DIVIDE AND CONQUER: STUDYING MICRORNA REGULATORY NETWORKS BY SEPARATION OF MICRORNA DIRECT TARGETS FROM INDIRECT CHANGES

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Ravi Krishnakant Patel
May 2020

DIVIDE AND CONQUER: STUDYING MICRORNA REGULATORY NETWORKS BY SEPARATION OF MICRORNA DIRECT TARGETS FROM INDIRECT CHANGES

Ravi Krishnakant Patel, Ph. D.
Cornell University 2020

Post-transcriptional regulation of gene expression is essential for cell function. MicroRNAs (miRNAs), small (~22nt) non-coding RNAs, are negative regulators of post-transcriptional expression. MiRNAs recruit the RNA induced silencing complex (RISC), which includes an Argonaute (AGO) protein, to partially complementary target sites located in 3' untranslated regions (UTRs), and induce accelerated mRNA decay and translational inhibition of target mRNAs. Each miRNA has potential to regulate many hundreds of mRNAs via this pathway. Although, miRNAs typically elicit a modest effect per mRNA, hundreds of such small changes result in a substantial cumulative impact on the transcriptome. Additionally, miRNA regulatory networks incorporate master regulators such as transcription factors downstream of miRNAs, enabling miRNAs to trigger substantial changes in gene regulatory networks. Understanding the biological impact of miRNAs requires knowledge of their targets, and robust distinction of miRNA direct targets from cascading downstream regulatory changes remains challenging.

In this work, I developed a simple experimental approach to robustly uncouple post-transcriptional and transcriptional changes using RNA-seq and Precision Run-On sequencing (PRO-seq), a method for profiling actively transcribing RNA polymerases. The net change in mRNA abundance results from changes in synthesis

and decay. I demonstrated that by subtracting the changes in mRNA synthesis (PRO-seq) from the changes in mRNA abundance (RNA-seq), a robust estimate of post-transcriptional regulation by miRNAs could be derived. I refer to this approach as CARP: Combined Analysis of RNA-seq and PRO-seq. Using CARP, I successfully separated true direct targets of specific miRNAs from the downstream indirect changes, which I validated using orthogonal assays such as Argonaute eCLIP-seq and ribosome profiling. Additionally, CARP analysis revealed that the majority of miRNAs used in the study elicited sizable indirect targeting at both transcriptional and post-transcriptional levels, which are often disregarded. Using motif enrichment analysis, I found candidate transcription factors underlying the miRNA-mediated indirect regulation at transcriptional level. I also demonstrated that CARP facilitates effective dissection of complex regulatory changes triggered by miRNAs. Furthermore, my analysis revealed that many miRNAs elicit discernible repression of target sites located in open reading frames (ORFs); the significance of ORF target sites is a potentially important aspect of miRNA biology, but the extent to which it occurs has been controversial. My data demonstrated that while ORF sites to certain miRNAs often mediate subtle repression, their likely role is in assisting the miRNA-mediated regulation of weaker 3'UTR sites to collectively elicit significant post-transcriptional repression of the target mRNA.

Overall, the tools I developed in my graduate work facilitate robust distinction of direct target from indirect regulatory changes aiding in the study of miRNA regulatory networks at the systems-level. Finally, I apply CARP and other genomic approaches to better understand biological roles of miRNAs in the immune response of CD8+ T-cells in mouse models.

BIOGRAPHICAL SKETCH

Ravi Krishnakant Patel was born in Nadiad, a small town in Gujarat, India. He grew up in a joint family consisting of eight members, including a younger brother and two younger cousins. He was fond of mathematics from the young age, and being the eldest, he had a responsibility of teaching mathematics to his brother and cousins while growing up. However, his passion for mathematics gradually died off due to lack of proper exposure in the small town of Nadiad. During the high school, he became interested in biology, genetics in particular. After all, this was the time when *Bacillus thuringiensis* (Bt) cotton was introduced in India; the excitement in the news media about the first genetically modified crop in India influenced his fascination towards genetics.

Ravi later went on to study biology for his undergraduate degree at Sardar Patel University, where he was exposed to molecular biology and biotechnology. During his master's studies in bioinformatics, he was introduced to computer programming, including mathematics by his computer science teacher. That is when Ravi reinvented his passion for mathematics. He found his niche in bioinformatics, the intersection of biology, mathematics and computer science. He also worked as a freelance computer programmer to pay for his college fees, which helped him learn invaluable computer programming skills.

After completing his studies, Ravi became a lecturer at the Maharaja Sayajirao

University (MSU) of Baroda, where he taught bioinformatics and computer programming to the master's students for a year. Later, he joined Dr Jain's laboratory at Jawaharlal Nehru University, where he worked on Chickpea genome annotation and transcriptome analysis. His work in the Jain lab resulted into several publications at international peer-reviewed journals. After working for almost three years in the Jain lab, Ravi joined the Sundar lab at UC Davis and worked on the discovery of tissue-specific long non-coding RNAs in rice using RNA-seq. By this time, Ravi had developed a strong interest in studying different modes of gene expression regulation.

In 2013, Ravi came to Cornell University to pursue his PhD in Genetics, Genomics & Development program. He joined the lab of Dr. Andrew Grimson to study the miRNA-mediated regulation of gene expression. During his graduate studies, Ravi has developed skills of a molecular biologist and has greatly enhanced his skills as a computational biologist. He hopes to apply these skills in his future work as a scientist to understand the modes of gene regulation that shape the biology of humans.

This dissertation is dedicated to the Almighty, my parents and Dr. Andrew Grimson.

ACKNOWLEDGMENTS

Dr. Andrew Grimson; you have been a phenomenal advisor and an exceptional mentor. Thank you for providing me opportunities to work on many different exciting research questions. I am also very grateful to you for your guidance, understanding and patience and for encouraging me when I felt low with my research.

My parents; thank you for your support and patience throughout my Ph.D. work and my life. Every success in my life is a result of your countless sacrifices and your unconditional love for me.

My wife; thank you for being on my side through last few intense months, for making delicious food for me and for making sure that I stay healthy. Your arrival in my life has brought so much joy and happiness.

My brother and sister-in-law; thank you for your support, love and faith in me.

Dr. Upendra Patel (uncle); you have been my guru, in professional and personal life. Thank you for your guidance and encouragement in tough times during my Ph.D.

Dr. John Lis; thank you for your advice and suggestions on my research work. You have inspired me to ask creative questions using my data.

Dr. Haiyuan Yu; thank you for all your suggestions and help with my research work.

Dr. Jen Grenier; you have been another advisor at Cornell, after Andrew. Thank you for all your help and support, and for providing me new opportunities to work on interesting research projects. You made my work much smoother.

Elizabeth Fogarty; thank you for being a very good friend and for all your support with my research and otherwise.

Dr. Rene Geissler; you have been my lab guru; I learned majority of my experimental skills from you. Thank you for all your help.

Dr. Stephanie Hilz; thank you for being my mentor and a very good friend and for all your help to make my journey of Ph.D. smoother.

Dr. Alfred Simkin; I have had such interesting conversations with you about science, computer programming and spirituality. Thank you for being a great colleague and a wonderful friend.

Jessica West and Ciarán Daly; thank you for all those very interesting scientific and intellectual conversations; I have learned so much while talking with you guys. And of course, thank you for so many fun-filled ping-pong games.

Former and current Grimson lab members; thank you for all your suggestions, feedback and advice with my research work, practice talks and papers. Your advice and constructive criticism have helped me become a better scientist.

Dr. Brian Rudd and the Rudd lab members; thank you for giving me opportunities to work on exciting research projects in immune system. Thank you being so helpful and patient with me, and for teaching a lot of immunology.

Dr. John Schimenti and Priti Singh; I had a fun collaboration with you. Thank you for giving me that opportunity.

Other members of the Cornell community whose role have been very important include, RNA Sequencing Core staff, BRC and Genomic core staff, and The MBG administrative staff.

The work presented in this dissertation was supported by funding from National Institute of Health (NIH; R01GM105668 and P50HD076210 (Core B)) and a Cornell Vertebrate Genomics Seed grant.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

6mer site	match to miRNA nucleotides 2-7
7mer-A1 site	6mer site with an A at the position opposite of nucleotide 1 of miRNA
7mer-m8 site	match to miRNA nucleotides 2-8
8mer site	7mer-m8 site with an A at the position opposite of nucleotide 1 of miRNA
Ago	argonaute
CARP	combined analysis of RNA-seq and PRO-seq
CDS	coding sequence
CLIP	cross-linking and immunoprecipitation
DGCR8	DiGeorge Syndrome Critical Region 8
edgeR	empirical Analysis of Digital Gene Expression Data in R
FACS	fluorescence activated cell sorting
FLP	flippase
gB	HSV1 Kb-restricted epitope gB498-505
gB-TI	transgenic mouse line in which all TCR receptors are specific to gB
GFP	green fluorescent protein
miRNA	microRNA
MPEC	memory precursor effector cell
nt	nucleotide
ORF	open reading frame
PCA	principal component analysis
PCR	polymerase chain reaction
poly(A)	poly-adenosine
PRO-seq	precision run-on sequencing
RBP	RNA-binding protein
RPM	Reads per million
SLEC	short-lived effector cell
TCR	T cell receptor
TE	translational efficiency
TN	true naïve
TNRC6	tri-Nucleotide Repeat Containing 6
UTR	untranslated region
VM	virtual memory
WT	wild-type

CHAPTER 1

Introduction

Post-transcriptional gene regulation

Regulation of gene expression is essential for cell function. The advent of genome-wide analysis tools has facilitated new insights into fundamental aspects of gene regulatory programs in variety of organisms. It has become clear that organismal complexity is not just a function of number of genes; mammals and nematodes have similar number of genes, yet nematodes have far fewer cell types (Straalen & Roelofs, 2012). A major portion of the complexity in higher eukaryotes derives from complex gene regulatory circuitries. While transcriptional regulation has been implicated as a major mode of gene regulation, the post-transcriptional regulation of gene expression plays an important role in various cellular processes, including development, metabolism and cancer progression, providing a critical layer of regulation that allow fine-tuning of gene expression (Schaefke *et al*, 2018). Between the synthesis and degradation, messenger RNAs (mRNAs) go through a variety of post-transcriptional processes, each of which can be regulated, and which include covalent modifications, nuclear export, sub-cellular localization, translation and decay. An mRNA is comprised of three regions, 5' untranslated region (5'UTR), open reading frame (ORF) and 3' untranslated region (3'UTR). While the ORFs code for proteins, and hence are occupied by the translational machinery, the UTRs are relatively more accessible and therefore provide vital opportunities for post-transcriptional regulation through the cis-regulatory elements residing in them. Given their larger lengths

(~1300nt in humans) and conservation of primary sequence, the 3'UTRs are rich in regulatory information (Zhao *et al*, 2011), and thus are a major site for post-transcriptional regulation in animals.

MicroRNAs and their biological importance

MicroRNAs, a class of small non-coding RNAs (21-24nt) are negative regulators of post-transcriptional expression, with widespread impact on the transcriptome. Several hundred miRNAs have been discovered in humans, of which many are conserved in other animals (Bartel, 2009). Each miRNA can potentially regulate hundreds of mRNAs, hence, collectively, miRNAs regulate nearly all mRNAs in humans, suggesting that essentially all biological processes are under control of miRNA-mediated regulation, including development. Indeed, the loss-of-function of miRNAs results in a variety of phenotypes, including disease conditions and developmental defects in mouse knock-out models and humans. Dysregulation of miRNAs has also been implicated in disorders like autoimmunity and cancer (Huang *et al*, 2011), whereas loss of some miRNAs is lethal. Below, I will review few important findings depicting the biological importance of miRNAs.

Some miRNAs are essential for the viability of an organism, for example, deletion of one of the duplicated copies of miR-1 results in 50% embryonic lethality with cardiac defects in mice (Zhao *et al*, 2007). Similarly, the loss of let-7 in worms and bantam in

flies is lethal for those animals. Moreover, A germline deletion of miRNA cluster, miR-17~92, results in skeletal and growth defects in humans (de Pontual *et al*, 2011). Similarly, a point mutation in the human miRNA, miR-96, which impairs production of mature miRNAs, results in autosomal dominant, progressive hearing loss in humans (Mencía *et al*, 2009). Furthermore, our lab's work in collaboration with the Rudd lab showed that the loss of either miR-29 or let-7 in adult CD8+ T-cells, a type of immune cell, alters the fate of the CD8+ T-cells in infection, mimicking cells of fetal origin (discussed in detail in chapter 3). Additionally, several miRNAs have been shown to act as tumor suppressors or oncogenes (Huang *et al*, 2011). These findings strongly suggest crucial role of miRNAs in human health.

MicroRNA biogenesis

MiRNAs are typically expressed as rather long primary transcripts containing a hairpin structure which contains the mature miRNAs. The hairpin is processed by the microprocessor complex, composed of Drosha and DiGeorge Syndrome Critical Region 8 (DGCR8) proteins, producing ~50nt pre-miRNA hairpin. Upon nuclear export, the pre-miRNA hairpin is further processed using Dicer to remove the apical loop, producing ~22nt double-stranded mature miRNAs. One strand of the duplex loads on Argonaute (AGO1-4) protein of RNA-induced silencing complex (RISC), which is called a guide strand as it guides the RISC to target mRNAs for post-

transcriptional repression, whereas the other strand gets degraded. It is important that the guide strand is processed precisely; a difference of even a single nucleotide on the 5' terminus of the guide strand alters the miRNA "seed" region (nucleotides 2-7) responsible for target recognition, leading to recognition of completely different repertoire of targets (the miRNA seed is discussed in more detail below). In addition to a miRNA and AGO, the mature RISC complex includes TNRC6 (commonly known as GW182), which plays a crucial role for miRNA-mediated post-transcriptional regulation (see below) (Kawamata & Tomari, 2010; Elkayam *et al*, 2017).

Mode of action

MicroRNAs negatively regulate gene expression via two pathways: accelerated mRNA decay and translational repression (Carthew & Sontheimer, 2009; Bartel, 2009). The miRNA-mediated mRNA decay takes place in two sequential steps: deadenylation and decapping (Chen *et al*, 2009). The recruitment of RISC to effective target sites in a transcript promotes recruitment of the decay machinery. This process is orchestrated by TNRC6 proteins, which recruit CCR4-NOT and PAN2-PAN3 deadenylation complexes to the target mRNAs, marking this step the first step of the decay process. The proteins in the deadenylation complexes catalyze trimming of poly(A) tail, exposing the 3' end of target mRNA for 3'-5' exonucleolytic decay. Deadenylation is

followed by removal of the 5' cap, which then makes 5' end of mRNAs accessible for 5'-3' degradation (Chen *et al*, 2009). Together, these processes expose both ends of transcripts leading to enhanced decay of miRNA targets (Chen & Shyu, 2011).

While the mechanism of miRNA-mediated mRNA decay is well understood, the area of research delineating mechanism of translational repression remains less developed and controversial (Wilczynska & Bushell, 2015). Various mechanisms have been proposed for miRNA-mediated inhibition of translation, including interruption in translational initiation and elongation, degradation of nascent polypeptide and ribosome dropoff (Cannell *et al*, 2008; Fabian & Sonenberg, 2012). However, recent experiments from multiple groups indicate that the translational initiation is a primary target for miRNA-mediated translational repression (Pillai *et al*, 2005; Iwasaki *et al*, 2009; Mathonnet *et al*, 2007).

Determining relative contributions and temporal precedence of these two pathways (accelerated decay and translational repression) has been an area of active research. A series of studies from different groups have demonstrated that translational repression precedes miRNA-mediated accelerated mRNA decay and that translational repression is not dependent on mRNA destabilization or poly(A) tail removal (Djuranovic *et al*, 2012; Mathonnet *et al*, 2007; Bazzini *et al*, 2012). Additionally, reports claim that miRNA-mediated translational inhibition is in fact a pre-requisite for target degradation (Wilczynska & Bushell, 2015). While translational inhibition is triggered

rapidly after introduction of miRNAs, the magnitude of repression caused by reduced translation is minimal. Importantly, the repression by translational inhibition does not last very long; the mRNA decay pathway takes over, likely within hours of miRNA expression (Eichhorn *et al*, 2014). Furthermore, recent studies established that mRNA destabilization dominates the overall miRNA-mediated repression (Eichhorn *et al*, 2014; Guo *et al*, 2010). However, it is important to acknowledge that majority of these experiments were performed using transient transfection of miRNAs in high excess, likely more than an order of magnitude higher compared to the endogenous expression of miRNAs. Thus, experiments performed using conditions that more closely approximate in vivo parameters are needed to corroborate these observations.

While most of the miRNA:target interactions result in post-transcriptional repression, some reports claim that miRNAs can activate translation in rare cellular contexts (Vasudevan *et al*, 2007), for example, coupling of AGO with fragile X mental retardation–related protein 1 (FXR1) in miRNA-dependent manner elicits translational upregulation of target genes upon cell cycle arrest (Vasudevan *et al*, 2007). Nevertheless, the prevalence of miRNA-mediated upregulation appears to be minimal, therefore, I will exclude these anomalies in the remainder of this dissertation.

Canonical miRNA target sites

MicroRNAs regulate post-transcriptional expression via recruitment of RISC to target messages in sequence-specific manner. The target recognition is facilitated via

Watson-Crick base-pairing between a short sequence on the 5' end of miRNAs, nucleotides 2 – 7 of a miRNA (the miRNA "seed"), and the target site (Lai, 2002; Lewis *et al*, 2003; Krek *et al*, 2005). The perfect match between the miRNA seed and the target site has been shown to be critical for miRNA-mediated repression; even a single nucleotide substitution often leads to loss of repression of target mRNAs (Kloosterman *et al*, 2004; Doench & Sharp, 2004; Brennecke *et al*, 2005). The critical nature of seed pairing is also strongly supported by the following findings from structural analysis of miRNA-bound AGO: when the miRNA is loaded on AGO, initially, only the positions 2 – 5 of miRNA (sub-seed region) are exposed and are available for pairing with the target (Nakanishi *et al*, 2012; Elkayam *et al*, 2012). Pairing of the sub-seed region with targets triggers conformational changes, exposing nucleotides 6 to 8 and 13 to 16 for further pairing (Schirle *et al*, 2014), supporting the critical role of seed-pairing. The pairing with nucleotides 9 through 12 negatively impacts the function of miRNA, because pairing with nucleotides 9 to 12 requires extensive conformational change in AGO:miRNA complex to accommodate for the helical turn in miRNA resulted from extended pairing, which would disrupt the interaction of AGO and miRNA.

