs in tissue specific Promoters

Comparison of TF occupancy in tissue specific promoters. Cells marked in yellow indicate the TF is enriched in that tissue, black TFs are circuited in that tissue, and cells marked in brown are both enriched and circuited in that tissue. FBCs in tissue specific promoters tend to occur with TFs enriched in other cell types.

However, I looked at genome wide FBCs consisting of TFs with expression enriched in alternative tissue lineages and found that the vast majority of these fate reinforcing FBCs occur within promoters that are active in multiple tissues (i.e. not within promoters that newly arise as a result of differentiation). I place each FBC (uniquely described as a TF, miRNA promoter combination) into one of the following four categories: Tissue-specific, tissue-selective, broadly active, and constitutive. Tissue-specific describing those FBCs that occur in only one tissue. Tissue-selective occurring in more than one tissue, but in no more than half. Broadly active occurring in most tissues, and constitutive occurring in every tissue. I find that unique FBCs usually occur in only a few tissues (Figure 2.16 **A**). By contrast, the promoters that this FBCs occur within are mostly constitutive or broadly active (Figure 2.16 **B,** promoters being classified

by the same criteria as previously stated for individual FBCs). This suggests that, as a general phenomenon, constitutive promoters are utilized in a tissue-specific manner to modulate the expression of TFs whose expression is enriched in alternative tissue lineages (thus also potentially reinforcing cell fate decisions in development).

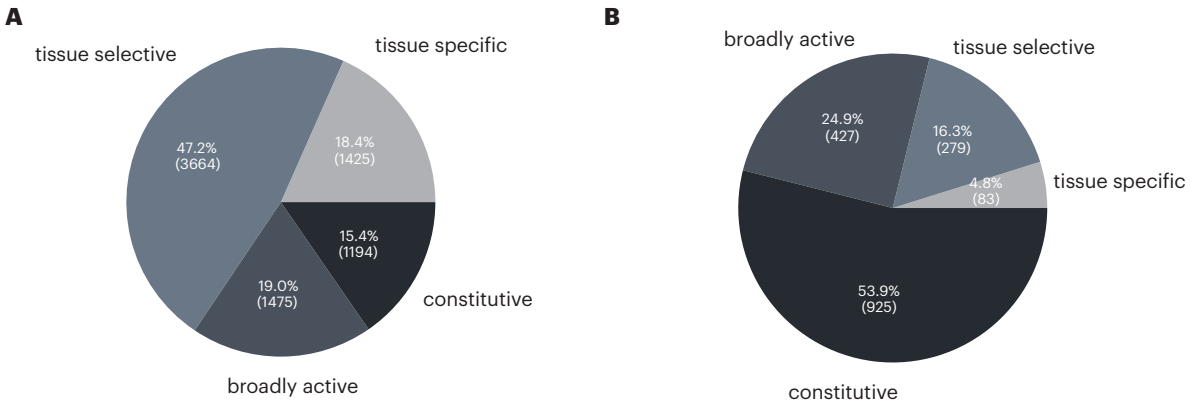

Figure 2.16 Cell fate reinforcing FBCs

A. Classification of fate reinforcing FBCs by the number of tissues in which they are present. Most fate reinforcing FBCs only occur in a few tissues.
B. Classification of the fate reinforcing FBC promoters. These promoters are active across a wide range of tissues.

2.2.7 FBCs reflect expression changes of key regulators during Hematopoietic differentiation

The Hematopoietic lineage is a good system for analyzing how regulatory networks change in response to differentiation. For this analysis, I paired RNAseq expression with DNase I sequencing for CD34+ (Hematopoietic progenitor) cells and several other derivative cells from the hematopoietic lineage, namely CD3+ (general T-cells), CD4+ (T-helper cells), CD8+ (Cytotoxic T-cells), CD56+ (Natural Killer Cells), CD14+ (Monocytes), and CD20+ (B-cells).

First, I sought to compare the landscape of accessible miRNA promoters across hematopoietic cells at different stages in development. Surprisingly, many more miRNA promoters are accessible in more terminally differentiation cell types (this is in contrast to the tendency overall of the genome of more differentiated cell types to be less accessible to nuclease cleavage (Stergachis et al. 2013)). Figure 2.17 summarizes the changes in promoter accessibility and FBC counts in each derivative cell type as compared to CD34+ cells. In general, relatively few promoters are lost in derivative cell types when compared to the number gained. As expected, the number of FBCs also increases in these cell types; the increases in accessible promoters providing more opportunities for FBCs to occur.

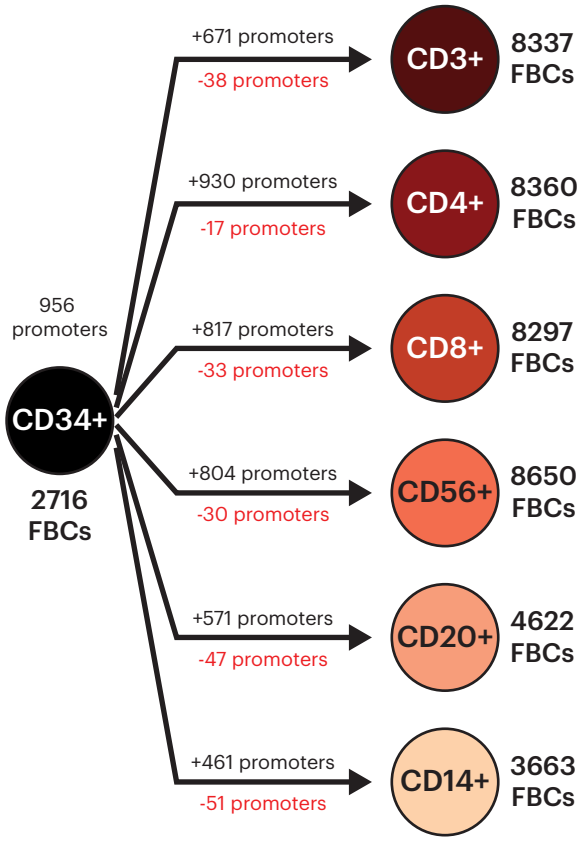

Figure 2.17 miRNA promoter activity in blood cells

Relative changes in the relative number of active promoters and FBCs observed in progenitor cells (CD34+) and derivative cells. A larger number of promoters are active in more differentiated cells.

I suggest that miRNAs function to reinforce cell fate decisions made during differentiation, and hematopoietic cells types are ideal to examine the interplay between FBCs and expression. I thus hypothesized that the number of FBCs involving TFs that are expressed highly in CD34+ would increase in derived cell types in which the expression of those genes is turned off. I looked at genes that met these expression criteria and found that those genes did occur in more FBCs in the derivative cell type. Figure 2.18 shows the genes that turn off through differentiation and also increase in circuit number. I also found that genes which turn on after

differentiation also increase in number of FBCs in which they occur (Figure 2.19). Though counterintuitive at first, this is consistent with the theory that FBCs serve to increase the "response time" of a given TF. Essentially the paired negative regulation allows a TF to be controlled by a more transcriptionally active promoter and reach its effective concentration more quickly than if it was controlled by a weaker promoter.

As a sanity check, I checked the expression levels of the mature miRNAs controlled by the promoters in each FBC as assayed by the NANOSTRING nCounter miRNA assay (Kulkarni 2011). The nCounter assay is primarily used for diagnostic purposes and is not ideal for studying miRNA regulation dynamics due to a limited probe set, but we still found cases where the TFs and miRNAs found in FBCs have dynamics consistent with a switching mechanism (TF goes from ON to OFF while the miRNA goes from OFF to ON (Figure 2.18) and vice versa (Figure 2.19)).

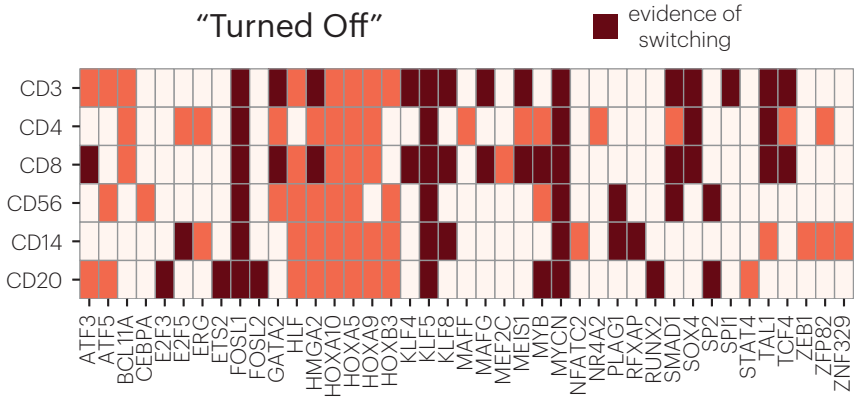

Figure 2.18 TFs turned off through differentiation

Genes shown in light red are expressed in CD34+ cells and are both, turned off and increase in FBC count in each derivative cell type. Genes shown in dark meet these criteria, but also show evidence of switching (FBC TF transitions from high to low expression, and the FBC miRNA transitions from low to high expression).

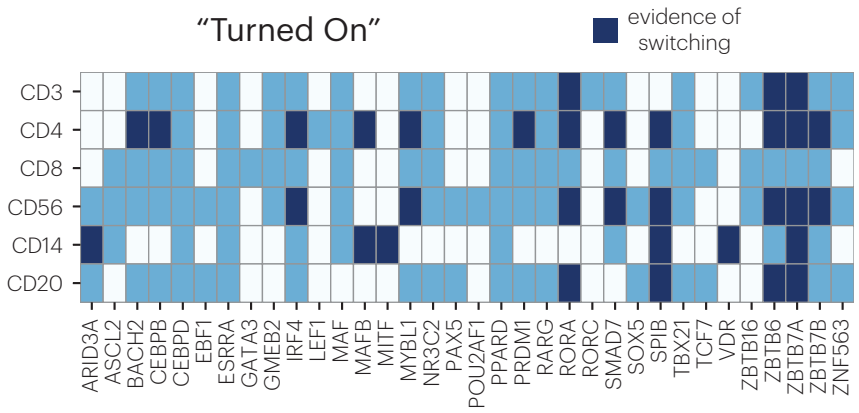

Figure 2.19 Genes turned on through differentiation

Genes shown in light blue are not expressed in CD34+ cells and are both turned on and increase in FBC count in each derivative cell type. Genes shown in dark blue meet these criteria and show evidence of switching opposite of that described for genes that turn off (FBC TF transitions from low to high expression, and the FBC miRNA transitions from high to low expression).