The target sites that match with miRNA seed and are located in 3'UTRs are called canonical target sites; non-canonical sites are discussed later. The functional canonical sites display the properties described below. Besides a match to the seed region,

pairing with nucleotide 8 of miRNA has positive impact, whereas pairing with nucleotides 9-12 negatively correlates with repression as explained above based on the structural studies of AGO. While the first nucleotide of a miRNA (m1) is typically embedded in AGO, and hence is unimportant for target recognition, AGO prefers nucleotide A in the target site across from the first nucleotide of miRNA (t1). This preference arises from the solvated surface pocket of AGO, which specifically binds to adenine (A) base of the target site at position 1 (Schirle *et al*, 2015). The base-pairing with the 3' end of a miRNA (3'-supplementary base pairing with miRNA nucleotides 13 to 16) has been shown to stabilize the pairing with seed, although these sites comprise less than 5% of the seed-matching sites (Grimson *et al*, 2007; Friedman *et al*, 2009; Lewis *et al*, 2005; Brennecke *et al*, 2005). Additionally, the 3'-supplementary pairing is beneficial only for certain miRNAs (McGeary *et al*, 2018), suggesting that the 3'-supplementary sites are rare. Based on these pairing preferences, the canonical sites are grouped in four site-types, as listed here in the decreasing order of conservation and their efficacy in mediating repression: (i) 8mer, (ii) 7mer-m8, (iii) 7mer-A1, and (iv) 6mer (Lewis *et al*, 2005). The 8mer sites base-pair with nucleotides 2-8 of the miRNA and contain an A at the position 1 (t1) across from the miRNA nucleotide 1 (m1). These sites typically mediate the strongest repression of all canonical sites. The 7mer-m8 sites include pairing with the miRNA nucleotides 2-8, whereas the 7mer-A1 sites contain nucleotide A at position 1 (t1), in addition to pairing with nucleotides 2-7 of the miRNA; these sites elicit intermediate repression.

The 6mer sites pair with nucleotides 2-7 of the miRNA and are least effective among these four site-types.

While the regulation of predicted targets generally correlates with the type of site that their 3'UTRs harbor, the canonical sites do not always elicit repression. Importantly, those canonical sites that do confer repression, the degree of regulation varies drastically depending on the 3'UTR context, suggesting the critical role of context in which the sites reside. A seminal work systematically characterized the 3'UTR context of functional miRNA sites (Grimson *et al*, 2007), which was pivotal in enhancing the understanding of miRNA-mediated regulation. The major properties of the 3'UTR context discovered included: (i) the AU content around seed-matching sites: the *local AU content* correlated positively with the repression of targets, perhaps because the lower thermodynamic stability of A-U pairing prevents the nucleotides around the site from participating in secondary structure formation, keeping the site accessible for miRNA binding. (ii) the location of sites in 3'UTR: sites towards the end of the 3'UTR tend to work better and sites too close to the stop codon affects miRNA activity negatively, perhaps due to hindrance from translating ribosomes. (iii) cooperative sites: two adjacent sites (13-35nt apart from each other) tend to act synergistically (Grimson *et al*, 2007; Saetrom *et al*, 2007; Broderick *et al*, 2011; Rinck *et al*, 2013). Consistent with cooperative effect of adjacent sites, a recent study demonstrated that one molecule of TNRC6 can interact with up to three AGO molecules, allowing

TNRC6 to bind multiple AGO molecules bound at adjacent sites and facilitating synergistic effect of adjacent sites by stabilizing the whole RISC (Elkayam *et al*, 2017). The study by Grimson et al was followed by a series of reports over the years recapitulating its findings (Baek *et al*, 2008; Agarwal *et al*, 2015; Kim *et al*, 2016; Agarwal *et al*, 2018; McGeary *et al*, 2018) and discovering additional new properties of functional sites (Garcia *et al*, 2011; Arvey *et al*, 2010; Ui-Tei *et al*, 2008; Agarwal *et al*, 2018; Saetrom *et al*, 2007). For example, Garcia et al found that the stronger seed pairing stability – the thermodynamic stability of pairing between seed and site – stabilizes the binding of AGO, leading to stronger repression (Garcia *et al*, 2011). These properties of canonical sites and the 3'UTR context have been proved pivotal in computational prediction of miRNA targets.

Non-canonical target sites

ORF sites

Traditionally, canonical target sites, 3'UTR sites with seed-match, are thought to make up most of the functional target site repertoire, whereas the target sites located in open reading frame (ORF) or 5'UTRs have been believed to be non-functional because of hindrance from translating ribosomes (Gu *et al*, 2009). However, studies profiling target sites that crosslink with AGO have found that ORF sites are bound by AGO and are as abundant as 3'UTR sites. However, the functionality of these sites has been controversial. While studies involving analysis of AGO-associated transcripts

assert that the AGO binding implicates function, most of these sites are demonstrated to mediate weak repression, if any, of target abundance and no change in translation. For example, when Fang and Rajewsky analyzed transcriptomic data from ten miRNA mis-expression experiments, they found that the experimentally identified ORF sites were not functional (Fang & Rajewsky, 2011). Similarly, Agarwal et al found when assessing ORF sites for hundreds of miRNAs that the ORF sites have minimal evidence of regulation, if any (Agarwal *et al*, 2015). However, since such studies assess aggregated effect of multiple miRNAs, they are likely to miss miRNA-dependent activity of ORF sites. Furthermore, additional studies claim that ORF sites primarily affect translation and not message stability (Hausser *et al*, 2013; Ni & Leng, 2015; Brümmer & Hausser, 2014). As you will see in chapter 2, I observed striking activity of ORF sites for specific miRNAs, and I did not observe any change in translation for messages containing ORF sites.

Seedless sites

Another group of non-canonical sites includes those sites that lack perfect pairing with miRNA seed (seedless sites), which is compensated by pairing with miRNA nucleotides outside the seed region (Bartel, 2018; Chipman & Pasquinelli, 2019). The analysis of mRNAs that crosslink with AGO claim that about half of the profiled AGO:mRNA interactions comprise seedless sites (Chi *et al*, 2009; Hafner *et al*, 2010; Helwak *et al*, 2013). However, these sites fail to mediate detectable post-transcriptional

repression (Agarwal *et al*, 2015), likely because the absence of seed-match prevents the critical conformational change in AGO, which is required for stabilization of miRNA:target complex. Therefore, reliance solely on AGO-crosslinking may results in false targets. Nonetheless, specific examples of different types of such sites have been reported to function in specific contexts. The only such sites that also exhibit detectable conservation are 3'-compensatory sites, which compensate for mismatches in seed pairing by base pairing with miRNA nucleotides around 13-16 (Bartel, 2009; Friedman *et al*, 2009). Reports have observed measurable repression caused by 3' compensatory sites (Brennecke *et al*, 2005), however, such sites are rare; the 3'-compensatory sites comprise less than 1% of conserved targets sites (Friedman *et al*, 2009). Another group of seedless sites with perfect pairing to the 3' half of a miRNA have also been shown to function when present only in the ORF, which presumably exists only for specific miRNAs (e.g. miR-20 and let-7b) and is present only in handful of genes (Zhang *et al*, 2018). These sites mediate translational repression without destabilizing target mRNA and their activity is independent of TNRC6 proteins of RISC, indicating that these sites employ a non-canonical pathway of miRNA-mediated repression. While several such individual examples of seedless sites have been reported, they are extremely rare. Hence, they are not discussed further in this dissertation.

In summary, the current view is that the majority of functional target sites are located within the 3'UTRs and are canonical seed-matching sites. Nevertheless, it is clear that exceptions exist, although the prevalence of such sites is uncertain. Finally, the degree to which miRNA-specific deviations from these general rules exist is an open question.

Finding targets

The biological impact of miRNAs is defined by the targets that they regulate. Since the discovery of miRNAs, finding true targets of miRNAs has been an overarching question in the field (Lee *et al*, 1993; Bartel, 2009). Identification of responsive miRNA targets has been a challenging task due to following two major parameters: (i) the magnitude of repression mediated by miRNAs is relatively small; even the strongest miRNA target sites often lead to only two-fold repression of target mRNAs. (ii) While miRNAs can potentially regulate large number of targets, only a small fraction is truly regulated in any given cell. Over the years, several labs have spent enormous amount of efforts to devise approaches to effectively identify targets of miRNAs. These approaches can be divided into two major classes: 1) computational prediction of miRNA targets, and 2) biochemical method to capturing AGO:target interactions.

Computational prediction of miRNA targets

Since miRNAs bind to targets in sequence-specific manner, the most attractive approach has been to computationally find matches to miRNA seed sequences in the transcriptome (Lewis *et al*, 2003, 2005; Brennecke *et al*, 2005; Krek *et al*, 2005). However, since the match is typically only 7-8nt long, one can randomly find a seed match at every 16,390 (for 7mers) or 65,542 (for 8mer) bases of the transcriptome, making a list of targets as long as few thousand targets per miRNA family, the majority of which would likely be false-positive predictions. A key advancement to solving this issue was the use of sequence conservation in finding miRNA target sites (Lewis *et al*, 2005). The 3'UTRs from multiple species were aligned and searched for 7-8mer elements that are conserved across multiple species. Incorporating the preferential sequence conservation in the prediction algorithm reduced the number of predicted targets by an order of magnitude; out of few thousand possible seed matches, the updated algorithms estimated 300-400 preferentially conserved targets per miRNA family (Lewis *et al*, 2005; Friedman *et al*, 2009). This advancement significantly improved detection of true miRNA targets. The algorithms that incorporate sequence conservation in their model include TargetScan (Grimson *et al*, 2007; Garcia *et al*, 2011; Agarwal *et al*, 2015), PicTar (Lall *et al*, 2006), miRanda (Betel *et al*, 2008) and MirTarget2 (Wang & El Naqa, 2008).

While the sequence conservation provided confidence in target prediction, later studies indicated a high rate of false-negatives using these methods. Beyond the conserved sites, studies showed that reporters containing non-conserved target sites can also elicit significant repression of reporters (Farh *et al*, 2005), further complicating the prediction of true target. These findings motivated detailed study of target site properties (Grimson *et al*, 2007; Saetrom *et al*, 2007; Arvey *et al*, 2010; Garcia *et al*, 2011). Incorporation of properties identified by Grimson et al (as described above) into the prediction algorithms along with sequence conservation metrics significantly improved the prediction efficacy while expanding the repertoire of effective targets (Grimson *et al*, 2007). Later research found that the increased thermodynamic stability of seed pairing and reduced abundance of targets positively affect the miRNA-mediated repression, and integrating these features further improved the prediction models (Garcia *et al*, 2011). The most recent studies showed that accessibility of target site as predicted using RNA-folding algorithms also helps increase prediction efficacy (Agarwal *et al*, 2015, 2018). A suit of prediction methods based on variety of statistical models have been developed that incorporate many of these properties of target sites (Peterson *et al*, 2014), including TargetScan (Agarwal *et al*, 2015), MirTarget2 (Wang & El Naqa, 2008), miRanda (Betel *et al*, 2008), DIANA microT (Paraskevopoulou *et al*, 2013) and PITA (Kertesz *et al*, 2007). The majority of these tools use machine learning methods that rely upon transcriptome profiling data (such as microarray and RNA sequencing (RNA-seq)) of cells in which specific

miRNAs are perturbed, for training and testing their models. Typically, the changes observed in mRNA abundance upon perturbing miRNA levels are used to rate the influence of each of the aforementioned features of miRNA target sites, for example, 8mer site-type is given high weight if a set of genes containing 8mer sites demonstrated strong repression of target abundance compared to those that lack 8mer sites. Once the models are trained on a subset of datasets, they are then tested on the remaining datasets; the majority of these tools demonstrate substantial precision and sensitivity (Reczko *et al*, 2012) Such prediction algorithms have immensely advanced the research on role of miRNAs in various diseases states and cellular processes. However, it is important to acknowledge that the majority of the transcriptome profiling datasets used for developing these predictions models were generated in combination with transient transfection of high amounts of miRNAs (Lim *et al*, 2005; Linsley *et al*, 2007; Grimson *et al*, 2007; Selbach *et al*, 2008), likely resulting into many fold higher expression of miRNAs compared to their physiological levels. This deviation from the *in vivo* miRNA levels likely results into changes in mRNA abundance that are typically not observed *in vivo*, affecting the efficacy of prediction models. Indeed, research have shown that as the stoichiometry between miRNA and their targets changes, the repertoire of true miRNA targets varies (Jin *et al*, 2015; Denzler *et al*, 2016). Additionally, another study showed that transfection of one miRNA at very high levels affects the regulation by endogenous miRNAs by competing for loading onto AGO (Khan *et al*, 2009), thereby polluting

the regulatory effects of the transfected miRNA in transcriptome profiling datasets. Hence, it is highly likely that the supraphysiological levels of miRNAs used for generating training datasets lead to higher false-positive rates for the prediction models. Studies assessing efficacy of prediction algorithms have demonstrated that the false-positive estimates for even the most effective approaches are as high as 50% (Pinzón *et al*, 2017).

Conventionally, the most popular and effective approach to refining false predictions of computational models has been to retain only those predicted targets that also display in vivo changes in mRNA abundance in response a perturbed miRNA. However, the in vivo changes in mRNA abundance are likely confounded by downstream regulatory changes stemming from miRNA-mediated regulation of direct miRNA targets. For example, miRNA-mediated repression of messages coding for a transcription factor alters transcription of the downstream targets of the transcription factor (indirect targets). Indeed, miRNAs and transcription factors are intertwined in gene regulatory networks (Hobert, 2008), strongly suggesting existence of confounding downstream changes regulated at transcription. This confounding effect of indirect targeting by miRNAs adds another layer of complexity in effective identification of miRNA targets. Separating the direct miRNA regulation and the downstream indirect changes, and quantifying contributions of each in controlling miRNA regulatory networks remain outstanding open questions.

Biochemical approaches

Biochemical approaches have also been developed to robustly identifying miRNA targets. Major efforts focused on genome-wide experimental profiling of AGO-bound targets and associated miRNAs. The key advancement in this area was to repurpose a previously developed technique called cross-linking and immunoprecipitation (CLIP) for profiling RNAs bound to RNA-binding proteins (RBP) of interest. Using antibodies for AGO in CLIP methods, several labs developed different versions of AGO-CLIP over the years (Chi *et al*, 2009; König *et al*, 2010; Hafner *et al*, 2010; Helwak *et al*, 2013; Kudla *et al*, 2011; Moore *et al*, 2015); I will review few important modifications here. In general, all methods used UV-cross-linking and immuno-precipitation of AGO, followed by protein digestion, library preparation and sequencing (Lin & Miles, 2019). While the sequencing allowed identification AGO-bound mRNAs, a major challenge was to determine the miRNAs that is guiding the RISC to the target mRNAs. Most CLIP approaches can be grouped into two sets based on the way they determine identity of the associated miRNA. The first set of approaches rely on the fact that during reverse transcription step in library preparation, the reverse transcriptase enzyme either stops or introduces mutations at the site of crosslink due to residual amino acids left after protein digestion, which is then profiled using sequencing of the ends of cDNA molecules (Chi *et al*, 2009; König *et al*, 2010; Hafner *et al*, 2010; Van Nostrand *et al*, 2016). The information about associated miRNAs is inferred using computational approaches to predict miRNA-

seed match near the site of crosslink, thus reducing the search space for target sites (De & Gorospe, 2017). The second set of methods performed an additional step of conjugating two adjacent RNA molecules (miRNA & target RNA) via single-stranded RNA-RNA ligation and sequencing chimeras (Helwak *et al*, 2013; Moore *et al*, 2015), followed by computational determination of miRNA and target mRNAs. Both sets of approaches improved miRNA target identification while providing experimental proof of miRNA and target mRNA interaction. However, these approaches are difficult to implement, and reproducibility is still an issue, even for the more mature CLIP-seq approaches (Kishore *et al*, 2011). Additionally, the results of these experiments are strongly affected by technical variations between the protocols, for example, the chimera-based approaches work efficiently only for highly expressed miRNAs and mRNA targets, missing large numbers of true targets (Lu & Leslie, 2016). Furthermore, studies, including our data, demonstrated that many targets identified using CLIP-based methods show no change in mRNA abundance, perhaps due to transient binding of AGO, indicating large problems with false-positives that are intrinsic to these approaches.

Conclusions

MicroRNAs are critical players of post-transcriptional regulation, which are necessary for proper functioning of cells in complex organisms. Understanding the biological

impact of miRNAs requires robust identification of miRNA targets. While collective efforts from several labs over the years have greatly advanced our understanding of miRNA target sites, the high false-positive rates of existing methods indicate that important pieces of information are still missing from the complete picture of miRNA regulation. An outstanding challenge in the field is to partition the miRNA direct targets from confounding downstream regulatory changes, and quantifying and delineating the contributions of direct and indirect (downstream) targets would enable a systematic understanding of miRNA regulatory networks. In this dissertation, I present my work towards separating the direct miRNA targets from downstream indirect changes.

I developed a simple approach to deconvoluting transcriptional and post-transcriptional changes using RNA-seq and PRO-seq. I call this approach Combined Analysis of RNA-seq and PRO-seq (CARP). To test the efficacy of this approach, I developed human cell lines expressing a specific miRNA that mimics parameters of endogenous miRNAs, facilitating identification of true miRNA targets. First, I establish the robustness of this approach by validating the identified miRNA targets using orthogonal assays, such as CLIP-seq and ribosome profiling. Using CARP, I partitioned the modes of regulation elicited by mRNAs, dissecting the complexity of miRNA regulatory networks. I identified candidate transcription factors underlying the downstream regulatory changes in response to specific miRNAs. I evaluated

efficacy of ORF sites for the seven miRNAs used this work and found miRNA-specific activities of the sites, for example, the ORF sites to miR-1 were completely ineffective, consistent with much of the current literature, whereas the ORF sites to miR-122 exhibited significant repression, exceeding that of many canonical target sites. Overall, I establish that CARP robustly deconvolutes post-transcriptional and transcriptional programs, facilitating the identification of miRNA direct targets and dissection of miRNA regulatory network. This new approach is particularly is suitable for deconvoluting miRNA regulatory networks *in vivo*.

REFERENCES

Agarwal V, Bell GW, Nam J-W & Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4:**

Agarwal V, Subtelny AO, Thiru P, Ulitsky I & Bartel DP (2018) Predicting microRNA targeting efficacy in Drosophila. *Genome Biol.* **19:** 152

Arvey A, Larsson E, Sander C, Leslie CS & Marks DS (2010) Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular Systems Biology* **6:** n/a-n/a

Baek D, Villén J, Shin C, Camargo FD, Gygi SP & Bartel DP (2008) The impact of microRNAs on protein output. *Nature* **455:** 64–71

Bartel DP (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136:** 215–233

Bartel DP (2018) Metazoan MicroRNAs. *Cell* **173:** 20–51

Bazzini AA, Lee MT & Giraldez AJ (2012) Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science* **336:** 233–237

Betel D, Wilson M, Gabow A, Marks DS & Sander C (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.* **36:** D149-153

Brennecke J, Stark A, Russell RB & Cohen SM (2005) Principles of microRNA-target recognition. *PLoS Biol.* **3:** e85

Broderick JA, Salomon WE, Ryder SP, Aronin N & Zamore PD (2011) Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA* **17:** 1858–1869

Brümmer A & Hausser J (2014) MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of post-transcriptional gene regulation. *Bioessays* **36:** 617–626

Cannell IG, Kong YW & Bushell M (2008) How do microRNAs regulate gene expression? *Biochem. Soc. Trans.* **36:** 1224–1231

Carthew RW & Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136:** 642–655

Chen C-YA & Shyu A-B (2011) Mechanisms of deadenylation-dependent decay. *Wiley Interdiscip Rev RNA* **2:** 167–183

Chen C-YA, Zheng D, Xia Z & Shyu A-B (2009) Ago-TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. *Nat. Struct. Mol. Biol.* **16:** 1160–1166

Chi SW, Zang JB, Mele A & Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* **460:** 479–486

Chipman LB & Pasquinelli AE (2019) miRNA Targeting: Growing beyond the Seed. *Trends in Genetics* **35:** 215–222

De S & Gorospe M (2017) Bioinformatic tools for analysis of CLIP ribonucleoprotein data. *Wiley Interdiscip Rev RNA* **8:**

Denzler R, McGeary SE, Title AC, Agarwal V, Bartel DP & Stoffel M (2016) Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on

Competing Endogenous RNA-Regulated Gene Expression. *Molecular Cell* **64:** 565–579

Djuranovic S, Nahvi A & Green R (2012) miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science* **336:** 237–240

Doench JG & Sharp PA (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.* **18:** 504–511

Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, Baek D, Hsu S, Ghoshal K, Villén J & Bartel DP (2014) mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular Cell* **56:** 104–115

Elkayam E, Faehnle CR, Morales M, Sun J, Li H & Joshua-Tor L (2017) Multivalent Recruitment of Human Argonaute by GW182. *Mol. Cell* **67:** 646-658.e3

Elkayam E, Kuhn C-D, Tocilj A, Haase AD, Greene EM, Hannon GJ & Joshua-Tor L (2012) The structure of human argonaute-2 in complex with miR-20a. *Cell* **150:** 100–110

Fabian MR & Sonenberg N (2012) The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.* **19:** 586–593

Fang Z & Rajewsky N (2011) The Impact of miRNA Target Sites in Coding Sequences and in 3'UTRs. *PLoS ONE* **6:** e18067

Farh KK-H, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB & Bartel DP (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310:** 1817–1821

Friedman RC, Farh KK-H, Burge CB & Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19:** 92–105

Garcia DM, Baek D, Shin C, Bell GW, Grimson A & Bartel DP (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol* **18:** 1139–1146

Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP & Bartel DP (2007) MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing. *Mol Cell* **27:** 91–105

Gu S, Jin L, Zhang F, Sarnow P & Kay MA (2009) Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. *Nat. Struct. Mol. Biol.* **16:** 144–150

Guo H, Ingolia NT, Weissman JS & Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466:** 835–840

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp A-C, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M & Tuschl T (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141:** 129–141

Hausser J, Syed AP, Bilen B & Zavolan M (2013) Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* **23:** 604–615

Helwak A, Kudla G, Dudnakova T & Tollervey D (2013) Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell* **153:** 654–665

Hobert O (2008) Gene Regulation by Transcription Factors and MicroRNAs. *Science* **319:** 1785–1786

Huang Y, Shen XJ, Zou Q, Wang SP, Tang SM & Zhang GZ (2011) Biological functions of microRNAs: a review. *J Physiol Biochem* **67:** 129–139

Iwasaki S, Kawamata T & Tomari Y (2009) Drosophila argonaute1 and argonaute2 employ distinct mechanisms for translational repression. *Mol. Cell* **34:** 58–67

Jin HY, Gonzalez-Martin A, Miletic AV, Lai M, Knight S, Sabouri-Ghomi M, Head SR, Macauley MS, Rickert RC & Xiao C (2015) Transfection of microRNA Mimics Should Be Used with Caution. *Front Genet* **6:** 340

Kawamata T & Tomari Y (2010) Making RISC. *Trends Biochem. Sci.* **35:** 368–376

Kertesz M, Iovino N, Unnerstall U, Gaul U & Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39:** 1278–1284

Khan AA, Betel D, Miller ML, Sander C, Leslie CS & Marks DS (2009) Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nat Biotech* **27:** 549–555

Kim D, Sung YM, Park J, Kim S, Kim J, Park J, Ha H, Bae JY, Kim S & Baek D (2016) General rules for functional microRNA targeting. *Nat Genet* **48:** 1517–1526

Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M & Zavolan M (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* **8:** 559–564

Kloosterman WP, Wienholds E, Ketting RF & Plasterk RHA (2004) Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res.* **32:** 6284–6291

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM & Ule J (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17:** 909–915

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M & Rajewsky N (2005) Combinatorial microRNA target predictions. *Nat. Genet.* **37:** 495–500

Kudla G, Granneman S, Hahn D, Beggs JD & Tollervey D (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **108:** 10010–10015

Lai EC (2002) Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30:** 363–364

Lall S, Grün D, Krek A, Chen K, Wang Y-L, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P, Kao H-L, Gunsalus KC, Pachter L, Piano F & Rajewsky N (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.* **16:** 460–471

Lee RC, Feinbaum RL & Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75:** 843–854

Lewis BP, Burge CB & Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20

Lewis BP, Shih I -hung, Jones-Rhoades MW, Bartel DP & Burge CB (2003) Prediction of mammalian microRNA targets. *Cell* **115:** 787–798

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS & Johnson JM (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773

Lin C & Miles WO (2019) Beyond CLIP: advances and opportunities to measure RBP–RNA and RNA–RNA interactions. *Nucleic Acids Research* **47:** 5490–5501

Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, Johnson JM, Cummins JM, Raymond CK, Dai H, Chau N, Cleary M, Jackson AL, Carleton M & Lim L (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell. Biol.* **27:** 2240–2252

Lu Y & Leslie CS (2016) Learning to Predict miRNA-mRNA Interactions from AGO CLIP Sequencing and CLASH Data. *PLoS Comput. Biol.* **12:** e1005026

Mathonnet G, Fabian MR, Svitkin YV, Parsyan A, Huck L, Murata T, Biffo S, Merrick WC, Darzynkiewicz E, Pillai RS, Filipowicz W, Duchaine TF & Sonenberg N (2007) MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science* **317:** 1764–1767

McGeary SE, Lin KS, Shi CY, Bisaria N & Bartel DP (2018) The biochemical basis of microRNA targeting efficacy Biochemistry Available at: http://biorxiv.org/lookup/doi/10.1101/414763 [Accessed November 22, 2019]

Mencía A, Modamio-Høybjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre LA, del Castillo I, Steel KP, Dalmay T, Moreno F & Moreno-Pelayo MA (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat. Genet.* **41:** 609–613

Moore MJ, Scheel TKH, Luna JM, Park CY, Fak JJ, Nishiuchi E, Rice CM & Darnell RB (2015) miRNA–target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature Communications* **6:** ncomms9864

Nakanishi K, Weinberg DE, Bartel DP & Patel DJ (2012) Structure of yeast Argonaute with guide RNA. *Nature* **486:** 368–374

Ni W-J & Leng X-M (2015) Dynamic miRNA–mRNA paradigms: New faces of miRNAs. *Biochemistry and Biophysics Reports* **4:** 337–341

Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T & Hatzigeorgiou AG (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research* **41:** W169–W173

Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L & Congdon CB (2014) Common features of microRNA target prediction tools. *Front. Genet.* **5:** Available at: http://journal.frontiersin.org/article/10.3389/fgene.2014.00023/abstract [Accessed November 24, 2019]

Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, Bertrand E & Filipowicz W (2005) Inhibition of Translational Initiation by Let-7 MicroRNA in Human Cells. *Science* **309:** 1573–1576

Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F & Seitz H (2017) microRNA target prediction programs predict many false positives. *Genome Res.* **27:** 234–245

de Pontual L, Yao E, Callier P, Faivre L, Drouin V, Cariou S, Van Haeringen A, Geneviève D, Goldenberg A, Oufadem M, Manouvrier S, Munnich A, Vidigal JA, Vekemans M, Lyonnet S, Henrion-Caude A, Ventura A & Amiel J (2011) Germline deletion of the miR-17∼92 cluster causes skeletal and growth defects in humans. *Nat. Genet.* **43:** 1026–1030

Reczko M, Maragkakis M, Alexiou P, Papadopoulos GL & Hatzigeorgiou AG (2012) Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data. *Front. Gene.* **2:** Available at: http://journal.frontiersin.org/article/10.3389/fgene.2011.00103/abstract [Accessed November 24, 2019]

Rinck A, Preusse M, Laggerbauer B, Lickert H, Engelhardt S & Theis FJ (2013) The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance. *RNA Biol* **10:** 1125–1135

Saetrom P, Heale BSE, Snøve O, Aagaard L, Alluin J & Rossi JJ (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.* **35:** 2333–2342

Schaefke B, Sun W, Li Y-S, Fang L & Chen W (2018) The evolution of posttranscriptional regulation. *WIREs RNA* **9:** e1485

Schirle NT, Sheu-Gruttadauria J, Chandradoss SD, Joo C & MacRae IJ (2015) Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. *Elife* **4:**

Schirle NT, Sheu-Gruttadauria J & MacRae IJ (2014) Structural basis for microRNA targeting. *Science* **346:** 608–613

Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R & Rajewsky N (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature* **455:** 58–63

Straalen NM van & Roelofs D (2012) An introduction to ecological genomics 2nd ed. Oxford ; New York: Oxford University Press

Ui-Tei K, Naito Y, Nishi K, Juni A & Saigo K (2008) Thermodynamic stability and Watson–Crick base pairing in the seed duplex are major determinants of the efficiency of the siRNA-based off-target effect. *Nucleic Acids Research* **36:** 7100–7109

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M & Yeo GW (2016) Robust transcriptome-wide discovery of RNA-

binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13:** 508–514

Vasudevan S, Tong Y & Steitz JA (2007) Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. *Science* **318:** 1931–1934

Wang X & El Naqa IM (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* **24:** 325–332

Wilczynska A & Bushell M (2015) The complexity of miRNA-mediated repression. *Cell Death Differ* **22:** 22–33

Zhang K, Zhang X, Cai Z, Zhou J, Cao R, Zhao Y, Chen Z, Wang D, Ruan W, Zhao Q, Liu G, Xue Y, Qin Y, Zhou B, Wu L, Nilsen T, Zhou Y & Fu X-D (2018) A novel class of microRNA-recognition elements that function only within open reading frames. *Nat Struct Mol Biol* **25:** 1019–1027

Zhao W, Blagev D, Pollack JL & Erle DJ (2011) Toward a systematic understanding of mRNA 3' untranslated regions. *Proc Am Thorac Soc* **8:** 163–166

Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, Muth AN, Tsuchihashi T, McManus MT, Schwartz RJ & Srivastava D (2007) Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell* **129:** 303–317

CHAPTER 2[1]

Robust partitioning of microRNA targets from downstream regulatory changes

[1]This work is a manuscript in preparation. The current authors on this manuscript are Ravi K Patel, Jessica D. West, Ya Jiang, Elizabeth A. Fogarty and Andrew Grimson. Ravi K Patel conceptualized the experiments, performed most of the experiments, analyzed the sequencing data, performed all bioinformatic analyses, interpreted data and prepared the manuscript and figures. Detailed description of author contributions can be found below.

ABSTRACT

The biological impact of microRNAs is determined by their targets, and robustly identifying direct miRNA targets remains challenging. Existing methods suffer from high false-positive rates and are unable to effectively differentiate direct miRNA targets from downstream regulatory changes. Here, we present a simple approach to deconvolute post-transcriptional and transcriptional changes using PRO-seq with RNA-seq. In combination, these methods allow us to systematically profile the regulatory impact of a miRNA. We refer to this approach as CARP: Combined Analysis of RNA-seq and PRO-seq. We apply CARP to multiple miRNAs and show that it robustly distinguishes direct targets from downstream changes, while greatly reducing false positives. We validate our approach using Argonaute eCLIP-seq and ribosome profiling, demonstrating that CARP defines a comprehensive repertoire of true targets. For certain miRNAs, we identify numerous effective target sites within the coding region. CARP facilitates the dissection of complex changes in gene regulatory networks triggered by miRNAs and identification of transcription factors that underlie downstream regulatory changes. Given the robustness of the approach, CARP is particularly suitable for dissecting miRNA regulatory networks *in vivo*.

INTRODUCTION

While transcriptional regulation account for much of gene regulation, post-transcriptional regulation represents an additional and consequential layer of regulation (Corbett, 2018, Mayr, 2017). MicroRNAs (miRNAs), a class of small non-coding RNAs, are one of the major trans-acting factors responsible for post-transcriptional regulation (Bartel, 2018). Together with an Argonaute (AGO) protein, miRNAs function primarily by binding to target mRNA transcripts and inducing mRNA decay and/or translational repression. In humans and other mammals, there are many hundreds of different microRNAs (miRNAs), which collectively regulate the majority of human mRNA transcripts (Friedman, Farh et al., 2009) and likely contribute to all gene regulatory pathways. Accordingly, identifying the targets of miRNAs is fundamental to understanding their biological functions, and a wide variety of genomic, biochemical and computational approaches have been developed to address this question (Bartel, 2009, Riffo-Campos, Riquelme et al., 2016, Roberts & Borchert, 2017). Despite intense efforts, even the most effective approaches suffer from high rates of false positives and/or negatives (Pinzon, Li et al., 2017).

The majority of miRNA target sites in bilaterian animals are found in 3' untranslated regions (3'UTRs), and comprise a short sequence with perfect complementarity to the 5' end of the miRNA, or miRNA seed (Bartel, 2018). Effective seed-matching target sites are often located within a region of 3'UTR sequence that contains additional

features, such as high local AU-content, which determine site efficacy (Grimson, Farh et al., 2007, Nielsen, Shomron et al., 2007). In addition to these canonical seed-matching sites, numerous other types of sites have been reported, including sites in coding sequence and 5'UTRs, and sites without perfect seed matches (Ecsedi, Rausch et al., 2015, Lytle, Yario et al., 2007, Schnall-Levin, Zhao et al., 2010, Stark, Lin et al., 2007). The extent to which such non-canonical sites contribute to the total targeting repertoire of a miRNA is unclear. Moreover, miRNA-specific parameters influence the targeting properties of certain miRNAs (Agarwal, Bell et al., 2015, Garcia, Baek et al., 2011). The earliest effective approaches to predicting and identifying mammalian miRNA targets used comparative genomics, and worked by cataloguing orthologous 3'UTR sequences whose capacity to basepair perfectly to a miRNA seed sequence is detectably conserved (Brennecke, Stark et al., 2005, Lall, Grun et al., 2006, Lewis, Burge et al., 2005); such approaches remain an important component of defining biologically consequential miRNA targets. In addition to conserved target sites, a large number of non-conserved sites also respond to their cognate miRNAs (Farh, Grimson et al., 2005). Non-conserved sites constitute the majority of total sites; therefore, conservation alone cannot be used to robustly identify target sites (Friedman et al., 2009). Numerous computational approaches exist which predict the strength or efficacy of miRNA target sites, with varying degrees of effectiveness (Riffo-Campos et al., 2016, Roberts & Borchert, 2017). In general, these approaches are built on experimental data in which the transcriptome or proteome is monitored

in response to high and transient exposure to an exogenous miRNA. The resulting data, aggregated over many experiments, is used to train a model that captures the response to a miRNA, and the results extrapolated to other miRNAs and other cell types. Such tools have played an important role in accelerating our understanding of miRNA biology. Biochemical techniques (including CLIP-based assays) have also been used to identify miRNA targets (Chi, Zang et al., 2009); however, these assays suffer from high levels of background, perhaps arising from the transient nature of AGO binding. Indeed, subsequent attempts to verify non-canonical target sites identified from CLIP have shown that they are largely ineffective (Agarwal et al., 2015).

Although approaches that identify or predict miRNA targets have continued to evolve and improve, the vast majority of both training and validation datasets rely upon cell culture experiments in which an exogenous miRNA is introduced transiently at high concentration (Bartel, 2009). Extending these approaches to *in vivo* settings, with miRNA knockouts, for example, has indicated that target prediction remains valuable but imperfect. Deviations between target prediction and *in vivo* miRNA-mediated regulation presumably derive from numerous sources. Biologically consequential miRNAs are enmeshed within complex gene regulatory networks, and the action of such a miRNA is likely to elicit substantial downstream changes beyond the direct targets (Gosline, Gurtan et al., 2016). For example, a miRNA may directly repress an mRNA encoding a transcription factor, thus altering the downstream targets of the

transcription factor and potentially confounding efforts to identify the direct targets of the initiating miRNA. Indeed, a large body of literature illustrates intimate mingling of miRNAs and transcription factors within gene regulatory networks (Hobert, 2008); thus, identifying transcription factors that direct downstream regulatory changes initiated by a miRNA is likely an important step towards understanding biological functions of miRNAs. This complexity in miRNA regulatory networks alone makes miRNA target prediction *in vivo* problematic. Three major challenges exist: (1) the relatively subtle regulation elicited by a miRNA, often less than 2-fold, (2) the large number of potential targets, often several hundred, and finally, (3) for consequential miRNAs, the extent of downstream changes.