2.3 Summary

Here I presented evidence that TF-miRNA feedback circuits are an integral part of gene regulatory networks across differentiation and development. I have shown that miRNA promoters are highly cell-type-specific, and their activity reflects differences in tissue states during differentiation and development. TF-miRNA FBCs constitute a *bona fide* network motif that is enriched in all tissues and cell types, and FBCs occur preferentially with TFs that drive the regulatory landscape of a given tissue or cell type. These results suggest that these TF-miRNA circuits serve as a "centrifugal governor" for lynchpin TFs, dampening their expression in

circumstances where overexpression could drastically change the expression profile of a particular cell or tissue type.

I've also shown that the expression of FBC factors is consistent with current hypotheses of miRNA function. With the fetal data, I show that FBCs tend to occur with TFs that are expressed in a particular cell type, but not enriched, which is consistent with FBCs being a potential mechanism for a "weak buffer" system that adjusts the expression of a given TF to physiologically relevant levels. I also observe, in the case of hematopoietic cell types, that circuit counts are consistent with TFs being turned off during differentiation. Interestingly, circuit counts also increase with genes that turn on during differentiation. This is not necessarily unexpected, as TFs increase in expression, they bind more miRNA promoters, which in turn makes them more likely to form FBCs. I suggest that this serves as a quick response mechanism, allowing new TFs to reach their effective concentration faster during differentiation. However, as this work currently lacks expression dynamics throughout the differentiation process, I can only speculate.

2.4 Materials & Methods

2.4.1 Annotation of miRNA promoters

High confidence miRNA TSSs from the integrated miRNA Expression Atlas(Rie et al. 2017) were merged with TSSs from transcriptions derived from Drosha dominant-negative experiments (Chang et al. 2015). These TSS were remapped to hg38 using the UCSC liftover tool and combined with TSS from GENCODE v26. I then used the BEDOPS (Neph, Stergachis, et al. 2012) closest-feature tool to identify the closest DHS from a nonoverlapping set of DHS

generated from the samples from ENCODE. TSSs not within 1KB of a DHS from the master list were removed from our analysis.

2.4.2 Promoter Clustering

I used BEDOPS to generate a genome genome-wide binary matrix for miRNA promoters. Our promoters were overlapped with DHS master list for the samples/conditions used in our analysis. A binary call was made for each promoter in each sample based on the whether it overlapped with a DHS peak by at least 25% in that sample. Promoters not active in any sample were removed from the matrix for this analysis. Clustering was done for both samples and promoters using Euclidean distance and Ward clustering.

2.4.3 Cross-Regulatory network generation

I modeled TF-miRNA networks as a graph consisting of three different kinds of edges, TF-primiRNA, primiRNA-miRNAs, and miRNA-TF, and deriving each type comes with its own considerations. For TF-primiRNA edges I defined a primiRNA as all mature miRNAs under the control of the same promoter. This varies slightly from the usual definition of primiRNA but captures the complete dynamics of miRNA expression. primiRNA-miRNA edges are calculated once, for every promoter, mapping them to the specific mature miRNA sequences under the control of that promoter. miRNA-TF edges are also calculated only once, and I used a combination of computationally predicted targets from TargetScan (Agarwal et al. 2015) 7.1 and experimentally derived targets from miRTarBase (Chou et al. 2017) 7.0. I required each experimentally derived miRNA target to also have a predicted binding site (conserved or nonconserved) for that miRNA, which was done to alleviate the potential for false positives inherent to both methodologies. To construct each graph, I scanned miRNA promoters for a

collection TF binding models (Mathelier et al. 2016; Jolma et al. 2013; Newburger and Bulyk 2009; Wingender et al. 1996; Kulakovskiy et al. 2017) using the FIMO (Bailey et al. 2009) tool with default parameters with a maximum p-value threshold of 1×10^{-5}. If a FIMO hit overlapped a DNase footprint by at least 3bp in a particular sample, an edge was added to the network between that TF-primiRNA combination. This process was repeated for all samples/conditions.

2.4.4 Identification of network motifs

TF-miRNAs were identified with a custom python script that identified cycles within these networks. Due to the tripartite nature of the underlying graph model, cycles only occur when a TF binds a promoter that controls the expression of a miRNA that in turn targets that same transcription factor. The significance of FBCs was calculated on a per-network basis by fitting a hypergeometric distribution using TFs and miRNAs found in that specific network. A p-value was calculated using SciPy. For networks only involving a single type of edge, such as the TF-TF regulatory networks, mfinder (Milo et al. 2002) was used for motif identification.

2.4.5 Processing of NANOSTRING nCounter expression

Nanostring data was processed as directed by the nCounter manual. Sample specific variation was removed via positive probe normalization, and samples with a normalization factor outside the expected range were removed from the analysis. In the cases where a binarized expression call was necessary, a given miRNA was considered as being expressed if it had an expression value of at least 2 standard deviations above the mean of the negative control probes for that sample. Tissue specific expression was assessed by averaging over all samples from the same tissue, which provided a more robust measurement at the cost of losing resolution to detect expression dynamics over developmental time.

2.4.6 ROADMAP RNAseq

Sequencing reads were aligned to the reference human genome (GRCh37/hg19) using TopHat (Trapnell, Pachter, and Salzberg 2009) algorithm25 using parameters "-p 4 -g 10 –segment-mismatches 0 –segment-length 18" and expression values were determined using the Cufflinks (Trapnell et al. 2010; Roberts, Trapnell, et al. 2011; Roberts, Pimentel, et al. 2011; Trapnell et al. 2013) algorithm10 with the comprehensive set of GENCODE v19 transcript models16. Expression calls were made using an FPKM threshold of 1.0 and genes with expression with a z-score of at least 2.0 are considered enriched.

Chapter 3: Novel Methods for the Automated Design of Transcriptional Activator-Like Effectors and the Analysis of Editing Outcomes

3.1 Introduction

Recent advances in editing technology have substantially improved our ability to make directed changes to DNA, making it possible for the first time in history to make targeted perturbations to regulatory networks in human cells. Though the CRISPR/Cas9 system is quickly becoming the dominant strategy for genome editing (Doudna and Charpentier 2014), the field itself actually goes back more than 20 years (Urnov 2018). Early work using meganucleases (Johnson and Jasin 2001) and zinc finger nucleases (Urnov et al. 2005) demonstrated the capability of making precise changes to endogenous loci by introducing DNA and a double-stranded break (DSB). The field of genome engineering started with advent of programmable

nucleases, created by tethering the DBD of zinc finger proteins to an endonuclease, which allowed researchers to make site-specific DSBs (Urnov et al. 2010). This concept of attaching a nuclease to a DBD was transferred to TAL effector proteins, which are easy to assemble and have a very predictable binding code (Bogdanove and Voytas 2011).

3.1.1 Transcriptional Activator-Like Effectiors

Transcriptional Activator-Like Effectors (TALEs) are a class of proteins first identified in a pathogenic bacterium of the Xanthomonas genus that infect plants (Schornack et al. 2006; Kay and Bonas 2009). In their native context, these proteins directly bind the promoter region of plant genes, affecting host expression. The functional unit of TAL proteins is a modular set of repetitive domains consisting of 33-35 amino acids that bind DNA in a sequence specific fashion. The nucleotide specificity of a given repeat domain is conferred by residues 12 and 13, also known as the "repeat-variable di-residue" or RVD. Previous work has reverse engineered the binding specificities of known RVDs, creating a platform for programmable DBDs (Moscou and Bogdanove 2009; Boch et al. 2009). Though naturally occurring TALE proteins contain an activating domain, these proteins can be fused to an endonuclease, most commonly the obligate homodimer nuclease FOK1, to yield engineered nucleases that specifically target DNA regions of interest (Cermak et al. 2011).

Though it is outside the scope of this work, our lab has developed a technique for quickly assembling TALENs and deploying them at scale. Although the CRISPR/Cas9 system is trivial to assemble and deploy, we find that TALENs offer a primary advantage of targeting specificity. Figure 3.1 **A** shows a comparison of possible target sites for TALENs and Cas9 in the functional regions of the human genome (hg38). I calculated these numbers by designing all possible TALEN dimer pairs (cut site was taken as just the center-most nucleotide in the spacer) and

CRISPR guides in GENCODE annotated genes and active regulatory regions. TALENs target roughly 91% of all nucleotides, while Cas9 only targets 12%. I also find that the distance between cut sites is significantly smaller using a TALEN based system than using CRISPR/Cas9 (Figure 3.1 **B**). I find this increase in targetability appealing, especially when it comes to fine-grain dissection of regulatory elements, and it is in this context that the development of the following computational tools lies.

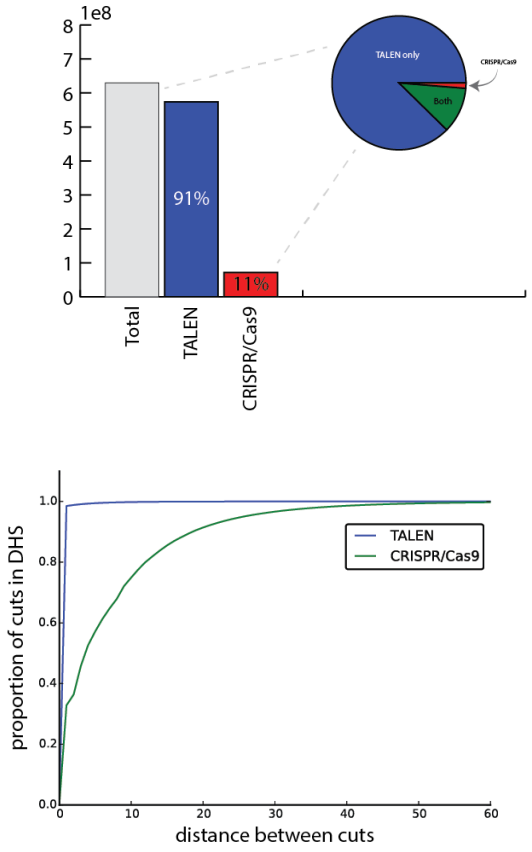

Figure 3.1 TALENs vs. Cas9

Systematic comparison of the targetability of TALENs vs CRISPR/Cas9.

A. Overall number of nucleotides within genes and active regulatory regions targetable by both genome editing platforms.
B. Plot describing the distance between cuts for each platform.

3.2 Results

3.2.1 DesignTools: a suite of tools for automated TAL Effector Design

Several computational tools exist for the design of TALEs and agents for genome editing (Montague et al. 2014; Doyle et al. 2012), however the process of designing these agents is still primarily a manual one. I wanted to develop a set of computational tools to generalize common editing tasks so that the design of editing agents could be done at large scale. I identified three common editing tasks that I wanted to automate. The first of these tasks is locus deletion, in which TALENs have previously been shown to be effective at creating large deletions in human cells (H. Lee, Kim, and Kim 2010; Hu et al. 2013; Kim et al. 2013; Cong et al. 2013). The second task is disruption of DNA function via multiple cuts spanning the entire length of a given DNA element. This task is based on a procedure known as functional footprinting (Vierstra et al. 2015), where it was shown to be effective in probing the function of regulatory DNA elements. The third task is introducing a single, highly specific double-stranded break to disrupt fine scale elements or introduce user provided DNA via homology directed repair (HDR), which shows promise as a mode of endogenous gene repair (Urnov et al. 2010). Thus, I create three running modes of DesignsTools; flank, strafe, and precision edit to address tasks one, two, and three respectively.

3.2.1.1 Creating an index of TALE seeds

Naturally occurring TALE binding sites show a significant preference for a T nucleotide in the first position (Moscou and Bogdanove 2009; Boch et al. 2009). Though this is not a hard requirement, it presents a beneficial motif to identify potential TALE monomers. Tabix (Sokol

and Ambros 2005; Y. Li et al. 2006; Cassidy et al. 2013) is a generalized tool for indexing tabular data for quick retrieval. As a pre-processing step, I wanted to exhaustively enumerate all possible starting positions for TALE monomers genome wide. To do this, I wrote a custom script that takes a FASTA file containing a sequence of interest, in our case I used the hg38 assembly of the human genome and produces a bed format file of T positions on both strands. This file can be easily queried for potential TALE monomers/dimer pairs by downstream design applications.

3.2.1.2 Scoring of TALE monomers

An integral part of any automated design pipeline is the ability to rank potential designs so that the most ideal design can be selected for a given task. For this, I developed a scoring method for TALE monomers based roughly on the criteria described in the literature (Streubel et al. 2012) as well as anecdotal guidelines obtained through past experiments in our lab. Our current scoring method consists of the following criteria: GC content, start strength, and balance. All three of these criteria leverage the fact that RVDs binding nucleotides G and C bind more tightly than those binding A and T. First, we calculate the overall GC content of each monomers. Then, we calculate the starting strength, which is defined as the GC content of the first three bases of the TALEN binding site. Finally, we calculate the average distance between Gs and Cs and divide it by the total length of the TALE monomer. Each of these yields a number between 0 and 1 and the score for a given monomer is the average of these three values. Scores for TALEN monomer pairs are an average of the two individual monomer scores.

3.2.1.3 Flank Mode

Flank mode is a mode for designing TALEN pairs for deleting a region of interest (ROI), which follows a few basic steps. First the ROI is padded outwards, both 5’ and 3’ of the ROI, to

derive regions where the cuts will be made. Then these padded regions are divided into subregions based on the number of designs desired (*i.e.* 2 subregions if the objective is to design 2 TALEN pairs per side). Finally, the highest scoring TALEN pairs are chosen from each subregion, ensuring all TALE monomers are unique. The resulting TALEN designs are output in 4 column bed format for easy review.

3.2.1.4 Strafe Mode

Strafe mode is a mode intended to design TALEN pairs to densely tile a genomic locus to functionally characterize elements contained within. First, the ROI is divided evenly into bins based on the desired resolution/spacing. By default, the tool places a cut site every 5bps. Then from left-to-right the highest scoring TALEN from each bin is selected in a fashion that, once again, avoids redundancy in TALE monomers. The resulting TALEN pairs are output for review/assembly.

3.2.1.5 Precision-edit Mode

This mode is centered around designing TALEN pairs for fine-grained editing experiments, such as endogenous gene correction. In this mode the ROI generally consists of a small region, as low as a single base pair, where a cut site is to be placed. This mode can be run in two distinct way, as a default edit or a scar-less edit. By default, the highest scoring n TALEN pairs are selected, based on the desired number of edits, avoiding redundancy. With scar-less editing, the TALEN pairs are designed so that one of the TALE monomers overlaps the ROI, so that introducing sequence changes via HDR disrupts rebinding of that monomer post edit.

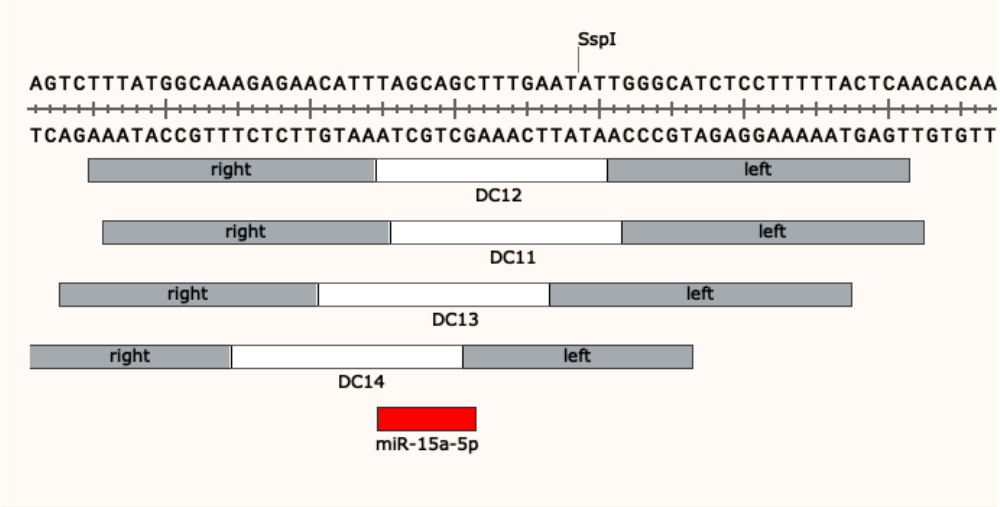

Figure 3.2 Precision Edit Example

Example TALENs designed automatically with precision edit. The left and right monomers are displayed in gray for each TALEN pair DC11-DC14 as well as the space in white. All TALEN pairs were designed such that their spacer sequence overlaps with the binding site for miR-15a-5p.

3.2.1.6 Adapting designs to epigenetic modifiers

Although these tools were written with targeted nucleases in mind, it only takes minor adjustment to adapt these tools for epigenomic modification. Our lab and others (Cong et al. 2012) have been able to achieve transcriptional repression by fusing a transcriptional repressor domain to the DNA Binding Domain of TALE proteins. This can be achieved by placing these fusion proteins in the promoter regions of protein coding genes. Strafe mode has been easily co-opted for this purpose by tiling TALEN pairs across an ROI and using the higher scoring TALE monomer of each pair. The resulting TALE monomers can then be assembled into an epigenome altering context. Though this application does not constitute a novel feature on its own, it does

illustrate the versatility of our platform for generating automated designs for TALE proteins for a wide variety of experiments.

3.2.2 Tools for large scale assembly and automated liquid handling

The most widely used method for TALE protein assembly is Golden Gate cloning (Engler, Kandzia, and Marillonnet 2008), and this method has been shown to produced high quality TALEs fast and effectively (Cermak et al. 2011; Sakuma and Yamamoto 2016). However, to produce TALE proteins at a large scale, it became necessary to adapt this method for assembly via robotic liquid handling. Golden Gate cloning is a stepwise assembly method that requires the use of many vectors as well as a bacterial growth step. To accomplish this, I wrote a set of scripts that maps TALE proteins designed with DesignTools into plate format for multiplexed Golden Gate ligation reactions and generates specific protocol for our liquid handling system to dispense reagents. This tool also generates expected digestion sizes for validation at both stages of the assembly process as described in previous work (Cermak et al. 2011). Though our specific liquid handling is conducted on the Labcyte Echo, this code can easily be adapted for generating protocols for any automated liquid handling system.

3.2.3 CLEAN-CUT: Calculation of the Length of Edits And Number of CUTs

A logical next step was to develop a platform where I could efficiently test the efficacy of assembled TALEN pairs in a high throughput fashion. Generally, an excess of TALE proteins is built to achieve a given editing task and TALENs with the highest predicted efficiency are used first in experiments. First the TALENs are assembled, transfected into K562 cells, and then the DNA region surrounding the cut site is PCR amplified (amplicon) and sequenced to assess editing efficiency. End-to-end amplicon sequencing is an appealing technique for assessing the

outcome of editing experiments, as it minimizes the coverage biases experienced with other types of short read sequencing (Ravi, Walton, and Khosroheidari 2018). I developed a two-tier system of sample multiplexing. The first layer of multiplexing is based on Illumina i5 and i7 adaptors, and the second is based on unique amplicon sequence. Samples can be assigned the same Illumina adaptor index combination as long as there is no overlap between the respective amplicons. Once the sequencing is complete, the flowcell is processed by standard Illumina software and the resulting FASTQ files then go on to the next step of our custom pipeline.

3.2.3.1 Data pre-processing

The first step of our quantification pipeline is to merge paired end reads into a single, high quality read and align this read to the expected wildtype sequence (usually assumed to be that of hg38 reference sequence). Specifically, the FASTQ files generated in the previous step are first merged using PEAR (Zhang et al. 2014). Then, via a custom script, reads are combined into a FASTA file in such a way that each entry is unique, and the ID field contains the number of times that sequence occurs in the sequencing reads (corrected for PCR duplicates). This is done so each sequence, which may be present in the raw sequencing reads > 10,000X, only needs to be aligned to the reference amplicon once. This combined FASTA file is used as input to the BLAT aligment tool (Kent 2002) and an alignment file (PSL format) is produced for each amplicon from each unique i5, i7 adaptor combination. These PSL files serve as the input to the CLEANCUT tool, where it is processed using one of three running modes: dimer-efficiency, deletion-efficiency, or clonal genotyping.