A popular and effective approach to identifying miRNA targets *in vivo* is to intersect lists of genes differentially expressed in response to a specific miRNA with lists of predicted targets. Importantly, both lists often include many hundreds of genes, thus random overlap alone will generate a substantial set of intersecting candidate direct miRNA targets. We reasoned that eliminating genes whose differential expression derives from transcriptional regulation might enable more robust delineation of the direct targets of a miRNA. Prior to our work, a conceptually similar approach has been developed: EISA (Exon–Intron Split Analysis) exploits intron-mapping reads in RNA-seq data to indicate levels of pre-mRNAs and thus serves as a proxy for transcriptional activity (Gaidatzis, Burger et al., 2015). The advantage of EISA is that it is straightforward to implement; nevertheless, pre-mRNA levels are not a direct

sensor of transcription, potentially compromising the accuracy of this method. Here, we use PRO-seq (Precision Run-On sequencing), a tool that directly monitors transcription across the genome (Kwak, Fuda et al., 2013), in combination with RNA-seq to robustly distinguish between direct miRNA targets and indirect effects arising from downstream regulation. We corroborate the efficacy of our approach using orthogonal genomic assays to measure AGO-miRNA binding to targets (AGO eCLIP-seq (Van Nostrand, Pratt et al., 2016)) and translational efficiency (ribosome profiling (Ingolia, Ghaemmaghami et al., 2009)). We use these data to investigate mechanisms of miRNA-mediated repression; for example, we quantify the contributions of miRNA-mediated mRNA decay and translational repression using miRNAs expressed at physiological levels. Additionally, we identify novel, effective miRNA target sites residing within the coding region; interestingly, such coding sites are only prevalent for a subset of miRNAs that we examine. Because PRO-seq also maps active enhancers, we identify candidate transcription factors associated with enhancers exhibiting altered transcriptional activity. We find that activities of many of these transcription factors are modulated by specific miRNAs, and which contribute to the downstream changes in transcriptional regulation. Using CARP to deconvolute regulation occurring at the level of transcription, post-transcription, or both, we demonstrate that combining RNA-seq and PRO-seq is a powerful approach to investigate complex transcriptional and post-transcriptional gene regulatory networks.

RESULTS

A system for measuring miRNA-mediated post-transcriptional regulation

Given the limitations of conventional methods for identifying direct miRNA targets, we sought to develop a novel experimental approach to robustly identify genes subject to post-transcriptional regulation. To discriminate between direct miRNA targets, which are regulated at the post-transcriptional level, and indirect targets, which are likely to be predominantly regulated at the transcriptional level, we used RNA-seq and PRO-seq to measure steady-state mRNA levels and transcription rates, respectively. Our strategy was to profile HEK293 cells in the presence and absence of specific miRNAs. We chose first to study miR-1 and miR-122 because they are well-studied human miRNAs not expressed in HEK293 cells. Unlike majority of previous studies that used high levels of transiently transfected miRNA, we elected to stably integrate miRNA hairpins embedded within the intron of a doxycycline-inducible GFP reporter, to more closely approximate *in vivo* expression levels. To ensure accurate processing of the mature miRNAs, we designed miRNA hairpins based on established sequence and structural features favored by the miRNA biogenesis machinery (Fang & Bartel, 2015). In order to approximate steady-state miRNA levels, we treated cells with doxycycline for seven days (Fig S2.1A; (Kingston & Bartel, 2019)). We confirmed high physiological levels and accurate processing of induced miRNAs using small-RNA sequencing (sRNA-seq; Fig S2.1B). Using quantitative PCR (qPCR) of miRNAs, we established that miR-1 and miR-122 are expressed within the

physiological range of other miRNAs found in HEK293 cells (Fig S2.1C): compared to miR-10, the mostly highly detected endogenous miRNA, miR-122 is expressed at a about the same level and miR-1 is two-fold higher. Furthermore, using sRNA-seq, the overall miRNA profile was unchanged upon induction of these exogenous miRNAs. Importantly, using reporter assays, we found that the induced miRNAs are functionally active (Fig S2.1E) and that the anticipated miRNA guide strand was chosen selectively for loading onto AGO (Fig S2.1F). Collectively, these results indicate that our synthetic miRNA expression system mimics endogenous miRNA, facilitating the analysis of miRNA targets while approximating normal *in vivo* parameters.

We performed PRO-seq and RNA-seq on cells expressing miR-1 or miR-122 and compared to control cells lacking a miRNA hairpin. Principal component analysis (PCA) of the RNA-seq data showed that replicate transcriptomes cluster tightly, indicating high data reproducibility, and that the transcriptomes of miR-1 and miR-122 expressing cells are distinct from each other and from the control cells (Fig 2.1B). When we performed the PCA analysis of the PRO-seq data, however, we found that the genome-wide transcriptional profile of control and miR-122 expressing cells were not well separated (Fig 2.1B), perhaps indicating that miR-122 does not elicit wide-spread changes in transcription. Indeed, transcription of 371 and 10 genes were significantly altered (q-value < 0.05) in response to miR-1 and miR-122, respectively, supporting this interpretation (Fig S2.1G).

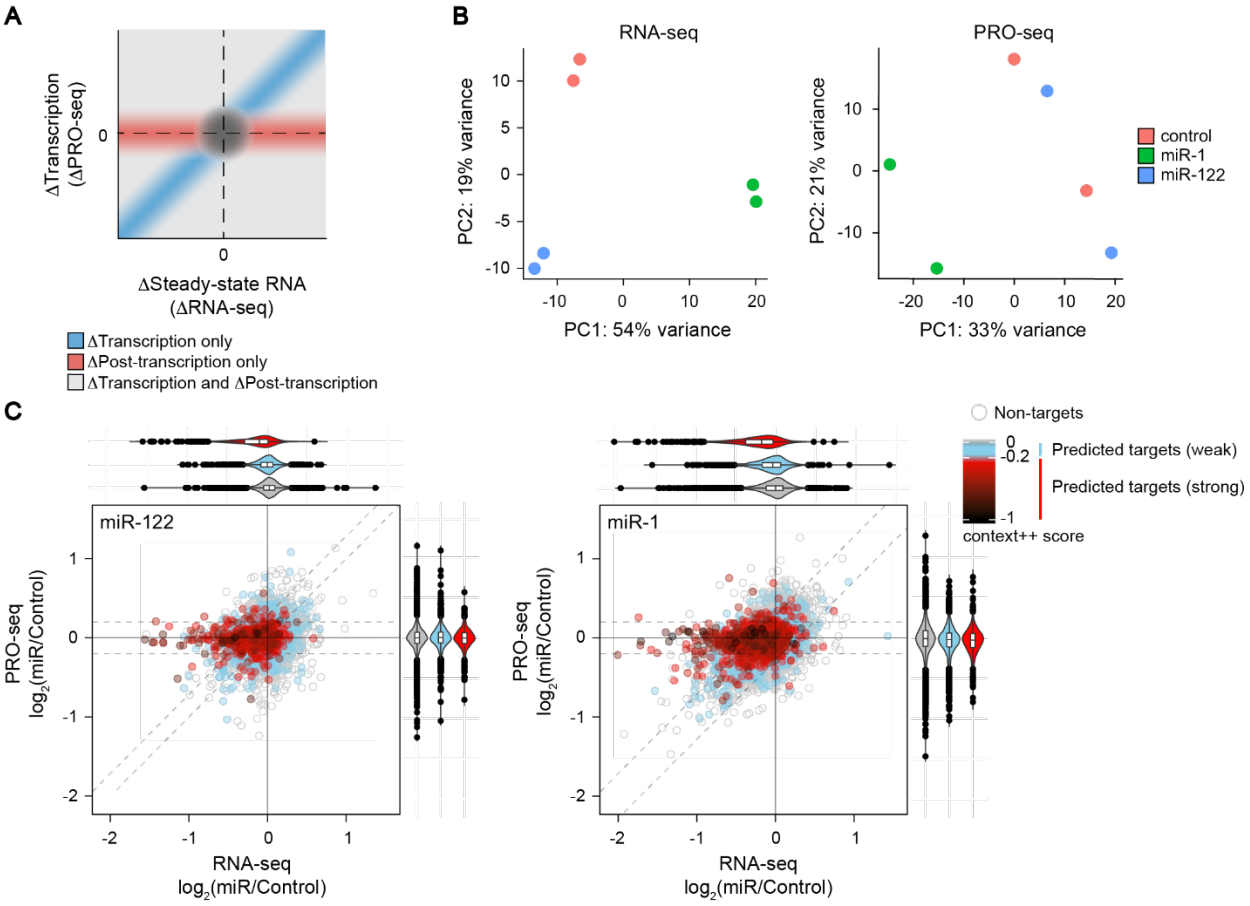

Figure 2.1. Combined analysis of RNA-seq and PRO-seq identifies genes subject to post-transcriptional regulation.

(**A**) Schematic representation of combined analysis of RNA-seq and PRO-seq data. Changes in transcription (y-axis; PRO-seq) and transcript levels (x-axis; RNA-seq) are plotted; genes within red and blue shaded regions represent those that are regulated exclusively by post-transcriptional and transcriptional regulation, respectively. Box in the bottom illustrates an equation for computing post-transcriptional changes by subtracting changes in transcriptional output (ΔPRO-seq) from changes in steady-state mRNA levels (ΔRNA-seq). (**B**) Principal component analysis of RNA-seq (D) and PRO-seq (E) data, for miR-1 (green) and miR-122 (blue) induced cell cells, and control cell lines without induced miRNAs (red). (**C**) Dot plots depicting changes in steady-state level (ΔRNA-seq; x-axis) and transcriptional output (ΔPRO-seq; y-axis) for all expressed genes (dots) upon miR-122 (left) or miR-1 (right) expression. The grey, green and red dots represent genes without predicted target sites, with predict weak targets and predicted strong targets of cognate miRNAs, respectively. The gradient of red color indicate predicted target site efficacy (context ++ score; add scale bar). The density plots on the top and right illustrate distribution of $\log_2$ fold-changes in ΔRNA-seq and ΔPRO-seq, respectively, for the different categories of

genes, color-coded as described above. Log_2 fold change of 0.2 in ΔPRO-seq and ΔRNA-seq – ΔPRO-seq is indicated using horizontal and diagonal dotted lines, respectively.

To investigate the effect of miRNA induction on transcriptional and post-transcriptional regulation, we compared changes in PRO-seq and RNA-seq across individual genes. The rates of synthesis and degradation together determine steady-state mRNA levels. Therefore, we reasoned that changes in PRO-seq signal (ΔPRO-seq), representing changes in transcription rates, subtracted from changes in RNA-seq (ΔRNA-seq), representing changes in steady-state RNA levels, would provide a quantitative readout for post-transcriptional regulation (Figure 1A). This combined analysis of RNA-seq and PRO-seq, referred to as CARP hereafter, identified many transcripts subjected to post-transcriptional repression in response to miR-122 or miR-1 in RNA-seq profiles without any changes in PRO-seq (Fig 2.1C). In fact, the majority of these transcripts contained target sites predicted to be strongly effective, as defined by TargetScan ((Agarwal et al., 2015); context++ score < -0.2; referred to as predicted strong targets hereafter). Correlating with predictions of site efficacy, most transcripts containing target sites predicted to be weakly effective, as defined by TargetScan context++ score of ≥ -0.2 (referred to as predicted weak targets hereafter), showed very subtle changes, if any, in RNA-seq or PRO-seq compared to genes without target sites (Fig 2.1C). Additionally, many genes, including confidently predicted targets, demonstrated concordant downregulation in both PRO-seq and RNA-seq in miR-1 expressing cells (Fig 2.1C), likely representing genes that are

regulated predominantly at the level of transcription, and not direct targets of miR-1. However, such concordant changes were minimal in miR-122 expressing cells, consistent with the PCA of PRO-seq profiles (Fig. 1B). Furthermore, the majority of genes were insensitive to induced miRNAs (63% and 79% genes in miR-1 and miR-122 samples, respectively, with $\log_2$ fold-change smaller than 0.2 in both RNA-seq and PRO-seq), as expected for single miRNA perturbation experiments (Fig. 1C). Included in these unchanged mRNAs were 45% and 60% of predicted targets of miR-1 and miR-122, respectively (47% and 55%, if considering conserved target sites), likely representing high false-positive rates of prediction algorithms, as reported previously (Krek et al. 2005; Lewis et al. 2005; Friedman et al. 2009). Taken together, CARP robustly quantifies post-transcriptional and transcriptional regulation, enabling a robust experimental framework for distinguishing the direct targets of a miRNA from the resulting downstream regulatory changes.

Identification and analysis of direct targets

Direct miRNA targets correspond to transcripts bound by miRNA-loaded AGO at target sites, resulting in accelerated decay and/or translational repression. To assess the ability of CARP to detect direct targets, we first performed a likelihood ratio test (LRT) (McCarthy, Chen et al., 2012) to identify transcripts that experience a significant post-transcriptional change – that is, a significant change in steady-state

mRNA level after accounting for any change in transcription. Consistent with the role of a miRNA as a negative regulator, most of the genes subject to post-transcriptional regulation (98% and 75%) exhibited repression in response to miR-122 or miR-1, respectively (Fig 2.2A). To evaluate further the properties of post-transcriptionally regulated genes, we assessed the presence of predicted target sites for the cognate induced miRNAs, using TargetScan v7.0 (Agarwal et al., 2015). The 3'UTRs of the majority of genes contained predicted target sites for miR-122 or miR-1 (90% and 63%, respectively), although we note that many additional predicted targets were not detectably repressed. To confirm that the post-transcriptional repression of the predicted targets is due to direct binding of RISC, we performed UV crosslinking and immunoprecipitation of AGO followed by sequencing (eCLIP-seq; (Van Nostrand et al., 2016)) to identify mRNAs bound by AGO. We reasoned that if the post-transcriptionally repressed genes containing predicted target sites are indeed direct targets, their 3'UTRs would exhibit increased binding of AGO in miR-122 or miR-1 expressing cells compared to control cells. Following quality trimming and mapping of eCLIP data, peaks of eCLIP reads, representing regions of the transcriptome occupied by AGO, were identified using CLIPper (Lovci, Ghanem et al., 2013). Although AGO binds predominantly at miRNA target sites in 3'UTRs (Chi et al., 2009), a majority (53%) of eCLIP peaks overlapped with introns, with only 17% peaks

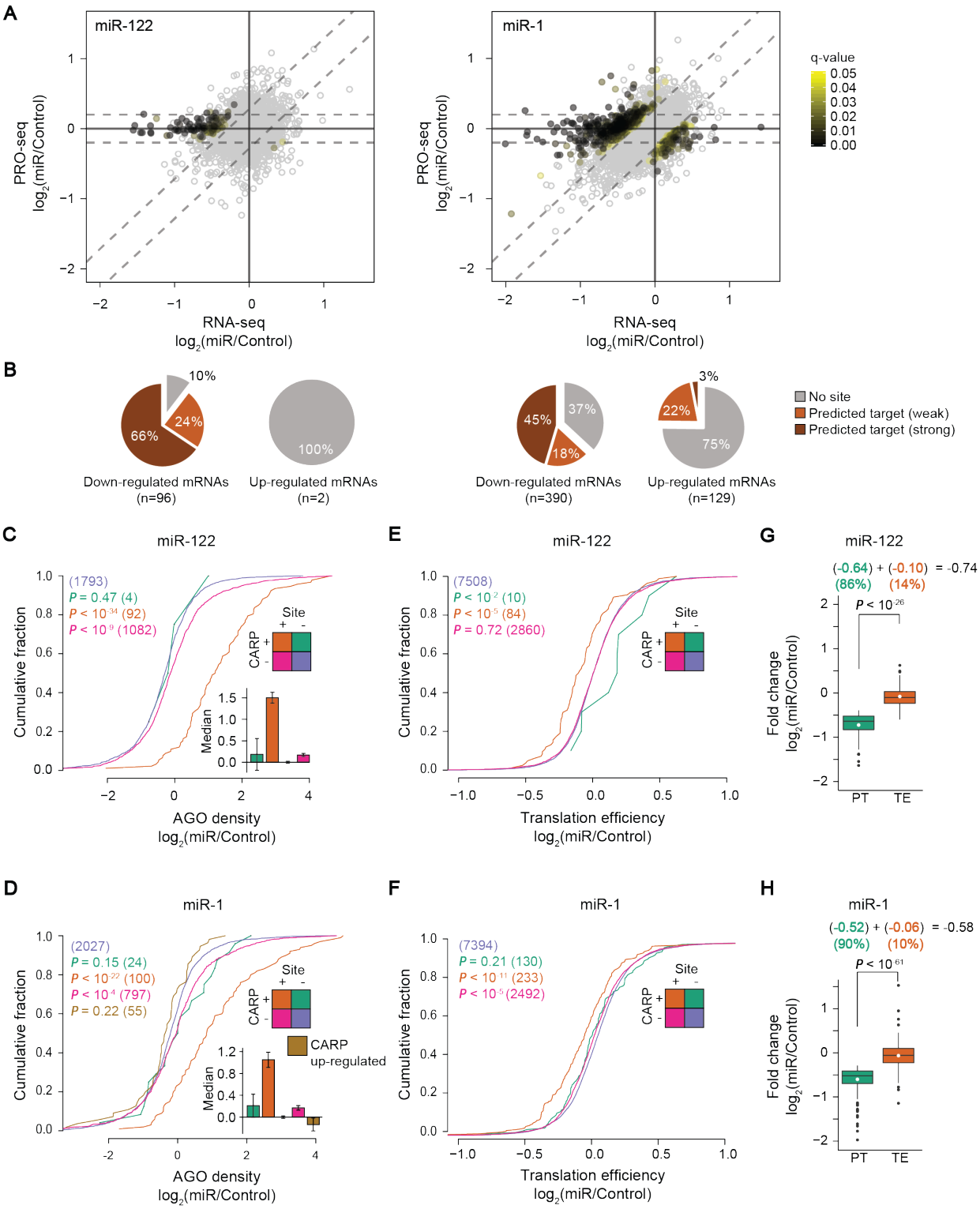

Figure 2.2. Identification and analysis of direct targets of cognate miRNAs. (**A,B**) Identification of post-transcriptional changes in miR-122 (A) and miR-1 (B)

expressing cells. The dot plots are as described in Fig 2.1F and Fig 2.1G, except that the filled and open dots illustrate genes with and without significant post-transcriptional change (q-value cut-off of 0.05), respectively. Scale bar denotes color-coded q-values. The numbers indicate counts of post-transcriptionally down-regulated (blue) and upregulated (red) genes in response to each miRNA. The fraction of predicted target sites (c1 and c2 denote sites predicted as stronger and weaker, respectively) in downregulated and upregulated transcripts is depicted in pie-charts s. (**C,D**) Cumulative distribution plots of AGO enrichment in miR-122 (C) or miR-1 (D) expressing cells, assessed by eCLIP signal within 3'UTRs. Genes were partitioned into four color-coded groups based on presence or absence of target site (*site*), and whether genes were determined to be post-transcriptionally repressed in response to miRNA indication (*CARP*). The group of non-target transcripts not subject to post-transcriptional regulation (blue) was used as the background set (blue) and compared to the remaining three groups (Wilcoxon Rank-Sum test). Each comparison P value is indicated, and colored according to the foreground group identity; parenthetical values denote number of eCLIP peaks in each group. (**E,F**) Cumulative distribution plots of translation efficiency in miR-122 (E) or miR-1 (F); otherwise as described in panels C and D. (**G,H**) Box plots comparing (x-axis) contributions of mRNA degradation and translational inhibition in miRNA-mediated post-transcriptional repression (y-axis) of direct targets of miRNAs. Mean and median $\log_2$ fold change indicated by white point and horizontal bar, respectively; overall distributions were compared (Wilcoxon Rank-sum test). The medians of $\log_2$ fold changes was used to quantify relative contributions of decay and translational repression.