3.2.3.2 Dimer-efficiency

The editing efficiency of a given TALEN dimer can be calculated as the number of reads showing evidence of having been cut by that TALEN pair divided by the total number of reads. Reads that show evidence of cutting mostly contain a small insertion or deletion (INDEL) with respect to the reference amplicon. These INDELs manifest as gaps in the PSL alignment file. DSBs resulting from a TALEN pair don't happen in a pre-defined place, instead they occur in a normal-like distribution centered around the midpoint of the spacer sequence between monomer binding sites, so any INDEL resulting from the TALEN cutting should overlap with the spacer sequence. To calculate the editing efficiency, I iterate over the alignments and record both the number of reads that contain INDELs overlapping the TALEN spacer and the total reads. The editing efficiency is then calculated as the number of reads with INDELs divided by the total read count. This mode also outputs formatted alignments as well as a plot of the cut sites and deletion profile over the length of the amplicon (Figure 3.3).

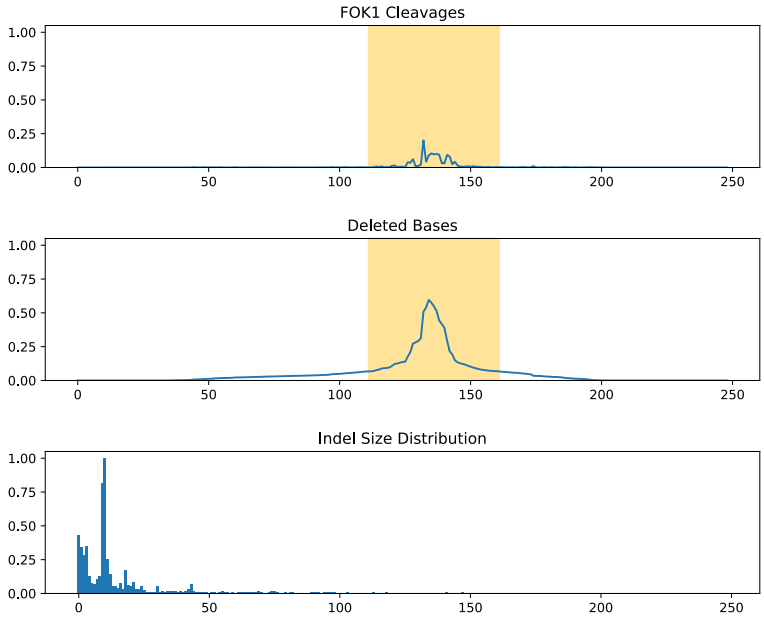

Figure 3.3 Dimer Deletion Profile

Example dimer profile for a TALEN targeting AAVS1. Plot show the distribution of FOK1 cleavages and deleted bases as well as a histogram of the INDEL sizes.

3.2.3.3 Deletion-efficiency

Deletion efficiency can be defined as the number of reads where the intervening sequence between two TALEN monomer pairs has been deleted divided by the total number of reads. Though conceptually simple, calculating the deletion efficiency has a few complications that need to be considered. The first complication is that deletions often cover large genomic regions (spanning several KB in some occasions), and the resulting amplicon would be too large to fit on a MiSeq sequencing read (300bp). This complication can be overcome by designing primers such that each end of the paired end sequence read extends into the TALEN sequence of one of the two TALEN pairs. This approach leads to the second complication though, which arises from

the fact that, given the same primers, a genomic region with a large deletion will be preferentially amplified via PCR compared to the full-sized region, leading to an overrepresentation of deletions in the sequencing pool, and an overestimation of overall deletion efficiency.

To overcome these challenges, we developed a method that I refer to as the three-primer assay. This three-primer assay involves computationally designing a set of three primers, two flanking the outer bounds of the predicted deletion and the third lying in the intervening sequence between dimer cut sites. These primers were designed to have similar melting/annealing temperatures and such that the resulting amplicon from the inner primers will be similar in size to that of the outer primers when the deletion occurs. Reads originating from the intervening primer are called as wildtype, while reads from the outermost primers are called as containing a deletion, and the deletion efficiency can then accurately be calculated as the number of reads containing deletions divided by total reads. Deletion efficiencies derived from this method are consistent with estimations based on PCR combined with more traditional readout methods, such as agarose gel.

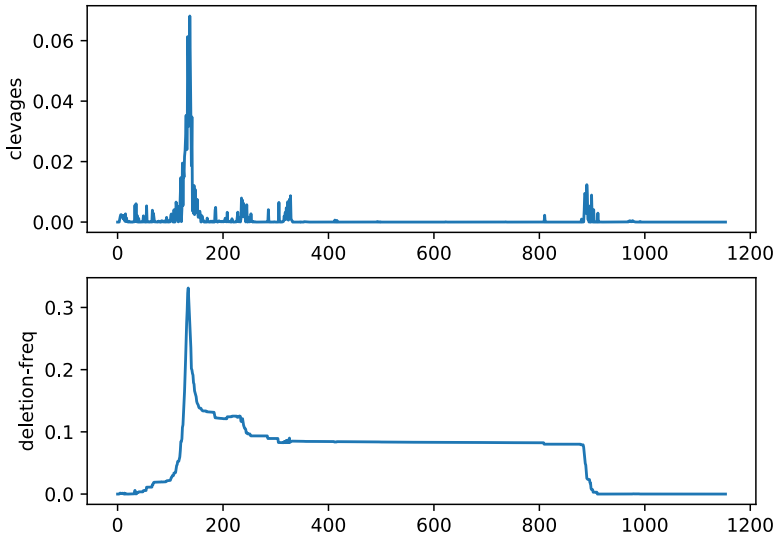

Figure 3.4 Deletion Plot

Anonymized example of an INDEL profile from a deletion experiment. On top is the cleavage profile of two TALEN dimers and on bottom is the deletion profile consistent with the intervening sequence between the TALEN pairs being deleted in a significant percentage of the sequencing reads.

3.2.3.4 Clonal Genotyping

Another hurdle to high-throughput genome engineering is the lack of computational tools that characterize populations derived from single-cell clones. To address this, I adapted our INDEL detection tools for the purpose of genotyping, proceeding in two steps. The first is to exhaustively enumerate and assess counts for all alleles in a population, and the second is to then make a statistically rigorous genotype call based on the chromosomal configuration that best fits the data.

To exhaustively enumerate all possible alleles, I conceptualized an allele "tagging" system that I call the mutation profile. The mutation profile is a unique identifier that consist of an ordered list of all mutations, INDELs and polymorphisms, for a given alignment based on where they occur in the reference amplicon. To derive the mutation profile, I iterate over the reads in two passes. The first pass is used to calculate the sequencing error at each position as a background model for polymorphisms. If a mismatch is detected, it is only considered a true polymorphism if it is present at a frequency higher than two times the standard deviation of the mean of the background counts, otherwise it is considered a sequencing error. For the second pass, I construct the indel profile by adding the type of mutation, as well as the position, to an ordered list as I traverse each alignment from left to right. If two alignments have the same mutational profile, they are considered the same allele and are counted together. At the end of both passes, I have counts for all alleles in the population.

If the sequencing reads are from a complex population there will be an exceedingly large number of alleles with a relatively small number of counts, but if the sequencing comes from a clonal population there should be a small number of alleles with relatively high counts(Luria and lbrück 1943). Operating under the assumption that the population is clonal, a genotyping call can then be made by the ratios of the counts of prominent alleles in the population. By default, I consider an allele as prominent only if it has a count higher than 10% of the total counts. Though simplistic, I have found that in practice this method yields a genotype call that is expected based on known chromosomal copy number of the cells commonly used in our lab (*i.e.* amplicons from loci with a known copy number of 3, have 3 distinct alleles). Figure 3.5 shows example genotypes from an experiment targeting an miRNA target site in the 3' UTR of YY1.

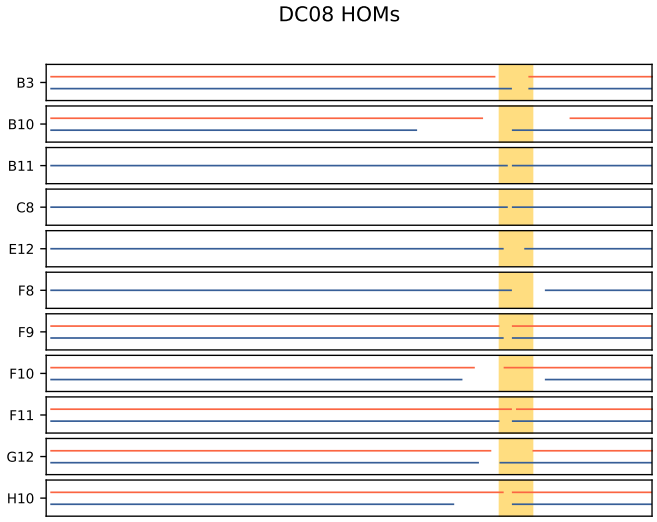

Figure 3.5 Homozygous Deletion Clones

Example Deletion Clones flagged by CLEANCUT as having INDELs disrupting the binding site for miR-19b-3p in the YY1 3' UTR. Different alleles are displaying in differing colors, and in the case of identical alleles only one is shown.

3.3 Summary

Here I present computational tools for conducting high throughput genome editing experiments. The first set of tools, DesignTools, is for the automated design and assessment of TAL effector proteins. DesignTools can effectively design TALEs for four distinct editing tasks and in our experience our current scoring metric does a decent job of avoiding monomers that are likely to bind/function poorly. The second set of tools creates instructions for the assembly of TALE monomers using Golden Gate Cloning for a robotic liquid handling system. I assert that this tool can easily be modified for a wide range of robots that are compatible with 96 or 384 well format. The final set of tools are a novel set of tools that broadly evaluate the results of editing experiments from amplicon sequencing data. This system can handle hundreds of

samples on a single flowcell and quickly quantify TALEN dimer cutting rate, dimer pair deletion efficiency, and genotype clonal populations. Taken together, this suite of tools represent a near comprehensive set of tools to conduct high-throughput TALE based editing experiments.