in 3'UTR (Fig S2.2A), implying high levels of background in the eCLIP-data. Further examination of the peak counts revealed that the replicates exhibited large variations (Fig S2.2B), a limitation associated even with more mature CLIP protocols using AGO. To remove irreproducible peaks, we calculated irreproducibility discovery rate (IDR; (Li, Brown et al., 2011)), a method appropriate for eCLIP data (Van Nostrand et al., 2016), and filtered peaks with less than 0.1 IDR, resulting in an average of 5,192 peaks per sample (Fig S2.2C). A majority (54%) of our reproducible peaks overlapped with 3'UTRs, with only 7% in introns (Fig S2.2D), suggesting that the intronic peaks

found in our unfiltered data and in previous reports (Moore, Scheel et al., 2015) likely correspond to background signal. To assess the quality of our eCLIP data further, we compared AGO enrichment and post-transcriptional repression of predicted targets as a function of predicted target site efficacy using TargetScan (Agarwal et al., 2015). As the predicted site efficacy increased, we observed an increase in AGO enrichment in 3'UTRs concomitant with an increase in post-transcriptional repression (Fig S2.2E), indicating that our filtered eCLIP data reflect high quality AGO occupancy profiles. Further analysis of AGO enrichment revealed that AGO is strongly enriched in 3'UTRs of down-regulated genes containing predicted target sites, with 85-90% of genes bound by more AGO in miRNA-expressing cells compared to control cells (Fig 2.2C,D). Strikingly, there was minimal evidence of AGO binding for predicted targets that CARP identified as not subject to post-transcriptional repression (Fig 2.2C,D). Collectively, these observations indicate that the intersection of genes distinguished as post-transcriptionally repressed by CARP, and genes containing predicted target sites in 3'UTR represents a set of high confidence direct targets of miRNAs.

Prior to our work, the conceptually equivalent approach EISA (Gaidatzis et al., 2015) used pre-mRNA levels, approximated by intronic reads in RNA-seq, as a proxy for transcriptional output. To compare the performance of EISA with CARP, we contrasted changes in transcriptional output measured using PRO-seq with that inferred using intronic reads from RNA-seq. We found that changes in intronic reads in response to miR-1 were poorly correlated ($r=0.26$; P-value $< 10^{-15}$) with changes in

PRO-seq measurements (Fig S2.2F). Because PRO-seq directly measures transcription by capturing transcriptionally engaged RNA polymerase (Kwak et al., 2013), our results indicate that the level of intronic reads is an imperfect measure of transcription. Furthermore, we observed reduced reproducibility in transcription estimated using intronic read compared to measurements made with PRO-seq (data not shown). Consistent with lower reproducibility, we found fewer genes (87%) with significant post-transcriptional change in EISA compared to CARP for miR-1 (Fig S2.2G). Equivalent results were observed for differential regulation elicited by miR-122. Finally, we found somewhat higher AGO enrichment in 3'UTRs of predicted targets identified using CARP compared to those in EISA (Fig S2.2H; P-value = 0.076). Collectively, these findings demonstrate that CARP outperforms EISA in quantification of post-transcriptional regulation and identification of miRNA direct targets. Nevertheless, it is clear that EISA, which requires only RNA-seq profiling, is easier to implement and is a valuable tool to define miRNA targets.

The major mode of miRNA-mediated repression is now believed to be accelerated mRNA decay with translational repression playing only a minor role (Guo, Ingolia et al., 2010), although this conclusion has been contested (Bethune, Artus-Revel et al., 2012), and exceptions based on cell type exist (Bazzini, Lee et al., 2012, Djuranovic, Nahvi et al., 2012, Mishima, Fukao et al., 2012). We revisited this important question, asking whether our robustly identified target set exhibited translational repression, and, the degree to which predicted targets that were CARP-negative underwent

translational repression. Thus, we performed ribosome profiling (Ingolia et al., 2009) to measure translation efficiency in miRNA-expressing cells compared to controls. We first assessed the translation efficiency of CARP-positive genes, which revealed that these direct miRNA targets experienced significant translational repression compared to non-targets (Fig 2.2E,F). However, the magnitude of translational repression was six- to nine-fold smaller than the contribution made by mRNA decay for miR-122 and miR-1, respectively (Fig 2.2G,H). We quantified contributions of translational regulation and decay in miRNA-mediated repression of direct targets and found that reduced mRNA levels explains most of the overall regulation observed, whereas translational repression contributes very little (14 and 10% translational, compared to 86 and 90% for miR-122 and miR-1, respectively (Fig 2.2G,H). These estimates are consistent with those obtained in previous studies using transiently transfected miRNAs (Guo et al., 2010). Additionally, we found that only a minority of direct targets (2 and 10) demonstrated a significant change in translational efficiency in miR-122 or miR-1 expressing cells, respectively (Fig S2.2I). These results indicate that miRNAs expressed at physiological levels repress targets predominantly via accelerated mRNA decay. Hence, CARP is well suited for identifying direct targets in most cellular contexts.

Next, we investigated whether the large number of predicted targets which lacked significant post-transcriptional repression via accelerated decay but exhibited weak AGO enrichment were instead regulated via translational repression. To test this

possibility, we assessed the translation efficiency of predicted targets that were not detected by CARP, and found no change in their translation efficiency compared to that of non-target mRNAs (Fig 2.2E,F). Despite the absence of statistically significant post-transcriptional change or change in translation efficiency (Fig 2.2E,F), these genes demonstrated weak but significant enrichment of AGO in 3'UTRs (Fig 2.2C,D). This enrichment was stronger for the subset that comprised predicted strong targets (Fig S2.2J). These observations suggest that the results of CLIP-based assays should be interpreted carefully.

Identification and analysis of indirect targets

In addition to direct targets, we also found a smaller portion (10% and 37%, for miR-122 and miR-1, respectively) of post-transcriptionally repressed genes that lacked predicted target sites (Fig 2.2A,B). This set represents either direct targets containing non-canonical target sites, or indirect targets of miRNAs for which the regulation is itself post-transcriptional. However, the absence of AGO enrichment within the 3'UTRs of these transcripts (Fig 2.2C,D) implies that they are indirect targets of miR-122 and miR-1, which are downregulated post-transcriptionally. Nevertheless, it was important to investigate these alternatives further. Notably, we also observed post-transcriptional upregulation of 129 genes upon miR-1 induction, which we assumed represent indirect targets of miR-1 (Fig 2.2A,B). Consistent with this interpretation,

almost all (97%) upregulated genes lacked predicted target sites and also lacked AGO enrichment (Fig 2.2B, D). Thus, our results indicate that indirect targeting triggered by miR-1 expands to both transcriptional and post-transcriptional regulation, presumably as a result of miRNA-mediated repression of transcripts coding for transcription factors and regulatory RNA-binding proteins, respectively. Indeed, the direct targets of miR-1 included many genes coding for transcription factors and RNA binding proteins (Fig S2.2K). Taken together, these data indicate that while the regulatory network of miR-122 is comprised primarily of direct targets in HEK293 cells, miR-1 elicits more complex responses involving both direct and indirect targets to regulate gene expression in these cells.

Partitioning modes of regulation elicited by miRNAs

To systematically investigate the utility of CARP, we examined relationships among three categories of genes: (1) genes exhibiting reduced mRNA abundance in response to either miR-122 or miR-1, which we measured using RNA-seq, (2) genes with confidently predicted microRNA target sites, which we determined using TargetScan (Agarwal et al., 2015), and (3) genes that we found to undergo significant post-transcriptional repression in response to miR-122 or miR-1 (Fig 2.3A). We compared these three gene sets, and depicted the results using Venn Diagrams (Fig 2.3A). A substantial fraction of predicted targets, for both miR-122 and miR-1 (72 and 57%,

respectively), do not exhibit evidence of significant changes in mRNA abundance or in post-transcriptional levels, consistent with established high rates of false positive predictions and the subtle nature of miRNA regulation (set *a* in Fig. 3A and B). Consistent with this interpretation, this set of genes exhibited no change in CARP, RNA-seq, PRO-seq or ribosome profiling (Fig 2.3B). Similarly, we found minimal evidence for miR-122 or miR-1 induced AGO binding to the 3'UTRs of this gene set. (Fig 2.3 C,D, S2.3A). We note that almost all (75 and 80%) of these transcripts exist as 3'UTR isoforms that contain the predicted target sites; thus, the absence of regulation cannot be attributed to alternative processing (Fig S2.3B). Taken together, these results imply that this category likely corresponds to false-positives predictions and/or targets that are not effective in this cell line.

The second subset of genes we considered were those that contain predicted target sites whose mRNA abundance is reduced in response to the cognate miRNA, and which we determined were regulated post-transcriptionally (set *b*). Importantly, this gene set exhibited the largest reduction in mRNA levels in response to the cognate miRNA (Fig 2.3B). Moreover, post-transcriptional regulation was the primary, and for many genes, sole factor responsible for repression (Fig 2.3B). Congruently, this set of transcripts also demonstrate the strongest AGO enrichment in their 3'UTRs in response to the induced miRNAs (Fig. 3C,D, Fig. S3A). This target set corresponds to direct miRNA targets. We note that translational repression for these set of targets is minimal (Fig 2.3B).

We next considered predicted target mRNAs that were downregulated in mRNA abundance but without significant post-transcriptional repression (set *c*). As a set, such genes exhibit minimal evidence of either miRNA-induced post-transcriptional repression or translational repression (Fig 2.3B). We recognize that a subset of these genes may represent *bona fide* direct targets, but with reduced site efficacy, thus reducing the ability to detect experimentally. Nevertheless, the average post-transcriptional repression is only 39 and 47% of that observed for direct targets defined by CARP for miR-1 and miR-122, respectively (set *b*). Moreover, although AGO-binding is significantly increased for these genes, the average magnitude of this increase is minimal (1.1/1.6-fold, compare to 2.5/2.2-fold for set *b*). It is important to note that in numerous conventional analyses of miRNA target sets, including our previous work (Wissink, Smith et al., 2015), all genes in set *c* would have been erroneously declared as direct targets.

An important class of genes are those that have lower steady-state mRNA levels but do not contain predicted target sites for miR-122 or miR-1 (set *d*, Fig 2.3A). Similar to set *c*, they exhibit a negligible magnitude of post-transcriptional regulation even though mRNA levels are significantly reduced in response to the induced miRNAs (1.2- and 1.1-fold repression is post-transcriptional levels for miR-122 and miR-1, respectively). Concordantly, in response to miR-1 or miR-122, we did not observe any significant AGO enrichment in 3'UTRs of these genes (Fig 2.3C,D), nor any evidence

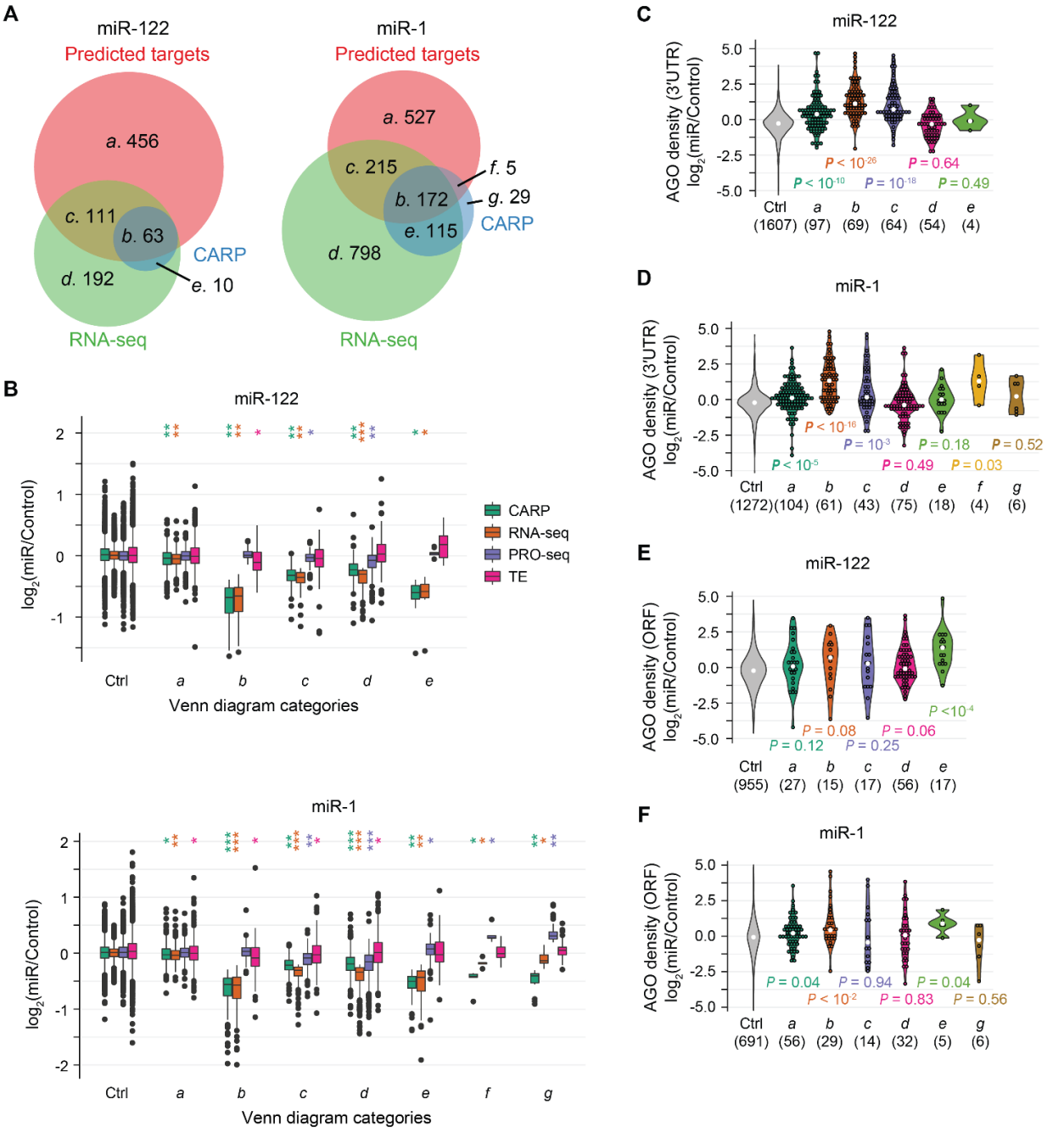

Figure 2.3. CARP enables discovery of different regulatory mechanisms
(**A**) Venn diagram analysis of genes partitioned according to presence or absence of predicted target site for cognate miRNAs (TargetScan; context++ score < -0.2; red set), genes experiencing reduced transcript levels (RNA-seq; q-value < 0.05 and log2FC < -0.2; green set) and post-transcriptional repression (CARP; q-value < 0.05 and log2FC < -0.2; blue set) in response to miR-122 (left) and miR-1 (right). Gene-set labels (*a-g*) and number of genes in each set are displayed. The gene set labels

correspond to: false positive predictions with and without changes in mRNA abundance (a and c, respectively), direct targets (b), indirect transcriptional targets (d), indirect post-transcriptional targets (e), genes subject to both direct and indirect regulation (f), and those subject to indirect transcriptional and post-transcriptional regulation (g). Area is proportional to number of genes in sets. (**B**) Boxplots illustrating distributions of log_2FC for indicated gene-sets in post-transcriptional levels (CARP; green), mRNA abundance (RNA-seq; orange), transcriptional output (PRO-seq; blue) and translation efficiency (Ribo-seq; pink) in miR-122 (top) and miR-1 (bottom) expressing cells. The control set "Ctrl" includes genes that lack significant change in RNA-seq and CARP (q-value > 0.05) whose 3'UTRs are devoid of predicted target sites. Gene-sets *a* through *g* correspond to gene-sets depicted in panels A and B. (**C,D**) Violin plots demonstrating distributions of AGO enrichments at eCLIP peaks in 3'UTRs of different gene-sets in miR-122 (C) and miR-1 (D) expressing cells. The control (Ctrl) gene-set is as described in panels B. The numbers (x-axis) indicate counts of eCLIP peaks found in 3'UTRs of each gene-sets. Statistical significance between each gene-set and the control set were determined using two-sided Wilcoxon rank-sum test. White dots depict median values and colored dots represent individual enrichment values. (**E,F**) Bubble plots summarizing average AGO enrichment in the coding region for each gene-set, across a range of significance (q-value) thresholds (y-axis) in miR-122 (E) and miR-1 (F) expressing cells. The q-values ranging from 0.1 to 0.01 were considered for summarizing Venn diagrams (A) at each cut-off. Size of the bubble corresponds to the number of genes in a set at given q-value cut-off and color indicates median AGO enrichment in coding region of those genes. Every gene-set (bubble) for a given q-value cut-off was compared to the corresponding control set (square) and statistical significance was calculated using two-sided Wilcoxon rank-sum test. * $10\text{-}2 < P < 10\text{-}5$.

of translational repression (Fig 2.3B). Interestingly, these genes also demonstrated subtle repression in transcriptional output. Perhaps this set depicts targets of feed-forward control, which are repressed minimally at both a transcriptional and post-transcriptional level, to ultimately exhibit effective repression in mRNA abundance. As a group, however, they exhibit characteristics indicative of indirect miRNA targets. We also identified genes that are significantly post-transcriptionally repressed, but which lack predicted miRNA target sites (set *e*). The magnitude of post-transcriptional

repression for this gene set matched set *b*, the set of confidently identified direct targets of miR-122/miR-1. We considered two possibilities to explain the observed post-transcriptional repression for set *e*: either they represented indirect targets whose regulation is post-transcriptional, or they represent direct targets lacking conventional miRNA target sites. We recognize that both scenarios might apply to different genes within the set. Notably, we did not observe AGO enrichment in the 3'UTRs of these genes (Fig 2.3C,D), consistent with lack of predicted 3'UTR target sites. Given the absence of AGO enrichment in the 3'UTR, we considered whether these genes might be repressed via target sites within their coding sequences. For miR-1, we detected no AGO2 enrichment within the coding region of genes in set *e* in response to miR-1 (Fig 2.3F), implying that these genes are indirect targets of miR-1, which are regulated post-transcriptionally, and independent of the miRNA pathway. We note that this interpretation is consistent with our observation that many genes post-transcriptionally upregulated in response to miR-1; that is, miR-1 induces widespread indirect post-transcriptional regulation. In contrast, we observed strong AGO enrichment in the coding region of miR-122 regulated genes in set *e* (Fig 2.3E, S2.3C), comparable in magnitude to that observed in the 3'UTR of *bona fide* direct targets (set *b;* Fig 2.3C). We note that set *e* encompasses only few (n=10) genes in response to miR-122. Nevertheless, these results suggest that miR-122 directly regulates a small group via target sites in the coding region.