3.4 Materials & Methods

3.4.1 Preprocessing of NGS Reads

Sequencing reads were processed in place using Illumina'a BCL2Fastq with a custom sample sheet designed for our 96-well sequencing format, where samples are fitted with a standard set of Illumina adaptors. This represents the first tier of sample multiplexing. Paired end reads are merged with PEAR (Zhang et al. 2014) under the default settings, and duplicate reads are filtered out via UMIs (Kivioja et al. 2012). A user defined spec sheet is then used to align reads using BLAT to relevant amplicons provided by the experimenter. An alignment file is created for each amplicon combination. This is the second tier of sample multiplexing.

3.4.2 Computational Tools

All aforementioned computational Tools are implemented in python 3.6 in a unix coding environment. Where possible I parallelized code to run on a standard SLURM cluster for high performance computing.

Chapter 4: Conclusion

TF-miRNA cross-regulatory networks provide an interesting view into how genes are regulated on a systems level. This work is the first to demonstrate that TF-miRNA negative

feedback circuits are a prominent network motif and demonstrates that these circuits likely play roles in enforcing cell-fate decisions and fine-tuning expression on a scale not previously appreciated. Our work shows that the feedback between TFs and miRNAs also play a major role in the process of cells transitioning from one state to another. While cells utilize these mechanisms to reliably reproduce differentiation and development across humans as a specifies, further understanding of this process provide insights into how the breakdown of these mechanisms could lead to developmental disorders, or diseases like cancer.

This work also demonstrates a previously under-appreciated role for negative regulation in human development. In essence, miRNAs function as signal amplifiers for TFs allowing a given TF to influence the expression of a large number of genes with a single regulatory interaction. With such a large effect, understanding how negative regulation influences transcriptional networks presents an interesting challenge. In their seminal work on network superfamilies *Milo et al.* describe a concept called the triad significance profile (TSP) (R Milo et al. 2002) which measures the enrichment of 3-node network motifs in a given network. They show that complex networks from different sources, including Drosophila regulatory networks and *C. elegans* neuron networks, all share a similar TSP and thus converge to a similar architecture. Previous work in our lab shows that TF regulatory network across a diverse set of cell types also converge to this similar architecture (Neph, Stergachis, et al. 2012). I hypothesized that networks of negative regulation, derived from miRNA interactions, would also converge to the same architecture. To test this hypothesis, I constructed what I refer to as the projected negative regulatory (PNR) network, which reduces TF->miRNA->TF interactions to TF->TF interactions, essentially removing the miRNA intermediate. In Figure 4.1 I show the TSP in PNR networks all converge to a common architecture that is distinct from that of TF-TF

networks and in fact more closely resembles the language superfamily *Milo et al* previously described (Ron Milo et al. 2004). This suggests that negative regulation serves a complementary role to regulated recruitment and exploring this relationship further presents an interesting challenge to gaining a wholistic view of human regulatory programs.

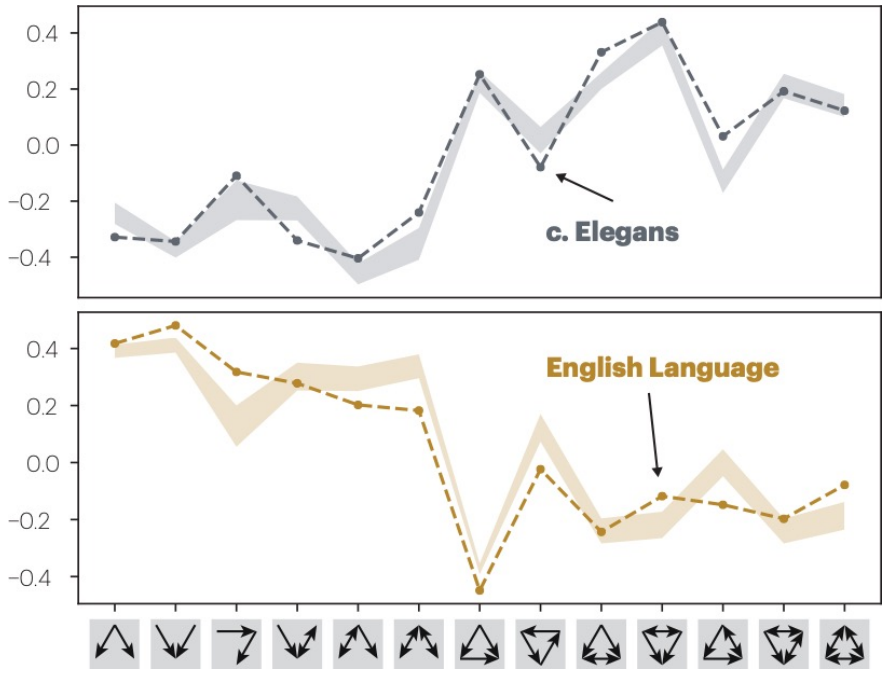

Figure 4.1 Triad Significance Profiles

The triad significance profiles of different types of regulatory networks. In the upper panel is the average TSPs of 156 samples across diverse cell types and conditions (after *Neph et al.* 2012(Neph, Stergachis, et al. 2012), *C. elegans* TSP shown for reference). In the bottom panel is the average TSP from the PNR networks across those same samples. The PNR TSP converges to a complementary architecture (TSP from the English language from the language superfamily shown for reference).

Several limitations of this work arise from the use miRNA target sites that have been computationally predicted by TargetScan, which sometimes do not represent genuine targets. I attempt to alleviate this by requiring our targets to have some experimental evidence, but sometimes, especially in the case of more high throughput methods of experimental validation, these targets also result from spurious interactions arising from the proximity based target

detection methods (Stork and Zheng 2016; Helwak et al. 2013). As the field gets better at predicting and validating miRNA targets, our knowledge of genuine miRNA targets will increase.

Another limitation derives from the fact that interactions inferred from DNase I footprinting lack sign, meaning that I cannot tell whether the interaction is activating or repressing gene expression. Any specific TF-miRNA interaction inferred by our method must be "taken with a grain of salt", however the general trends described here are still interesting.

4.1 Future Directions

I am currently in the process of experimentally validating some of the FBCs discovered by our analysis. To date I have used our TALEN pipeline to disrupt the miRNA binding sites in the UTR of 3 TFs that participate in FBCs. I then tested the protein levels of these TFs via western blot, observing an increase in overall protein level. Finding clonal knockouts involved building a large number of TALENs, and screening lots of clones, which would not have been possible without our aforementioned computational tools for TALEN design, assembly, and clone assessment. I am currently in the process of using our assembly pipeline to design TALE proteins disrupting TF binding sites in miRNA promoters.

I am also currently interested in improving our ability to computationally predict the efficacy of TALENs. Because of our high throughput pipeline, the amount of TALENs for which I have efficiency data for is ever increasing. This creates the opportunity for a data driven scoring metric, and the use of machine learning to determine the "rules" of TALEN editing.

Bibliography

Agarwal, Vikram, George W Bell, Jin-Wu Nam, and David P Bartel. 2015. "Predicting Effective MicroRNA Target Sites in Mammalian MRNAs." *ELife* 4: e05005. doi:10.7554/elife.05005 .

Amit, Ido, Manuel Garber, Nicolas Chevrier, Ana Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, et al. 2009. "Science-Sciencemag-Org.Offcampus.Lib.Washington.Edu 11/24/2019, 7:35:03 PM.Pdf." *Science* 326 (5950): 257–263. doi:10.1126/science.1179050 .

Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. 2009. "MEME Suite: Tools for Motif Discovery and Searching." *Nucleic Acids Research* 37 (suppl_2): W202–W208. doi:10.1093/nar/gkp335 .

Bartel, David P. 2004. "MicroRNAs Genomics, Biogenesis, Mechanism, and Function." *Cell* 116 (2): 281–297. doi:10.1016/s0092-8674(04)00045-5 .

Bartel, David P. 2009. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell* 136 (2): 215–233. doi:10.1016/j.cell.2009.01.002 .

Basso, Katia, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. 2005. "Reverse Engineering of Regulatory Networks in Human B Cells." *Nature Genetics* 37 (4): 382–390. doi:10.1038/ng1532 .

Betel, Doron, Manda Wilson, Aaron Gabow, Debora S Marks, and Chris Sander. 2008. "The MicroRNA.Org Resource: Targets and Expression." *Nucleic Acids Research* 36 (suppl_1): D149–D153. doi:10.1093/nar/gkm995 .

Boch, Jens, Heidi Scholze, Sebastian Schornack, Angelika Landgraf, Simone Hahn, Sabine Kay, Thomas Lahaye, Anja Nickstadt, and Ulla Bonas. 2009. "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors." *Science* 326 (5959): 1509–1512. doi:10.1126/science.1178811 .

Bogdanove, Adam J, and Daniel F Voytas. 2011. "TAL Effectors: Customizable Proteins for DNA Targeting." *Science* 333 (6051): 1843–1846. doi:10.1126/science.1204094 .

Bossard, P, and KS Zaret. 1998. "GATA Transcription Factors as Potentiators of Gut Endoderm Differentiation." *Development (Cambridge, England)* 125 (24): 4909–4917.

Bracken, Cameron P, Philip A Gregory, Natasha Kolesnikoff, Andrew G Bert, Jun Wang, Frances M Shannon, and Gregory J Goodall. 2008. "A Double-Negative Feedback Loop between ZEB1-SIP1 and the MicroRNA-200 Family Regulates Epithelial-Mesenchymal Transition." *Cancer Research* 68 (19): 7846–7854. doi:10.1158/0008-5472.can-08-1942 .

Carro, Maria, Wei Lim, Mariano Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P

Sulman, et al. 2010. “The Transcriptional Network for Mesenchymal Transformation of Brain Tumours.” *Nature* 463 (7279): 318–325. doi:10.1038/nature08712 .

Cassidy, Justin J, Aashish R Jha, Diana M Posadas, Ritika Giri, Koen Venken, Jingran Ji, Hongmei Jiang, Hugo J Bellen, Kevin P White, and Richard W Carthew. 2013. “MiR-9a Minimizes the Phenotypic Impact of Genomic Diversity by Buffering a Transcription Factor.” *Cell* 155 (7): 1556–1567. doi:10.1016/j.cell.2013.10.057 .

Cermak, Tomas, Erin L Doyle, Michelle Christian, Li Wang, Yong Zhang, Clarice Schmidt, Joshua A Baller, Nikunj V Somia, Adam J Bogdanove, and Daniel F Voytas. 2011. “Efficient Design and Assembly of Custom TALEN and Other TAL Effector-Based Constructs for DNA Targeting.” *Nucleic Acids Research* 39 (12): e82–e82. doi:10.1093/nar/gkr218 .

Chang, Tsung-Cheng, Mihaela Pertea, Sungyul Lee, Steven L Salzberg, and Joshua T Mendell. 2015. “Genome-Wide Annotation of MicroRNA Primary Transcript Structures Reveals Novel Regulatory Mechanisms.” *Genome Research* 25 (9): 1401–1409. doi:10.1101/gr.193607.115 .

Chen, Jian-Fu, Elizabeth P Murchison, Ruhang Tang, Thomas E Callis, Mariko Tatsuguchi, Zhongliang Deng, Mauricio Rojas, et al. 2008. “Targeted Deletion of Dicer in the Heart Leads to Dilated Cardiomyopathy and Heart Failure.” *Proceedings of the National Academy of Sciences* 105 (6): 2111–2116. doi:10.1073/pnas.0710228105 .

Chou, Chih-Hung, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, et al. 2017. “MiRTarBase Update 2018: A Resource for Experimentally Validated MicroRNA-Target Interactions.” *Nucleic Acids Research* 46 (D1): gkx1067-. doi:10.1093/nar/gkx1067 .

Cirillo, Lisa A, Clifton E McPherson, Pascale Bossard, Kimberly Stevens, Sindhu Cherian, Eun Shim, Kirk L Clark, Stephen K Burley, and Kenneth S Zaret. 1998. “Binding of the Winged-helix Transcription Factor HNF3 to a Linker Histone Site on the Nucleosome.” *The EMBO Journal* 17 (1): 244–254. doi:10.1093/emboj/17.1.244 .

Cong, Le, Ann F Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” *Science* 339 (6121): 819–823. doi:10.1126/science.1231143 .

Cong, Le, Ruhong Zhou, Yu-chi Kuo, Margaret Cunniff, and Feng Zhang. 2012. “Comprehensive Interrogation of Natural TALE DNA-Binding Modules and Transcriptional Repressor Domains.” *Nature Communications* 3 (1): 968. doi:10.1038/ncomms1962 .

Consortium, The. 2004. “The ENCODE (ENCyclopedia Of DNA Elements) Project.” *Science* 306 (5696): 636–640. doi:10.1126/science.1105136 .

Cosma, Maria, Tomoyuki Tanaka, and Kim Nasmyth. 1999. “Ordered Recruitment of Transcription and Chromatin Remodeling Factors to a Cell Cycle– and Developmentally Regulated Promoter.” *Cell* 97 (3): 299–311. doi:10.1016/s0092-8674(00)80740-0 .

Covert, Markus W, Eric M Knight, Jennifer L Reed, Markus J Herrgard, and Bernhard O Palsson. 2004. “Integrating High-Throughput and Computational Data Elucidates Bacterial Networks.” *Nature* 429 (6987): 92–96. doi:10.1038/nature02456 .

Damiani, Devid, John J Alexander, Jason R O’Rourke, Mike McManus, Ashutosh P Jadhav, Constance L Cepko, William W Hauswirth, Brian D Harfe, and Enrica Strettoi. 2008. “Dicer Inactivation Leads to Progressive Functional and Structural Degeneration of the Mouse Retina.” *The Journal of Neuroscience* 28 (19): 4878–4887. doi:10.1523/jneurosci.0828-08.2008 .

Davidson, Eric H, Jonathan P Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, et al. 2002. “A Provisional Regulatory Gene Network for Specification of Endomesoderm in the Sea Urchin Embryo.” *Developmental Biology* 246 (1): 162–190. doi:10.1006/dbio.2002.0635 .

Davis, Robert L, Harold Weintraub, and Andrew B Lassar. 1987. “Expression of a Single Transfected CDNA Converts Fibroblasts to Myoblasts.” *Cell* 51 (6): 987–1000. doi:10.1016/0092-8674(87)90585-x .

Davis, Tigwa H, Trinna L Cuellar, Selina M Koch, Allison J Barker, Brian D Harfe, Michael T McManus, and Erik M Ullian. 2008. “Conditional Loss of Dicer Disrupts Cellular and Tissue Morphogenesis in the Cortex and Hippocampus.” *The Journal of Neuroscience* 28 (17): 4322–4330. doi:10.1523/jneurosci.4815-07.2008 .

Doudna, Jennifer A, and Emmanuelle Charpentier. 2014. “The New Frontier of Genome Engineering with CRISPR-Cas9.” *Science* 346 (6213): 1258096. doi:10.1126/science.1258096 .

DOUGLAS, HC, and DC HAWTHOE. 1964. “ENZYMATIC EXPRESSION AND GENETIC LINKAGE OF GENES CONTROLLING GALACTOSE UTILIZATION IN SACCHAROMYCES.” *Genetics* 49: 837–844.

Doyle, Erin L, Nicholas J Booher, Daniel S Standage, Daniel F Voytas, Volker P Brendel, John K VanDyk, and Adam J Bogdanove. 2012. “TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: Tools for TAL Effector Design and Target Prediction.” *Nucleic Acids Research* 40 (W1): W117–W122. doi:10.1093/nar/gks608 .

Dunn, S.-J., G Martello, B Yordanov, S Emmott, and AG Smith. 2014. “Defining an Essential Transcription Factor Program for Naïve Pluripotency.” *Science* 344 (6188): 1156–1160. doi:10.1126/science.1248882 .

Dynan, William S, and Robert Tjian. 1983. “The Promoter-Specific Transcription Factor Sp1 Binds to Upstream Sequences in the SV40 Early Promoter.” *Cell* 35 (1): 79–87. doi:10.1016/0092-8674(83)90210-6 .

Engler, Carola, Romy Kandzia, and Sylvestre Marillonnet. 2008. “A One Pot, One Step, Precision Cloning Method with High Throughput Capability.” *PLoS ONE* 3 (11): e3647.

doi:10.1371/journal.pone.0003647 .

Friedman, Nir. 2004. “Inferring Cellular Networks Using Probabilistic Graphical Models.” *Science* 303 (5659): 799–805. doi:10.1126/science.1094068 .

Friedman, Robin C, Kyle Farh, Christopher B Burge, and David P Bartel. 2009. “Most Mammalian MRNAs Are Conserved Targets of MicroRNAs.” *Genome Research* 19 (1): 92–105. doi:10.1101/gr.082701.108 .

Galas, David J, and Albert Schmitz. 1978. “DNAase Footprinting a Simple Method for the Detection of Protein-DNA Binding Specificity.” *Nucleic Acids Research* 5 (9): 3157–3170. doi:10.1093/nar/5.9.3157 .

Graf, Thomas, and Tariq Enver. 2009. “Forcing Cells to Change Lineages.” *Nature* 462 (7273): 587–594. doi:10.1038/nature08533 .

Greenfield, Alex, Christoph Hafemeister, and Richard Bonneau. 2013. “Robust Data-Driven Incorporation of Prior Knowledge into the Inference of Dynamic Regulatory Networks.” *Bioinformatics* 29 (8): 1060–1067. doi:10.1093/bioinformatics/btt099 .

Gregory, Philip A, Andrew G Bert, Emily L Paterson, Simon C Barry, Anna Tsykin, Gelareh Farshid, Mathew A Vadas, Yeesim Khew-Goodall, and Gregory J Goodall. 2008. “The MiR-200 Family and MiR-205 Regulate Epithelial to Mesenchymal Transition by Targeting ZEB1 and SIP1.” *Nature Cell Biology* 10 (5): 593–601. doi:10.1038/ncb1722 .

Gross, DS, and WT Garrard. 1988. “Nuclease Hypersensitive Sites in Chromatin.” *Annual Review of Biochemistry* 57 (1): 159–197. doi:10.1146/annurev.bi.57.070188.001111 .

Harbison, Christopher T, Benjamin D Gordon, Tong Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, et al. 2004. “Transcriptional Regulatory Code of a Eukaryotic Genome.” *Nature* 431 (7004): 99–104. doi:10.1038/nature02800 .

Helwak, Aleksandra, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. 2013. “Mapping the Human MiRNA Interactome by CLASH Reveals Frequent Noncanonical Binding.” *Cell* 153 (3): 654–665. doi:10.1016/j.cell.2013.03.043 .

Hornstein, Eran, and Noam Shomron. 2006. “Canalization of Development by MicroRNAs.” *Nature Genetics* 38 (Suppl 6): S20–S24. doi:10.1038/ng1803 .

Hu, Ruozhen, Jared Wallace, Timothy J Dahlem, David Grunwald, and Ryan M O’Connell. 2013. “Targeting Human MicroRNA Genes Using Engineered Tal-Effector Nucleases (TALENs).” *PLoS ONE* 8 (5): e63074. doi:10.1371/journal.pone.0063074 .

Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. 2011. “Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-Seq.” *Genome Research* 21 (7): 1160–1167. doi:10.1101/gr.110882.110

.

Ivey, Kathryn N, and Deepak vastava. 2010. “MicroRNAs as Regulators of Differentiation and Cell Fate Decisions.” *Cell Stem Cell* 7 (1): 36–41. doi:10.1016/j.stem.2010.06.012 .

Iwafuchi-Doi, Makiko, and Kenneth S Zaret. 2016. “Cell Fate Control by Pioneer Transcription Factors.” *Development* 143 (11): 1833–1837. doi:10.1242/dev.133900 .

Johnson, RD, and M Jasin. 2001. “Double-Strand-Break-Induced Homologous Recombination in Mammalian Cells.” *Biochemical Society Transactions* 29 (2): 196–201. doi:10.1042/bst0290196 .