In response to miR-1, but not miR-122, we observed a small number of genes that are

both confidently predicted targets and significantly repressed post-transcriptionally, but without a detectable change in mRNA abundance (Fig 2.3A,B; set *f*). We hypothesized that these genes are direct targets, but also transcriptionally upregulated, resulting in no net change in transcript levels. Consistent with this idea, we observed strong AGO enrichment, comparable to that seen in set *b*, in 3'UTRs for such genes (Fig 2.3D). The weaker statistical significance of AGO enrichment is likely due to the small number of genes in this category (average fold changes of 2.51 and 2.45, for sets *b* and *f,* respectively). Importantly, we also observed increased transcription of these genes (Fig 2.3B), confirming transcriptional upregulation triggered by miR-1. This set of transcripts illustrate an interesting class of direct targets that are invisible to studies reliant on RNA-seq alone.

The final possible class of genes corresponds to those exhibiting post-transcriptional repression without any significant change in mRNA abundance and without predicted target sites (Fig 2.3A,B, set *g*). We observed this class of genes in response to miR-1 alone. Consistent with the absence of target sites, we did not observe AGO2 enrichment in their 3'UTRs (Fig 2.3D), nor within the coding region (Fig 2.3F) of these genes, suggesting that these genes are indirect targets that are repressed post-transcriptionally, reminiscent of the genes in set *e*. Similar to set *f*, these genes also exhibited transcriptional upregulation, the impact of which is masked by post-transcriptional repression, resulting in no change in mRNA abundance (Fig 2.3B). Collectively, this set of genes represents indirect targets that are regulated at both the

transcriptional and post-transcriptional levels.

It is important to acknowledge that partitioning genes into sets and subsets according to RNA-seq and PRO-seq signals necessitates use of statistical thresholds. Accordingly, we examined whether our interpretations (relating to Fig 2.3A) are robust over a range of reasonable thresholds (Fig S2.3A,C). Overall, our observations remain consistent across a series of statistical thresholds. Notably, even when we assessed the most lenient threshold (q-value < 0.1), we found several predicted targets experiencing reduced mRNA abundance without significant change in post-transcriptional levels (set *c*), suggesting that the large number of set *c* genes we found in Fig 2.3A and 2.3B are not specific to the threshold we used there. Taken together, our approach identifies not only the robust direct targets but also discovers various complex regulatory mechanisms of gene regulation which facilitate a detailed description of miRNA's roles in gene regulatory networks.

MicroRNA-specific targeting of sites located in coding regions

The prevalence and efficacy of miRNA target sites within coding sequence is unclear and controversial. Whereas certain studies indicate that such sites exert a negligible effect, perhaps due to ribosome-mediated removal of RISC from translating ORFs (Gu, Jin et al., 2009), others indicate that such sites are often effective in mediating repression (Hausser, Syed et al., 2013). Here, we revisited this important question,

using our improved ability to define direct miRNA targets.

We observed AGO enrichment within ORFs in response to miR-122 for a small set of genes (Fig 2.3E,F; set *e*); importantly, this set of genes was devoid of 3'UTR sites, and exhibited no evidence of AGO enrichment within the 3'UTR (Fig 2.3C,D). In addition, this set of genes were repressed post-transcriptionally in response to miR-122, thus, it seemed plausible that miR-122 was directly repressing these transcripts using target sites within the ORFs. To systematically examine miRNA-mediated regulation of ORF target sites, we grouped genes based on the location of potential target sites within the transcript and compared the degree of post-transcriptional regulation. In response to miR-122, transcripts containing miR-122 target sites in the ORF were significantly repressed, and the average efficacy of these sites exceeded the efficacy of those predicted to be weakly effective when located within 3' UTRs (Fig 2.4A). For miR-122, we also found that post-transcriptional repression grows stronger with increasing number of miR-122 target sites in ORFs (Fig S2.4A). Consistent with activity of miR-122 ORF sites, we observed AGO enrichment in ORFs containing miR-122 target sites, and the enrichment correlated with the number of sites (Fig S2.4B; data not shown). In contrast, we found no evidence for effective miR-1 target sites within ORFs (Fig. 4A), even when we examined transcripts harboring multiple miR-1 ORF sites (Fig S2.4A), nor did we detect significant AGO enrichment in ORFs containing potential miR-1 target sites (Fig S2.4B). Sites within 5'UTRs were ineffective for both miRNAs.

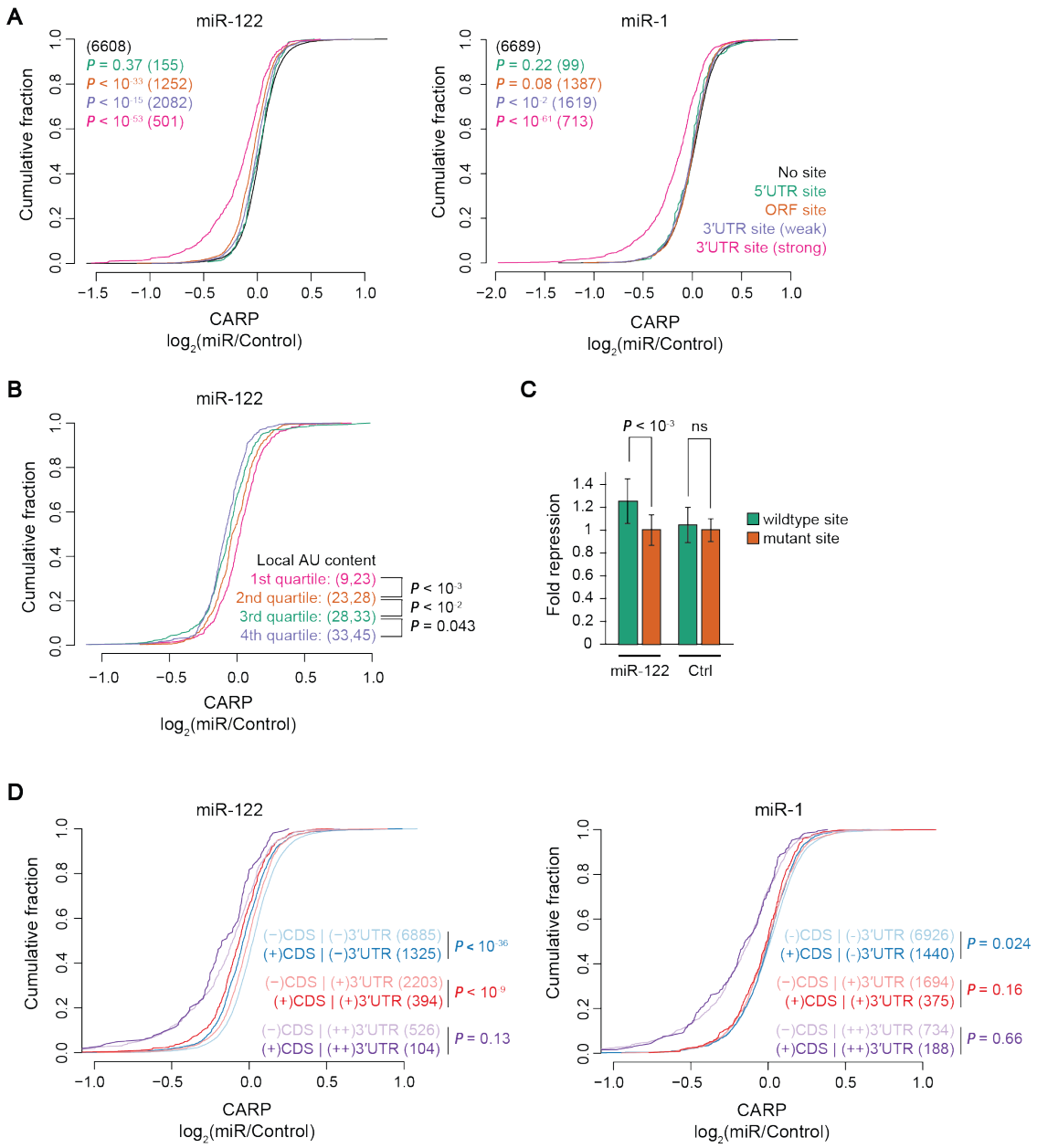

Figure 2.4. MicroRNA-specific targeting of sites located in coding regions. (**A**) Cumulative distributions of post-transcriptional regulation (change in RNA abundance after accounting for changes in transcription; x-axis) of transcripts in response to miR-122 (left) or miR-1 (right). Distributions for mRNAs containing sites in either 3'UTR, ORF or5'UTR , or not site are plotted (pink, blue, etc, respectively). The genes containing predicted 3'UTR sites were divided into two groups according to predicted site efficacy (pink and purple, stronger and weaker sites, respectively). Each group of transcripts containing target sites were compared to the control group without any sites (Wilcoxon rank sum test. Each comparison P value is indicated, and

colored according to the foreground group identity; parenthetical values denote number of genes in each category. (**B**) Cumulative distributions comparing the post-transcriptional regulation of miR-122 targets partitioned by local AU content around the predicted sites; otherwise as described in Panel A. The mRNAs containing target sites to miR-122 were grouped into quartiles based on the number of A+U nucleotides in a 60 nucleotides window centered on the predicted site. The range (in nucleotides) of local AU content of mRNAs in each quartile are presented in parentheses. The observed regulation for each pair of neighboring quartiles was compared (Wilcoxon rank sum test) (P-value). (**C**) Reporter assay showing miR-122-mediated repression of a gene containing a single miR-122 target site in the coding region. The bars represent fold-repression of the firefly luciferase reporter construct containing an intact site normalized to an otherwise identical construct containing a mutated site in presence of miR-122 or non-specific miRNAs. The statistical significance of differences was calculated using Wilcoxon rank sum test (P-value). (**D**) Cumulative distributions of post-transcriptional regulation for mRNAs without 3'UTR sites [(-)3'UTR], predicted weak targets [(+)3'UTR] and predicted stronger targets [(++)3'UTR], with [(+)ORF] or without [(-)ORF] ORF sites. Indicated comparison of regulation (Wilcoxon rank sum test) of difference between mRNA with and without ORF sites was calculated using (P-value). The number of mRNAs in each category are listed in parenthesis.

Next, we investigated whether the ORF sites share properties associated with functional 3'UTR sites. Canonical 3'UTR sites differ in their average efficacy depending on nucleotide sequence and pairing potential with the 5' terminus of the miRNA (referred to as the miRNA seed (Lewis et al., 2005)); we examined 8mer, 7mer-m8 and 7mer-A1 sites (ordered from strong to weaker sites; (Grimson et al., 2007)) located within ORFs. Additionally, miR-122 has been shown to have marginal efficacy on a type of non-canonical target site, referred to as a G-bulged site (Luna, Barajas et al., 2017). We found a consistent pattern for ORF sites to miR-122, with highest average repression elicited by 8mer sites followed by 7mer-m8 and 7mer-A1 sites, with least repression for G-bulged sites (Fig S2.4C). Beyond the type of seed

match, a major additional determinant of 3'UTR target site efficacy is local AU content, with higher AU content correlating with increased strength of miRNA-mediated repression. We found that ORF sites to miR-122 embedded in AU-rich regions exhibit increased post-transcriptional repression, with average repression correlating with local AU content (Fig 2.4B). No such relationships between site type or AU content were observed for miR-1 (Fig S2.4C,D).

To validate the efficacy of miR-122 ORF sites, we selected a subset of miR-122 target sites and generated luciferase reporters containing translational fusions between luciferase and a short region (encoding 26 amino acids) of endogenous sequence containing a potential miR-122 site. We selected sites located in regions spanning a range of AU contents. We also generated negative control variants of these reporters, which contained two to three synonymous mutations within the target site, designed to inactive the site. Reporter assays using the wild-type and control constructs in the presence of miR-122 or a control miRNA indicated repression of wild-type reporters which was specific to miR-122 (Fig 2.4C). However, only two of the sites tested mediated significant repression (Fig S2.4E), consistent with the reduced efficacy of ORF sites compared to 3'UTR sites observed in genome-wide analysis (Fig 2.4A). Importantly, the local AU content and the amount of repression were positively correlated (Pearson's $r = 0.32$) (Fig S2.4F). These results provide additional evidence that miR-122 regulates post-transcriptional expression by targeting ORF sites and that the effective ORF sites share properties of functional 3'UTR sites.

Given the modest impact of ORF sites in post-transcriptional regulation, we wondered whether such sites might function in concert with 3'UTR sites. To answer this question, we evaluated three groups of genes, 1) genes without 3'UTR sites, 2) genes with 3'UTR sites predicted to be weakly effective (context score > 0.2; (Agarwal et al., 2015)) and 3) genes with 3'UTR sites predicted to be more effective (context score < 0.2) , and asked if post-transcriptional regulation of these three groups changed depending on presence or absence of ORF sites. As expected, for transcripts without miR-122 3'UTR sites, those that contained ORF sites were significantly downregulated compared to those without (Fig 2.4D). Interestingly, transcripts containing predicted weak 3'UTR sites were significantly more repressed when their ORFs also contained miR-122 sites (Fig 2.4D). In contrast, the presence of ORF sites in transcripts containing predicted strong 3'UTR sites did not provide additional benefit, particularly for those genes that already are strongly repressed (genes with $\log_2$FC < -0.3). Consistent with our earlier results, we did not see any impact of miR-1 ORF sites on post-transcriptional regulation for transcripts with or without sites within the 3'UTR (Fig 2.4D). Collectively, these observations suggest that miR-122 ORF sites potentiate miRNA-mediated repression for transcripts containing marginal sites within their 3'UTR.

It has been suggested that effective miRNA target sites in ORFs elicit translational repression as a larger component of total repression, when compared to sites within 3'UTRs (Hausser et al., 2013). Our data, however, indicate that targeting of ORF sites

by miR-122 mediated no significant effect on translation (Fig S2.4G). To examine further whether translation status influences the mode of miRNA-mediated repression of ORF sites, we grouped genes based on the degree of translation, approximated using the ratio of ribosome profiling signal and expression in RNA-seq. We observed, irrespective of the amount of translation, a gradual increase in post-transcriptional repression with increasing number of miR-122 ORF sites (Fig S2.4H). Taken together, these results suggest that regulation triggered by miRNAs is mechanistically equivalent whether sites are located in the ORF or the 3'UTR, and not detectably altered by the translation status of the transcript.

Utility of CARP

To establish the general efficacy of CARP, we extended our approach to additional miRNAs. We reanalyzed our miRNA profiling data and selected five more miRNAs (miR-133a, miR-155, miR-302a, miR-372 and miR-373) whose expression is absent or negligible in HEK293 cells. We prepared HEK293 cell lines stably expressing each of these miRNAs. Following induction of miRNAs for one week, we performed RNA-seq and PRO-seq to quantify the global changes in mRNA abundance and transcriptional output and compared the RNA-seq and PRO-seq fold-changes. Similar to the observations for miR-1 and miR-122, the predicted targets of these miRNAs exhibited reduced mRNA abundance without any change in transcription, and

predicted weak targets experienced only a subtle change (Fig S2.5A). Using the likelihood ratio test (McCarthy et al., 2012), we identified gene sets demonstrating significant changes in post-transcriptional levels and compared these genes with sets of predicted targets. The majority of post-transcriptionally down-regulated genes (63-97%) contained predicted target sites for the cognate miRNA (Fig 2.5A – left panel), whereas a majority of the up-regulated genes lacked target sites, consistent with the suitability of CARP in estimating post-transcriptional regulation. Notably, the PRO-seq data revealed different degrees of altered transcription, with certain miRNAs eliciting widespread changes in transcription (Fig S2.5B), which can be attributed to variable degrees of indirect targeting triggered by different miRNAs. These results indicate that CARP effectively deconvolutes post-transcriptional from transcriptional regulation.

Next, we investigated relationships between the three categories of genes, as defined in Fig. 3, for all 7 miRNAs examined in this study; thus, we partitioned genes by whether they were repressed post-transcriptionally, exhibited reduced mRNA levels, and presence of a predicted target site. Consistent with results we observed for miR-1 and miR-122, a large number of confidently predicted targets (set *a*) were not under the control of miRNAs (Fig 2.5B). Similarly, attributes of other sets of genes (sets *b*-*g*) were consistent with the results observed in Fig. 3. In particular, direct miRNA targets identified by CARP (set *b*) exhibit strong post-transcriptional repression without any changes in transcription. The union of sets *b* and *c* represents hundreds of predicted

targets that exhibit reduced mRNA abundance in response to the cognate miRNA, all of which would be considered as direct targets by conventional approaches. However, the improved resolution provided by CARP indicate that only a fraction (33%; set *b*) of those genes exhibit evidence of significant post-transcriptional repression, as opposed to only subtle changes for set *c* genes, for which transcriptional regulation (indirect regulation) is contributing to the observed reduction in transcript abundance (Fig 2.5B - heatmaps). Consistent with this interpretation, the set *b* genes were predicted to contain target sites of lower efficacy than those in set *d* (Fig S2.5C). Nevertheless, many genes in set *c* contained target sites with predicted efficacy equivalent to those in set *d* (Fig 2.4D, S2.5C). Thus, CARP enables partitioning of predicted targets with equivalent sites into true miRNA direct targets and likely downstream indirect targets. These results suggest that CARP offers substantial improvement over existing approaches. We do not rule out the possibility that some of set *d* genes could be true direct targets that CARP is missing because of failing significance cutoff resulted from inherent noise in their measurements. We believe that this type of false-negative would be more prevalent for those miRNAs whose dysregulation does not lead to many indirect changes at transcription, such as miR-122, where incorporating PRO-seq would contribute more noise in absence of transcriptional change. Nonetheless, for the majority of miRNAs that we examined, we observed widespread change in transcription (Fig S2.5B), and hence it is likely that CARP does not miss many true direct targets.