Johnston, Robert J, Sarah Chang, John F Etchberger, Christopher O Ortiz, and Oliver Hobert. 2005. “MicroRNAs Acting in a Double-Negative Feedback Loop to Control a Neuronal Cell Fate Decision.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (35): 12449–12454. doi:10.1073/pnas.0505530102 .

Johnston, Robert J, and Oliver Hobert. 2005. “A Novel C. Elegans Zinc Finger Transcription Factor, Lsy-2, Required for the Cell Type-Specific Expression of the Lsy-6 MicroRNA.” *Development* 132 (24): 5451–5460. doi:10.1242/dev.02163 .

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. “DNA-Binding Specificities of Human Transcription Factors.” *Cell* 152 (1–2): 327–339. doi:10.1016/j.cell.2012.12.009 .

Kay, Sabine, and Ulla Bonas. 2009. “How Xanthomonas Type III Effectors Manipulate the Host Plant.” *Current Opinion in Microbiology* 12 (1): 37–43. doi:10.1016/j.mib.2008.12.006 .

Kemmeren, Patrick, Katrin Sameith, Loes van de Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O’Duibhir, et al. 2014. “Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors.” *Cell* 157 (3): 740–752. doi:10.1016/j.cell.2014.02.054 .

Kent, James W. 2002. “BLAT—The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656–664. doi:10.1101/gr.229202 .

Kim, Yongsub, Jiyeon Kweon, Annie Kim, Jae Chon, Ji Yoo, Hye Kim, Sojung Kim, et al. 2013. “A Library of TAL Effector Nucleases Spanning the Human Genome.” *Nature Biotechnology* 31 (3): 251–258. doi:10.1038/nbt.2517 .

Kivioja, Teemu, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. 2012. “Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers.” *Nature Methods* 9 (1): 72–74. doi:10.1038/nmeth.1778 .

Klein, Allon M, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. 2015. “Droplet Barcoding for Single-

Cell Transcriptomics Applied to Embryonic Stem Cells." *Cell* 161 (5): 1187–1201. doi:10.1016/j.cell.2015.04.044 .

Kornberg, Roger D. 1977. "Structure of Chromatin." *Annual Review of Biochemistry* 46 (1): 931–954. doi:10.1146/annurev.bi.46.070177.004435 .

Kulakovskiy, Ivan V, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, et al. 2017. "HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP-Seq Analysis." *Nucleic Acids Research* 46 (D1): gkx1106-. doi:10.1093/nar/gkx1106 .

Kulkarni, Meghana M. 2011. "Current Protocols in Molecular Biology." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 25: 25B.10.1-25B.10.17. doi:10.1002/0471142727.mb25b10s94 .

Laganière, Josée, Geneviève Deblois, Céline Lefebvre, Alain R Bataille, François Robert, and Vincent Giguère. 2005. "Location Analysis of Estrogen Receptor α Target Promoters Reveals That FOXA1 Defines a Domain of the Estrogen Response." *Proceedings of the National Academy of Sciences of the United States of America* 102 (33): 11651–11656. doi:10.1073/pnas.0505575102 .

Lee, Hyung, Eunji Kim, and Jin-Soo Kim. 2010. "Targeted Chromosomal Deletions in Human Cells Using Zinc Finger Nucleases." *Genome Research* 20 (1): 81–89. doi:10.1101/gr.099747.109 .

Lee, Rosalind C, Rhonda L Feinbaum, and Victor Ambros. 1993. "The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14." *Cell* 75 (5): 843–854. doi:10.1016/0092-8674(93)90529-y .

Lee, Tong, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, et al. 2002. "Transcriptional Regulatory Networks in *Saccharomyces Cerevisiae*." *Science* 298 (5594): 799–804. doi:10.1126/science.1075090 .

Lewis, Benjamin P, Christopher B Burge, and David P Bartel. 2005. "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets." *Cell* 120 (1): 15–20. doi:10.1016/j.cell.2004.12.035 .

Li, Heng. 2011. "Tabix: Fast Retrieval of Sequence Features from Generic TAB-Delimited Files." *Bioinformatics* 27 (5): 718–719. doi:10.1093/bioinformatics/btq671 .

Li, Yan, Fay Wang, Jin-A Lee, and Fen-Biao Gao. 2006. "MicroRNA-9a Ensures the Precise Specification of Sensory Organ Precursors in Drosophila." *Genes & Development* 20 (20): 2793–2805. doi:10.1101/gad.1466306 .

Lin, Ying, Qiong Zhang, Hong-Mei Zhang, Wei Liu, Chun-Jie Liu, Qiubai Li, and An-Yuan Guo. 2015. "Transcription Factor and MiRNA Co-Regulatory Network Reveals Shared and

Specific Regulators in the Development of B Cell and T Cell." *Scientific Reports* 5 (1): 15215. doi:10.1038/srep15215 .

Lu, Leina, Liang Zhou, Eric Z Chen, Kun Sun, Peiyong Jiang, Lijun Wang, Xiaoxi Su, Hao Sun, and Huating Wang. 2012. "A Novel YY1-MiR-1 Regulatory Circuit in Skeletal Myogenesis Revealed by Genome-Wide Prediction of YY1-MiRNA Network." *PLoS ONE* 7 (2): e27596. doi:10.1371/journal.pone.0027596 .

Luger, Karolin, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. 1997. "Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution." *Nature* 389 (6648): 251–260. doi:10.1038/38444 .

Luria, SE, and lbrück. 1943. "Mutations of Bacteria from Virus Sensitivity to Virus Resistance." *Genetics* 28 (6): 491–511.

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–1214. doi:10.1016/j.cell.2015.05.002 .

Malik, Sohail, and Robert G Roeder. 2000. "Transcriptional Regulation through Mediator-like Coactivators in Yeast and Metazoan Cells." *Trends in Biochemical Sciences* 25 (6): 277–283. doi:10.1016/s0968-0004(00)01596-6 .

Martinez, Natalia J, Maria C Ow, Inmaculada M Barrasa, Molly Hammell, Reynaldo Sequerra, Lynn Doucette-Stamm, Frederick P Roth, Victor R Ambros, and Albertha Walhout. 2008. "A C. Elegans Genome-Scale MicroRNA Network Contains Composite Feedback Motifs with High Flux Capacity." *Genes & Development* 22 (18): 2535–2549. doi:10.1101/gad.1678608 .

Martínez-Antonio, Agustino, Sarath Janga, and Denis Thieffry. 2008. "Functional Organisation of Escherichia Coli Transcriptional Regulatory Network." *Journal of Molecular Biology* 381 (1): 238–247. doi:10.1016/j.jmb.2008.05.054 .

Mathelier, Anthony, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, et al. 2016. "JASPAR 2016: A Major Expansion and Update of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 44 (D1): D110–D115. doi:10.1093/nar/gkv1176 .

Milo, R, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* 298 (5594): 824–827. doi:10.1126/science.298.5594.824 .

Milo, Ron, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. "Superfamilies of Evolved and Designed Networks." *Science* 303 (5663): 1538–1542. doi:10.1126/science.1089167 .

Molin, Alessandra, and Barbara Camillo. 2018. “How to Design a Single-Cell RNA-Sequencing Experiment: Pitfalls, Challenges and Perspectives.” *Briefings in Bioinformatics*. doi:10.1093/bib/bby007 .

Montague, Tessa G, José M Cruz, James A Gagnon, George M Church, and Eivind Valen. 2014. “CHOPCHOP: A CRISPR/Cas9 and TALEN Web Tool for Genome Editing.” *Nucleic Acids Research* 42 (W1): W401–W407. doi:10.1093/nar/gku410 .

Moscou, Matthew J, and Adam J Bogdanove. 2009. “A Simple Cipher Governs DNA Recognition by TAL Effectors.” *Science* 326 (5959): 1501–1501. doi:10.1126/science.1178817 .

Moss, Eric G, Rosalind C Lee, and Victor Ambros. 1997. “The Cold Shock Domain Protein LIN-28 Controls Developmental Timing in C. Elegans and Is Regulated by the Lin-4 RNA.” *Cell* 88 (5): 637–646. doi:10.1016/s0092-8674(00)81906-6 .

Mukherji, Shankar, Margaret S Ebert, Grace XY Zheng, John S Tsang, Phillip A Sharp, and Alexander van Oudenaarden. 2011. “MicroRNAs Can Generate Thresholds in Target Gene Expression.” *Nature Genetics* 43 (9): 854–859. doi:10.1038/ng.905 .

Murchison, Elizabeth P, Paula Stein, Zhenyu Xuan, Hua Pan, Michael Q Zhang, Richard M Schultz, and Gregory J Hannon. 2007. “Critical Roles for Dicer in the Female Germline.” *Genes & Development* 21 (6): 682–693. doi:10.1101/gad.1521307 .

Myers, Lawrence C, and Roger D Kornberg. 2000. “MEDIATOR OF TRANSCRIPTIONAL REGULATION.” *Biochemistry* 69 (1): 729–749. doi:10.1146/annurev.biochem.69.1.729 .

Neph, Shane, Scott M Kuehn, Alex P Reynolds, Eric Haugen, Robert E Thurman, Audra K Johnson, Eric Rynes, et al. 2012. “BEDOPS: High-Performance Genomic Feature Operations.” *Bioinformatics* 28 (14): 1919–1920. doi:10.1093/bioinformatics/bts277 .

Neph, Shane, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. 2012. “Circuitry and Dynamics of Human Transcription Factor Regulatory Networks.” *Cell* 150 (6): 1274–1286. doi:10.1016/j.cell.2012.04.040 .

Newburger, Daniel E, and Martha L Bulyk. 2009. “UniPROBE: An Online Database of Protein Binding Microarray Data on Protein–DNA Interactions.” *Nucleic Acids Research* 37 (suppl_1): D77–D82. doi:10.1093/nar/gkn660 .

Niu, Zhivy, Ankang Li, Shu X Zhang, and Robert J Schwartz. 2007. “Serum Response Factor Micromanaging Cardiogenesis.” *Current Opinion in Cell Biology* 19 (6): 618–627. doi:10.1016/j.ceb.2007.09.013 .