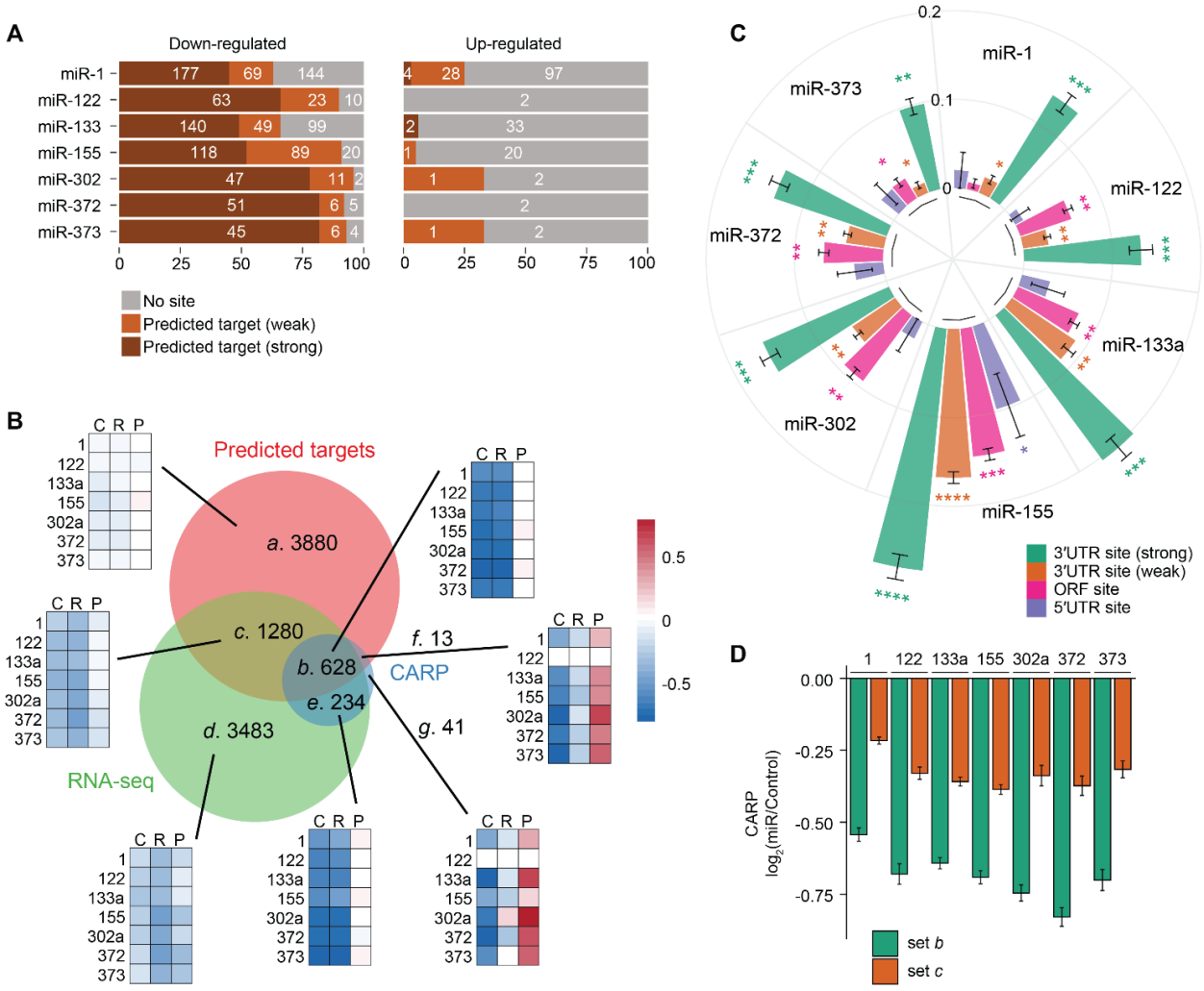

Figure 2.5. Versatility of CARP.

(**A**) Stacked bar plots of the fractions (X-axis) and numbers of mRNAs without seed match, predicted weak targets and predicted stronger targets in each of the two sets: post-transcriptionally downregulated and upregulated genes in response to different miRNAs (Y-axis). (**B**) Venn diagram with areas proportional to the number of genes in each set aggregated across all seven miRNAs, otherwise same as Fig 2.3A, B. Heatmap depicts the median of $\log_2$ (miR/Control) in post-transcriptional levels (CARP; column C), mRNA abundance (RNA-seq; column R) and transcriptional output (PRO-seq; column P) for each miRNA tested (rows). Scale bar denotes color-coded for $\log_2$ fold-changes. There is one heatmap for each set of the Venn diagram. (**C**) Circular bar graph depicting median of $\log_2$ fold-change in response to labeled/indicated miRNAs for transcripts containing potential target sites in either the 3'UTR, ORF or 5'UTR. mRNAs containing 3'UTR sites were divided into two groups according to predicted efficacy (green and orange, weaker and stronger, respectively). The plotted median values were normalized by the median log2 fold-change of

mRNAs without seed match to facilitate comparison between different miRNAs. The ± standard error is represented by error bars. (**D**) Comparison of post-transcriptional changes for context++– matched set *b* and set *c* genes.

In response to most miRNAs, we identified a small cohort of genes (set *f*) that experienced direct miRNA-mediated post-transcriptional down-regulation and indirect transcriptional upregulation, resulting in minimal or no net change in mRNA abundance. Such genes cannot be identified using transcriptome profiling alone, and such targets may represent an important and underappreciated component of miRNA biology.

We observed miRNA-dependent activity of ORF sites in response to miR-122 but not miR-1 (Fig 2.4), thus, we systematically examined whether this activity extended to other miRNAs. We evaluated the efficacy of ORF sites for the new set of miRNAs and compared them with the efficacy of 3'UTR sites of predicted strong targets and predicted weak targets and 5'UTR sites (Fig 2.5C). We observed significant differences between miRNAs in the activity of ORF sites; the miRNAs miR-133, miR-155, miR-302 and miR-372 triggered post-transcriptional downregulation of genes containing ORF sites, whereas the influence of miR-1 and miR-373 on messages containing ORF sites was negligible. In particular, ORF sites for miR-122, miR-133a, miR-155, miR-302a and miR-372 were of comparable or greater efficacy to predicted weak 3'UTR sites. We did not observe any significant activity of 5'UTR sites for most miRNAs; the strongest evidence for effective 5'UTR sites was for miR-155, although this class of sites was weaker than sites within the ORF. Taken together, these data indicate

unexpected variability in the efficacy of ORF sites between different miRNAs, although the average efficacy of these sites was substantially lower than targets identified by CARP that also contain 3'UTR sites predicted to be effective.

Taken together, in addition to robust identification of direct targets, CARP is able to confidently exclude a large number of false-positive predictions and selectively identify false-negatives of conventional target identification approaches. These results demonstrate that CARP can effectively provide a more rigorous list of direct targets by removing large number of false-positives of existing approaches.

Analysis of miRNA regulatory networks

Distinguishing direct miRNA targets from downstream regulatory changes is critical to gaining a systems-level understanding of miRNA gene regulatory networks. Additionally, identifying the specific direct targets (e.g., transcription factors) whose regulation results in these downstream effects is an essential component in understanding the biological roles of miRNAs. In addition to quantifying transcriptional output of genes, PRO-seq also captures and quantifies active enhancers across the genome, and has been used to identify transcription factors contributing to cell state changes (Danko, Hyland et al., 2015). Thus, we explored whether our PRO-seq data could be used to identify transcription factors contributing to the downstream regulatory changes triggered by miRNAs.

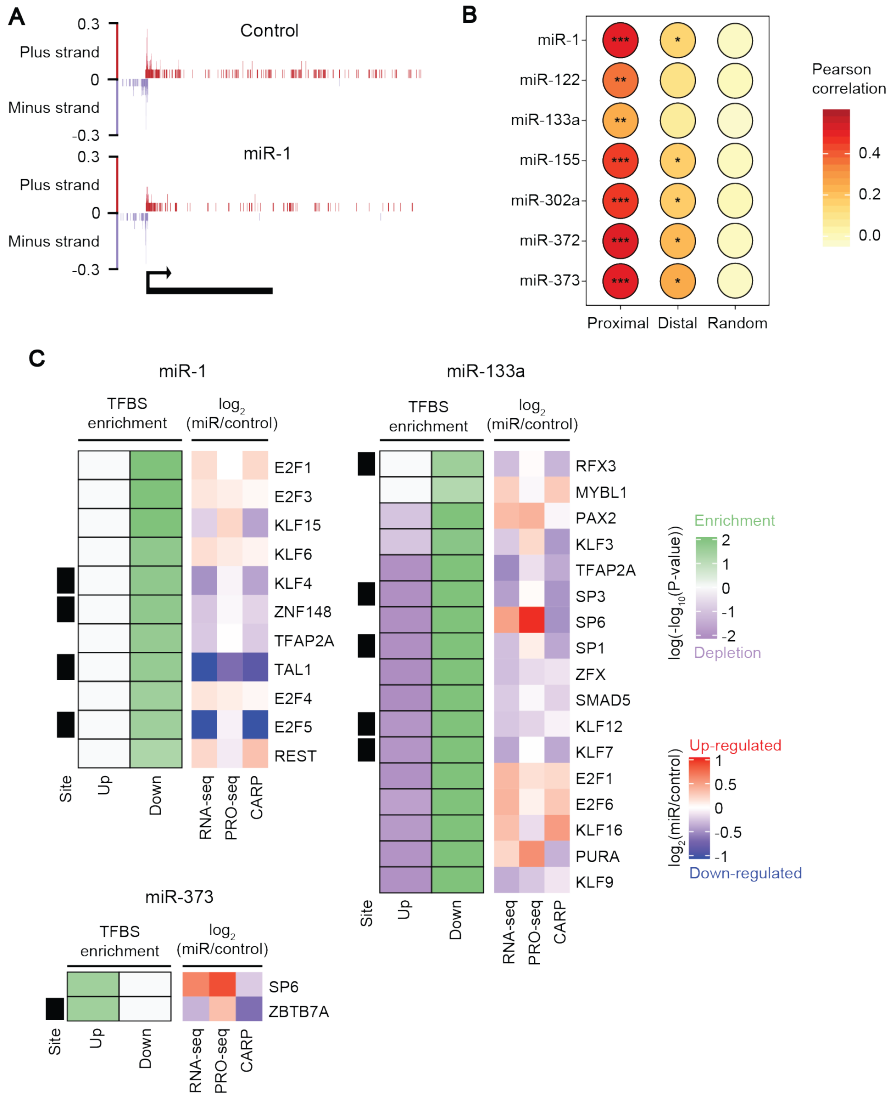

Figure 2.6. Systems-level understanding of miRNA regulatory network using CARP.

(**A**) Genome browser view of PRO-seq signal (counts per million; Y-axis) at a locus of OSMR gene (bottom) demonstrating repression in promoter activity and transcriptional output in response to miR-1. The PRO-seq signal on plus and minus strand of the genome are depicted using red and blue colors, respectively. The divergent transcription is shown at the promoter. (**B**) Heatmap illustrating Pearson correlations between changes in transcriptional activity at proximal or distal peaks with changes in transcriptional output at the nearest gene. The correlation P-value is depicted using a dot (.) or stars (*). Dot (.): 0.01 < P < 10-10; *: 10-10 < P < 10-50; **: 10-50 < P < 10-100; ***: P < 10-100. (**C**) Enrichment (green) or depletion (blue) of transcription factor binding motifs in significantly upregulated (Up) or

downregulated (Down) dREG peak-centers in response to respective miRNAs. The significance is color-coded and was determined with a binomial test using randomly selected open chromatin regions as background, followed by FDR correction. Those binding motifs are depicted that had FDR value less than 0.001. For the included transcription factors, their log2 fold-change in RNA-seq, PRO-seq or CARP are color-coded using blue (downregulated) and red (upregulated) colors. Presence of seed match for the respective miRNAs in the 3'UTR of transcription factors is depicted using black boxes (left).

In order to identify differential enhancer activity in response to miRNA induction, we first identified enhancers using dREG, a tool for prediction of regulatory elements from PRO-seq data (Wang, Chu et al., 2019). We grouped the dREG peaks into two sets; first, proximal peaks, defined as those that are close (within 1.5kb upstream and 0.5kb downstream of) to annotated transcription start sites, and second, the remaining, distal, peaks. We found a strong correlation between the changes in transcriptional activity at dREG peaks and the changes in transcriptional output of the nearest gene (Fig 2.6A,B). This enrichment was stronger for proximal peaks, many of which overlap with the promoter. These results indicate that we can effectively capture changes in transcriptional activity at DNA regulatory elements which contribute to the transcriptional regulation of nearby genes.

Next, we investigated if the differential activity at regulatory elements can be attributed to altered activities of specific transcription factors. Towards this end, we gathered the subset of dREG peaks exhibiting significantly increased or decreased transcriptional activity, and then performed binomial tests to evaluate if the centers of either of these sets of dREG peaks are significantly enriched for putative binding sites

for specific transcription factors, as compared to a background set of randomly selected dREG peaks. In miR-1 expressing cells, we found eleven candidate transcription factors whose putative binding sites were enriched in downregulated peaks, whereas none were enriched in upregulated peaks. Many of these transcription factors exhibited post-transcriptional repression, and many of those contained miR-1 predicted target sites in their 3'UTRs (Fig 2.6C), suggesting that miR-1 downregulates these genes directly which in turn reduces the activity of the encoded transcription factors. We did not find any candidate transcription factors in miR-122 expressing cells, consistent with our previous observations that there is limited indirect targeting at transcription for miR-122. For other miRNAs we examined, we successfully identified transcription factors that likely contribute to the indirect regulation we observed, including for miR-133a, despite the relatively lesser extent of indirect targeting elicited by miR-133a (Fig S2.5B). Perhaps this result indicates that while the miR-133a-mediated repression of transcription factors leads to altered activity at dREG peak, those changes are not being translated to the change in transcriptional output of nearby genes, consistent with reduced correlation between change in activity at proximal peaks and change in transcriptional output of nearby genes (miR-133a in Fig 2.6B). While we did not find any candidate transcription factors in miR-155, miR-302a or miR-372 expressing cells, likely because there were very few dREG peaks with significant change in activity, we found enrichment of two candidate transcription factors, ZBTB7A and SP6, in miR-373 expressing cells, whose putative

binding sites were enriched in those dREG peaks that exhibit increased transcriptional activity (Fig 2.6C). ZBTB7A contains target sites for miR-373 and has been shown to act as a transcriptional repressor, indicating that miR-373-mediated post-transcriptional repression of ZBTB7A promotes transcriptional upregulation of its target genes.

To more comprehensively assess the transcription factors we identified as responsible for miRNA-indeuced indirect regulation, we assessed whether their activities were restricted to those regulatory sites with significant differential activity or extends to globally influence regulatory elements. To investigate this question, we examined all dREG peaks containing putative binding sites for a given transcription factor and compared their activity to the control sample (Fig S2.6A-C). We found that the regulatory effect of many of these candidate transcription factors is widespread, for example, transcriptional activity at KLF15 binding sites was downregulated in miR-1 expressing cells. Similarly, we found increased activity at putative binding sites of ZBTB7A in miR-373 expressing cells. We note that the changes in activity at many dREG peaks is modest, consistent with miRNAs acting to modulate, or fine tune, levels of targets. We believe this additional attribute of PRO-seq data will be of significant value for gaining a comprehensive understanding of the impact of miRNAs at the systems-level, especially in *in vivo* conditions, where the dysregulation of miRNAs would trigger meaningful changes, perhaps of higher magnitude, in order to regulate the phenotypic changes.

DISCUSSION

This study was motivated by the assumption that many miRNAs with consequential functions work, in part, by eliciting substantial downstream regulatory changes, and that most such changes would occur via transcriptional regulation. That is, miRNAs likely function, in part, by controlling one or more transcription factors. Our data, and many published studies, corroborate this assumption. Identifying the direct targets of a miRNA, and distinguishing these direct targets from downstream changes (indirect targets) is challenging. These challenges derive from (i) the subtle nature of miRNA regulation, (ii) false positives and false negatives in target prediction, (iii) the large number of potential targets (iv) the potential for an extensive number of downstream indirect targets. Inevitably, some downstream targets will possess putative target sites, and will often be erroneously considered as direct targets. Our work highlights the extent of such errors, together with many other aspects of miRNA-controlled regulatory networks that are difficult to parse without directly measuring transcriptional regulation that is triggered by a miRNA. The combination of miRNA target prediction and RNA-seq profiling is routinely applied to the study of miRNAs. This study, and others, demonstrates that adding PRO-seq (or related tools) that measure transcription across the genome provides far more reliable definition of the target set of a miRNA, with significant added insights into the overall regulatory network controlled by a miRNA.

An important aspect of this study is the reliance on cell lines that ectopically express miRNAs within the physiological range, rather than using miRNA transfection, as is common in earlier work examining the regulatory impact of miRNAs. Indeed, perhaps the extent of false positives that exist amongst predicted targets derives, in part, from the reliance on training datasets that used transiently transfected miRNAs that likely exceed physiological levels. Another important aspect of this study is the use of multiple genomic tools to profile the regulatory changes mediated by a miRNA – namely, RNA-seq, PRO-seq, ribosome profiling and CLIP-seq, albeit for only miR-1 and miR-122, generating a comprehensive dataset of orthogonal approaches that encapsulates almost all aspects of miRNA-mediated targeting and regulation . Finally, we have provided combined RNA-seq and PRO-seq data for five additional miRNAs; in total, these data serve as an ideal resource to aid in our understanding of miRNA target sites and in further improvement of prediction algorithms.