PTASHNE, MARK. 1967. “Specific Binding of the λ Phage Repressor to λ DNA.” *Nature* 214 (5085): 232–234. doi:10.1038/214232a0 .

Ptashne, Mark. 2005. “Regulation of Transcription: From Lambda to Eukaryotes.” *Trends in*

Biochemical Sciences 30 (6): 275–279. doi:10.1016/j.tibs.2005.04.003 .

Ptashne, Mark. 2013. “Epigenetics: Core Misconcept.” *Proceedings of the National Academy of Sciences* 110 (18): 7101–7103. doi:10.1073/pnas.1305399110 .

Ramsköld, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, et al. 2012. “Full-Length MRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells.” *Nature Biotechnology* 30 (8): 777–782. doi:10.1038/nbt.2282 .

Ravi, Rupesh, Kendra Walton, and Mahdieh Khosroheidari. 2018. “Disease Gene Identification, Methods and Protocols.” *Methods in Molecular Biology (Clifton, N.J.)* 1706: 223–232. doi:10.1007/978-1-4939-7471-9_12 .

Reece-Hoyes, John S, Alos Diallo, Bryan Lajoie, Amanda Kent, Shaleen Shrestha, enath Kadreppa, Colin Pesyna, Job Dekker, Chad L Myers, and Albertha JM Walhout. 2011. “Enhanced Yeast One-Hybrid Assays for High-Throughput Gene-Centered Regulatory Network Mapping.” *Nature Methods* 8 (12): 1059–1064. doi:10.1038/nmeth.1748 .

Reinhart, Brenda J, Frank J Slack, Michael Basson, Amy E Pasquinelli, Jill C Bettinger, Ann E Rougvie, Robert H Horvitz, and Gary Ruvkun. 2000. “The 21-Nucleotide Let-7 RNA Regulates Developmental Timing in Caenorhabditis Elegans.” *Nature* 403 (6772): 901–906. doi:10.1038/35002607 .

Rie, Derek, Imad Abugessaisa, Tanvir Alam, Erik Arner, Peter Arner, Haitham Ashoor, Gaby Åström, Magda Babina, Nicolas Bertin, and Maxwell A Burroughs. 2017. “An Integrated Expression Atlas of MiRNAs and Their Promoters in Human and Mouse.” *Nature Biotechnology* 35 (9): 872. doi:10.1038/nbt.3947 .

Roberts, Adam, Harold Pimentel, Cole Trapnell, and Lior Pachter. 2011. “Identification of Novel Transcripts in Annotated Genomes Using RNA-Seq.” *Bioinformatics* 27 (17): 2325–2329. doi:10.1093/bioinformatics/btr355 .

Roberts, Adam, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. 2011. “Improving RNA-Seq Expression Estimates by Correcting for Fragment Bias.” *Genome Biology* 12 (3): R22. doi:10.1186/gb-2011-12-3-r22 .

Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. “Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing.” *Nature Methods* 4 (8): 651–657. doi:10.1038/nmeth1068 .

Sakuma, Tetsushi, and Takashi Yamamoto. 2016. “TALENs, Methods and Protocols.” *Methods in Molecular Biology (Clifton, N.J.)* 1338: 61–70. doi:10.1007/978-1-4939-2932-0_6 .

Schornack, Sebastian, Annett Meyer, Patrick Römer, Tina Jordan, and Thomas Lahaye. 2006.

"Gene-for-Gene-Mediated Recognition of Nuclear-Targeted AvrBs3-like Bacterial Effector Proteins." *Journal of Plant Physiology* 163 (3): 256–272. doi:10.1016/j.jplph.2005.12.001 .

Sokol, Nicholas S, and Victor Ambros. 2005. "Mesodermally Expressed Drosophila MicroRNA-1 Is Regulated by Twist and Is Required in Muscles during Larval Growth." *Genes & Development* 19 (19): 2343–2354. doi:10.1101/gad.1356105 .

Stark, Alexander, Julius Brennecke, Natascha Bushati, Robert B Russell, and Stephen M Cohen. 2005. "Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3′UTR Evolution." *Cell* 123 (6): 1133–1146. doi:10.1016/j.cell.2005.11.023 .

Stegle, Oliver, Sarah A Teichmann, and John C Marioni. 2015. "Computational and Analytical Challenges in Single-Cell Transcriptomics." *Nature Reviews Genetics* 16 (3): 133–145. doi:10.1038/nrg3833 .

Stergachis, Andrew B, Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L Paige, Benjamin Vernot, et al. 2013. "Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes." *Cell* 154 (4): 888–903. doi:10.1016/j.cell.2013.07.020 .

Stork, Cheryl, and Sika Zheng. 2016. "Genome-Wide Profiling of RNA-Protein Interactions Using CLIP-Seq." *Methods in Molecular Biology (Clifton, N.J.)* 1421: 137–151. doi:10.1007/978-1-4939-3591-8_12 .

Streubel, Jana, Christina Blücher, Angelika Landgraf, and Jens Boch. 2012. "TAL Effector RVD Specificities and Efficiencies." *Nature Biotechnology* 30 (7): 593–595. doi:10.1038/nbt.2304 .

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell* 126 (4): 663–676. doi:10.1016/j.cell.2006.07.024 .

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. "MRNA-Seq Whole-Transcriptome Analysis of a Single Cell." *Nature Methods* 6 (5): 377–382. doi:10.1038/nmeth.1315 .

Thompson, Dawn, Aviv Regev, and Sushmita Roy. 2015. "Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution." *Annual Review of Cell and Developmental Biology* 31 (1): 399–428. doi:10.1146/annurev-cellbio-100913-012908 .

Trapnell, Cole, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. 2013. "Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq." *Nature Biotechnology* 31 (1): 46–53. doi:10.1038/nbt.2450 .

Trapnell, Cole, Lior Pachter, and Steven L Salzberg. 2009. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics* 25 (9): 1105–1111. doi:10.1093/bioinformatics/btp120 .

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–515. doi:10.1038/nbt.1621 .

Tsang, John, Jun Zhu, and Alexander van Oudenaarden. 2007. “MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals.” *Molecular Cell* 26 (5): 753–767. doi:10.1016/j.molcel.2007.05.018 .

Urnov, Fyodor D. 2018. “Genome Editing B.C. (Before CRISPR): Lasting Lessons from the ‘Old Testament.’” *The CRISPR Journal* 1 (1): 34–46. doi:10.1089/crispr.2018.29007.fyu .

Urnov, Fyodor D, Jeffrey C Miller, Ya-Li Lee, Christian M Beausejour, Jeremy M Rock, Sheldon Augustus, Andrew C Jamieson, Matthew H Porteus, Philip D Gregory, and Michael C Holmes. 2005. “Highly Efficient Endogenous Human Gene Correction Using Designed Zinc-Finger Nucleases.” *Nature* 435 (7042): 646–651. doi:10.1038/nature03556 .

Urnov, Fyodor D, Edward J Rebar, Michael C Holmes, Steve H Zhang, and Philip D Gregory. 2010. “Genome Editing with Engineered Zinc Finger Nucleases.” *Nature Reviews Genetics* 11 (9): 636–646. doi:10.1038/nrg2842 .

Vaquerizas, Juan M, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. 2009. “A Census of Human Transcription Factors: Function, Expression and Evolution.” *Nature Reviews Genetics* 10 (4): 252–263. doi:10.1038/nrg2538 .

Vierstra, Jeff, Andreas Reik, Kai-Hsin Chang, Sandra Stehling-Sun, Yuanyue Zhou, Sarah J Hinkley, David E Paschon, et al. 2015. “Functional Footprinting of Regulatory DNA.” *Nature Methods* 12 (10): 927–930. doi:10.1038/nmeth.3554 .

WADDINGTON, CH. 1942. “CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS.” *Nature* 150 (3811): 563–565. doi:10.1038/150563a0 .

Walhout, Albertha. 2006. “Unraveling Transcription Regulatory Networks by Protein–DNA and Protein–Protein Interaction Mapping.” *Genome Research* 16 (12): 1445–1454. doi:10.1101/gr.5321506 .

Walker, Emily, Minako Ohishi, Ryan E Davey, Wen Zhang, Paul A Cassar, Tetsuya S Tanaka, Sandy D Der, et al. 2007. “Prediction and Testing of Novel Transcriptional Networks Regulating Embryonic Stem Cell Self-Renewal and Commitment.” *Cell Stem Cell* 1 (1): 71–86. doi:10.1016/j.stem.2007.04.002 .

Wang, Huating, Ramiro Garzon, Hao Sun, Katherine J Ladner, Ravi Singh, Jason Dahlman, Alfred Cheng, et al. 2008. “NF-KB–YY1–MiR-29 Regulatory Circuitry in Skeletal Myogenesis and Rhabdomyosarcoma.” *Cancer Cell* 14 (5): 369–381. doi:10.1016/j.ccr.2008.10.006 .

Wang, Yangming, Rostislav Medvid, Collin Melton, Rudolf Jaenisch, and Robert Blelloch. 2007. “DGCR8 Is Essential for MicroRNA Biogenesis and Silencing of Embryonic Stem Cell Self-Renewal.” *Nature Genetics* 39 (3): 380–385. doi:10.1038/ng1969 .

Wightman, Bruce, Ilho Ha, and Gary Ruvkun. 1993. “Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern Formation in C. Elegans.” *Cell* 75 (5): 855–862. doi:10.1016/0092-8674(93)90530-4 .

Wingender, E, P Dietze, H Karas, and R Knüppel. 1996. “TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites.” *Nucleic Acids Research* 24 (1): 238–241. doi:10.1093/nar/24.1.238 .

Yun, Kyuson, and Barbara Wold. 1996. “Skeletal Muscle Determination and Differentiation: Story of a Core Regulatory Network and Its Context.” *Current Opinion in Cell Biology* 8 (6): 877–889. doi:10.1016/s0955-0674(96)80091-3 .

Zhang, Jiajie, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. 2014. “PEAR: A Fast and Accurate Illumina Paired-End ReAd MergeR.” *Bioinformatics* 30 (5): 614–620. doi:10.1093/bioinformatics/btt593 .