The precision with which target sites are identified by CARP allows us to more reliably examine non-canonical miRNA targeting; that is, targeting that extends beyond seed-type sites within 3'UTRs. In particular, we have found that certain miRNAs (e.g., miR-122) have large numbers of target sites within coding sequence, whose efficacy is comparable to many canonical 3'UTR sites. Importantly, we validated this observation using AGO CLIP-seq. In contrast, the suite of target sites for other miRNAs including miR-1 and miR-373, were restricted to their conventional location within 3'UTRs. We note that other studies have also hinted at miRNA

specific differences in targeting propensities (REFS). Although there is certainly precedence for miRNA coding sites, we are not aware of compelling evidence that certain miRNAs possess large numbers of effective sites within coding sequence, nor studies that robustly compare the extent of such targeting between a cohort of different miRNAs. Thus, two important conclusions from this study are that the targeting properties of miRNAs are not uniform, and that for some miRNAs, a substantial fraction of their regulatory impact is mediated by target sites within coding sequence. It is also worth noting that for miR-122, mRNAs that possess sites in both the coding sequence and the 3'UTR are markedly repressed. Future studies will be needed to decipher the mechanistic bases for these observations.

The relative contributions of translational repression and accelerated decay to miRNA-mediated regulation is an important question, both because of mechanistic implications, and, pragmatically, due to the reliance on transcriptome profiling in the vast majority of miRNA studies, including CARP. Here, we have investigated two miRNAs, miR-1 and miR-122, using a suite of tools that provides an ideal dataset to rigorously quantify the relative contributions of decay and translational regulation elicited by miRNAs. We find no targets that are exclusively or predominantly regulated by translation, and overall, we estimate that less than 14% of total regulation occurs via translational regulation.

We have identified a set of high confidence direct targets, across multiple miRNAs, which show little or no change in mRNA abundance. Our data indicate that these

targets are simultaneously directly repressed and indirectly activated. Identifying such targets previously was challenging. Overall, such targets constitute X% of the mRNA subject to miRNA-mediated regulation. We argue that the study of these miRNA targets is important in order to understand how miRNAs regulate a complete gene regulatory network. Interestingly, not all of the indirect targeting we observed in such cases was transcriptional; we also found a set of genes that are indirectly regulated at the post-transcriptional level. It is important to note that indirect regulation by miRNAs is unlikely to account for this class of targets, because we saw no AGO enrichment in mRNAs subject to post-transcriptional regulation that lack canonical target sites. Although this study was not intended to investigate the biological roles of any of the miRNAs we used, the frequency with which we found miRNA targets whose repression is balanced suggests that such targets may also be common in native regulatory pathways involving miRNAs.

The primary motivation for this study was to demonstrate the utility of PRO-seq, in combination with RNA-seq, to robustly identify post-transcriptional regulation, and thereby robustly identify miRNA targets. However, in addition to genome-wide quantification of transcription, an integral aspect of PRO-seq data is genome-wide quantification of enhancer activities. This additional feature greatly amplifies the utility of PRO-seq in understanding the regulatory impact of a miRNA. Presumably, many gene regulatory pathways incorporate both miRNAs and transcription factors, and the combination of RNA-seq with PRO-seq is clearly well justified in deconvoluting

miRNA-mediated direct changes from the downstream indirect changes. This combination identifies genes subject to transcriptional and post-transcriptional control, and a profile of active enhancers genome-wide, which. We show that CARP provides a framework for simultaneously measuring regulation occurring at multiple stages of gene expression, which could prove to be a powerful approach for teasing apart gene regulatory networks in vivo.

MATERIALS AND METHODS

Cell culture

Flp-In T-REx 293 (HEK293-derived; Thermo Fisher Scientific) and HEK293T (ATCC) cells were used for all experiments in this study. Cells were cultured at 37°C in a humidified incubator containing 5% CO_2. and maintained in DMEM (Life Technologies) containing 10% FBS (Sigma-Aldrich) and 1% penicillin/streptomycin (Life Technologies); Flp-In T-REx 293 cells were also supplemented with 100μg/ml Zeocin. Cells were passaged every 2-4 days. Cell lines were not tested for mycoplasma contamination.

Synthetic miRNA hairpin constructs

Expression cassettes consisting of the doxycycline-inducible cytomegalovirus (CMV) promoter (derived from the ThermoFisher T-REx system), the intron-containing EF1α 5'UTR, and *Aequorea coerulescens* GFP were cloned into a lentiviral transfer

vector containing a neomycin selective resistance cassette. To clone synthetic miRNA hairpins into the intron, XbaI and XmaI restriction sites were introduced. An artificial hairpin backbone ("A5" from (Fang & Bartel, 2015) was used and the sequence of the mature miRNA was replaced with the sequences of the seven miRNAs used in this study: hsa-miR-1, hsa-miR-122, hsa-miR-133, hsa-miR-155, hsa-miR-302a, hsa-miR-373 and hsa-miR-373. To insert the hairpin sequences, a pair of complementary oligonucleotides was designed containing the sequences with terminal restriction enzyme sites XbaI and XmaI. The oligos were annealed, extended, digested and ligated into the GFP vector. The parent GFP vector lacking a miRNA hairpins was used as a negative control for all experiments.

Generation of stable cell lines expressing specific miRNAs

Lentiviral production

Lentiviruses were generated by transfecting the lentiviral transfer vectors described above along with packaging and envelope plasmids (PAX2 and VSV-G), with Lipofectamine 2000 (Thermo Fisher Scientific) in HEK293T cells. The media was replaced 24 hours later with fresh DMEM containing 30% FBS, and lentiviral supernatants were collected 24 hours later.

Lentiviral transduction and miRNA induction

To generate stable cell lines expressing miRNA constructs, Flp-In T-REx 293 cells were transduced and selected with 1 mg/mL Geneticin for six days, after which the Geneticin concentration was lowered to 0.8mg/ml. To ensure steady-state miRNA

levels, miRNA-GFP cassettes were induced by adding 1μg/mL doxycycline every day for seven days.

Biological replicates were transduced and maintained separately for all experiments. Two biological replicates were generated for experiments employing miR-1 and miR-122, while three replicates were generated for miR-133, miR-155, miR-302a, miR-373 and miR-373 experiments.

PRO-seq

Library preparation

Biological duplicates of cells expressing miR-1, miR-122 and empty vector control were harvested after seven days of induction with doxycycline. For each sample, cells were scraped from one 10am plate in ice-cold PBS. A portion of the cells (20%) was set aside for RNA-seq and the remaining 80% was used for PRO-seq following a protocol adapted from (Mahat *et al*, 2016). Briefly, nuclei were isolated using a buffer containing 0.05% tween-20 and incubated with biotin-labelled nucleotides in a nuclear run-on reaction along with sarkosyl. The total RNA was extracted using Trizol and fragmented using NaOH hydrolysis for 20 mins on ice. The biotin-labelled fragments of nascent RNAs were enriched using Streptavidin M280 beads (Invitrogen) followed by ligation of 3' ends with pre-adenylated DNA adapters (App-GATCGTCGGACTGTAGAACTCTGAAC/3InvdT/) using T4 RNA Ligase 2, truncated K227Q (NEB) in absence of ATP. Following another round of biotin-

enrichment, the 5' ends of the RNA were modified and ligated with 5' RNA adapter (CCUUGGCACCCGAGAAUUCCA). The cDNA was generated by reverse transcription of RNA molecules using SuperScript III RTase (Invitrogen). The libraries were PCR amplified, size selected using PAGE and sequenced on Illumina NextSeq 500.

For miR-133a, miR-155, miR-302a, miR-372 and miR-373, libraries were prepared in biological triplicate (along with the empty vector control) as described above except for the following modifications, adapted from (Booth *et al*, 2018). A 3' RNA adapter (p NNNNNNXXXXXXNNGAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/) containing sample barcodes (Xs) was ligated to the RNA 3' ends using T4 RNA ligase I (Invitrogen), allowing us to pool the samples, streamlining subsequent steps in the protocol. We used unique molecular indexes (UMIs) in both the 3' and 5' RNA adapters (5' adapter: CCUUGGCACCCGAGAAUUCCANNNNN) to minimize ligation bias and facilitate removal of PCR duplicates from the sequencing data.

Data analysis

For miR-1, miR-122, and the empty vector control, the single-end sequencing reads were trimmed to remove adapter sequences using fastx_clipper (-a TGGAATTCTCGGGTGCCAAGG -l 0 -Q33) and cutadapt (-a TGGAATTCTCGGGTGCCAAGG -m 0 -f fastq) for first and the second batch, respectively. All trimmed sequences shorter than 15 nucleotides were removed and

the remaining reads were reverse complemented. The reads originating from PhiX or rRNAs were removed using bowtie-based alignments (REF) to respective reference sequences. The resulting high quality reads were aligned to hg19 genome using bowtie (-p7 -v2 -m1 -q), and the resultant BAM files were used for further analysis.

For miR-133a, miR-155, miR-302a, miR-372 and miR-373 and the empty vector control, reads were first demultiplexed using the in-line sample barcode. The UMI sequences were removed from each read and placed into the read name. Subsequent processing steps were identical as above, with an additional removal of PCR duplicates, defined as reads containing identical UMI sequences and mapping locations.

To calculate transcriptional output from PRO-seq, we considered only gene-body reads, removing all reads originating from promoter or enhancer elements. These transcriptional regulatory elements were predicted using dREG. Bedtools was used to remove reads mapping to these regulatory elements, and the remaining gene-body reads were counted using featureCounts (-F SAF -s 1 -Q 50) of Subread package.

RNA-seq

Library preparation

Cells scraped from 10cm plates in PBS were pelleted and resuspended in 1ml Trizol. Directional RNA-seq libraries were prepared from 1000ng total RNA per sample using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs), with initial polyA+ isolation, by the Transcriptional Regulation and

Expression Facility at Cornell University.

Data analysis

Raw reads were trimmed to remove adapter sequences using fastx_clipper (parameters: -l 15 -Q33 -a GATCGGAAGAGCACACGTCTGAACTCCAGTC) and were aligned to the human genome (hg19) using tophat v2.1.1 (REF; parameters: --library-type fr-firststrand). The featureCounts was used to count reads mapping to exons (parameters: -F SAF -s 2 -Q 50) and introns (parameters: -F SAF -s 2 -Q 50 --fracOverlap 1).

AGO2 eCLIP

Library preparation

Biological duplicates of cells stably integrated with miR-1, miR-122, or the empty vector control were passaged in the presence of 1μg/ml doxycycline for seven (replicate 1) or eight (replicate 2) days. For each sample, two ~70% confluent 10cm plates were UV crosslinked on ice at 400mJ/cm^2, washed with PBS, lifted from the plates, pooled, pelleted, snap frozen in liquid nitrogen and stored at -80°C. For preparing eCLIP libraries, cell pellets were thawed, lysed and prepared with a protocol adapted from (Van Nostrand et al., 2016). Briefly, cell pellets were thawed and lysed in 1ml lysis buffer (50mM Tris-HCl pH 7.4, 100mM NaCl, 0.5% Igepal CA-630) supplemented with protease inhibitor cocktail III (EMD Millipore), treated with RNase I (Ambion), Turbo DNase (Thermo Fisher), and clarified. At all points,

RiboLock RNase inhibitor (Thermo Fisher) was used instead of Murine RNase Inhibitor. For immunoprecipitation, 10μg Ago2 antibody (Anti-AGO2 clone 11A9, MABE253, EMD Millipore) was bound to 100μL washed Dynabeads Protein G. The clarified lysate (950μl) was added to washed antibody-coupled beads and rotated at 4C for 4hr. The IP was washed 2x with high salt wash buffer (50mM Tris-HCl pH7.4, 1M NaCl, 1mM EDTA, 0.5% Igepal CA-630), 1x wash buffer (20mM Tris-HCl pH 7.4, 10mM MgCl2, 0.2% Tween-20), and then washed with 1x FastAP buffer. Beads were treated with FastAP (Thermo Fisher), Turbo DNase, and T4 PNK (NEB). A 3' RNA adapter (5' P-TGGAATTCTCGGGTGCCAAGG/3InvdT) was ligated to the RNA on-bead, and resuspended in LDS sample buffer (Thermo Fisher). Inputs and 10% IP were run on SDS-PAGE, transferred to nitrocellulose membrane, and visualized with western blot to verify pulldown. For preparing eCLIP libraries, 90% of the IP was run on SDS-PAGE and transferred to nitrocellulose membrane. For each sample, the membrane was cut from ~97kDa up to ~275kDa to isolate AGO2-RNA complexes. Membrane slices were treated with proteinase K and urea, extracted with acid phenol:chloroform and cleaned up as described in (Van Nostrand). RNA was reverse transcribed with SuperScript III (Thermo Fisher) with the RT primer 5' CCTTGGCACCCGAGAATTCCA. cDNA was cleaned up and a 3' DNA adapter (5' P-NNNNNNNNNNGATCGTCGGACTGTAGAACTCTGAAC/3InvdT) containing a 10nt unique molecular identifier (UMI) was ligated to the 3' end of the

cDNA, and cleaned up again as described in (Van Nostrand). Approximately 90% of the cDNA was amplified for 16 cycles using sample-specific illumina-compatible primers, ethanol precipitated and run on an 8% PAGE gel. The library smear from ~160bp to ~400bp was cut from the gel, purified, quantified, pooled, and sequenced on the Illumina NextSeq500 with the 75bp kit.

Data analysis

Following removal of UMI sequences from the read sequence, the reads were subjected to adapter trimming using cutadapt (parameters: -a TGGAATTCTCGGGTGCCAAGG -m 18). All the reads mapping to ribosomal RNAs using bowtie2-based alignment were removed and the remaining reads were aligned to the human genome (hg19) using tophat (parameters: -g 1 -p 2 --library-type fr-secondstrand). The alignments were processed using samtools (REF: Li H.*, Handsaker B.*). The PCR duplicates as defined by reads containing same UMI barcode and mapping location, were removed using in-house Perl script. Peaks were called with CLIPPER, and reproducible peaks were identified using IDR. To avoid noise from nonspecific background signal (REF: Friedersdorf & Keene), we calculated AGO density for each peak by computing a ratio of normalized read counts in miRNA-expression cells to that in control cells. Because increased AGO binding results in reduced mRNA levels due to miRNA repression, this measurement of AGO density is underestimated.

Ribosome profiling

Library preparation

Ribosome profiling libraries were prepared with the illumine TruSeq Ribo Profile (Mammalian) Kit. RNA-seq libraries for normalized ribosomal footprints were prepared in parallel. Biological duplicates of cells stably integrated with miR-1, miR-122, or the control were induced with 1μg/ml doxycycline for seven days. To stall ribosomes, cells in 10cm plates were incubated in media supplemented with 100 μg/ml cycloheximide for two minutes at 37°C. Cells were washed in ice-cold PBS, lifted from the plates, pelleted, and lysed in 800μL mammalian lysis buffer on ice for 10min. Lysate was clarified at 13,000rpm, and split into two tubes: (1) 100μL for preparing total RNA libraries, and (2) 200μL for prepare ribosome footprint libraries. Library preparation was performed according to the protocol. Ribosomes were treated with 60U TruSeq Ribo Profile Nuclease to generate ribosome-protected fragments and isolated via size-exclusion with an illustra MicroSpin S-400 HR column (GE healthcare). Ribosomal RNA was depleted from both ribosome-protected fragment and RNA-seq libraries using the illumina Ribo-Zero Gold Kit (Human/Mouse/Rat) (cat #MRZG12324) according to the protocol. Libraries were sequenced on the illumina NextSeq500 with the 75bp kit.

Data analysis

The raw reads were trimmed to remove adapter sequences using cutadapt (-a AGATCGGAAGAGCACACGTC -m 18). Trimmed reads originating from rRNA

were removed using bowtie2 with default parameters for RNA-seq datasets and with "-L 20" and other default parameters for Ribo-seq datasets. Remaining reads were mapped to the hg19 genome and gencode v19 annotated transcripts using following parameters: --no-novel-juncs --transcriptome-index indexFile -p 3 --library-type fr-firststrand. Reads mapping to coding region excluding the ends (initial 45nt and ending 15nt) were counted using featureCounts (-F SAF -s 1 -Q 50 -T 10 --fracOverlap 1) software of Subread (v1.5.1) package. Translation efficiency was calculated by computing ratio of RBF reads and RNA-seq reads.

Luciferase reporter assays

For verifying miRNA activity of stably-integrated miRNAs (Fig S2.1E), a sequence containing two miR-1 target sites was excised from pAG76 (Fahr et al) inserted into pmirGLO Dual-Luciferase vector (Promega) using restriction enzyme sites for SacI and XbaI. The miR-1 sites were disrupted as control. The constructs used in Fig S2.1F contained a perfectly complementary site inserted in pmirGLO vector at XbaI and SalI restriction sites.

For assaying miR-122 ORF target sites, candidate sites were chosen. A 78nt region centered on the candidate miR-122 ORF site was cloned into the coding region of firefly luciferase in pmirGLO. A short linker (amino acid sequence GGGSGGGS) was added to firefly luciferase after the last amino acid, followed by the 26 amino acid

sequence taken from the ORF site, followed by a stop codon. As a control, 2-4nt synonymous mutations were introduced within the miRNA seed sequence to disrupt the site, with attempts to maintain similar codon usage.

For all reporter assays, $1x10^5$ HEK293 cells were seeded per well in 24-well plate 24 h prior to transfection. For the experiments in Fig 2.1 and S1, the HEK293 cells expressing specific miRNAs were transfected with 30 ng of pmirGlo reporter plasmids using 0.5 μL of Lipofectamine 2000 (Life Technologies) and were harvested 24 h after transfection. For the experiments in Fig 2.4 and S4, cells were transfected with 100 nM miR-122 mimics (sequences in SFX) and 6 ng pmirGlo reporter plasmid and harvested 24 h after transfection. Assays were performed using the dual-luciferase reporter assay kit (Promega) and a Veritas Microplate Luminometer (Turner Biosystems).

Data and code availability

The RNA-seq, PRO-seq, Ribo-seq, eCLIP-seq and smallRNA-seq data reported in this paper are available from GEO (GSE140367).

Author Contributions R.P and A.G. conceived this project. R.P. performed most of the experiments and all of the analyses. J.W. performed eCLIP and ribosome profiling experiments, Y.J. assisted R.P with PRO-seq experiments. R.P. and A.G. wrote the paper with assistance from J.W.

Acknowledgements This work was supported by R01GM105668 from NIH and a Cornell Vertebrate Genomics Seed grant to A.G. We thank members of the Grimson laboratory for helpful discussions, and John Lis and his laboratory for helpful discussions regarding PRO-seq protocols. CORE

SUPPLEMENTAL FIGURES

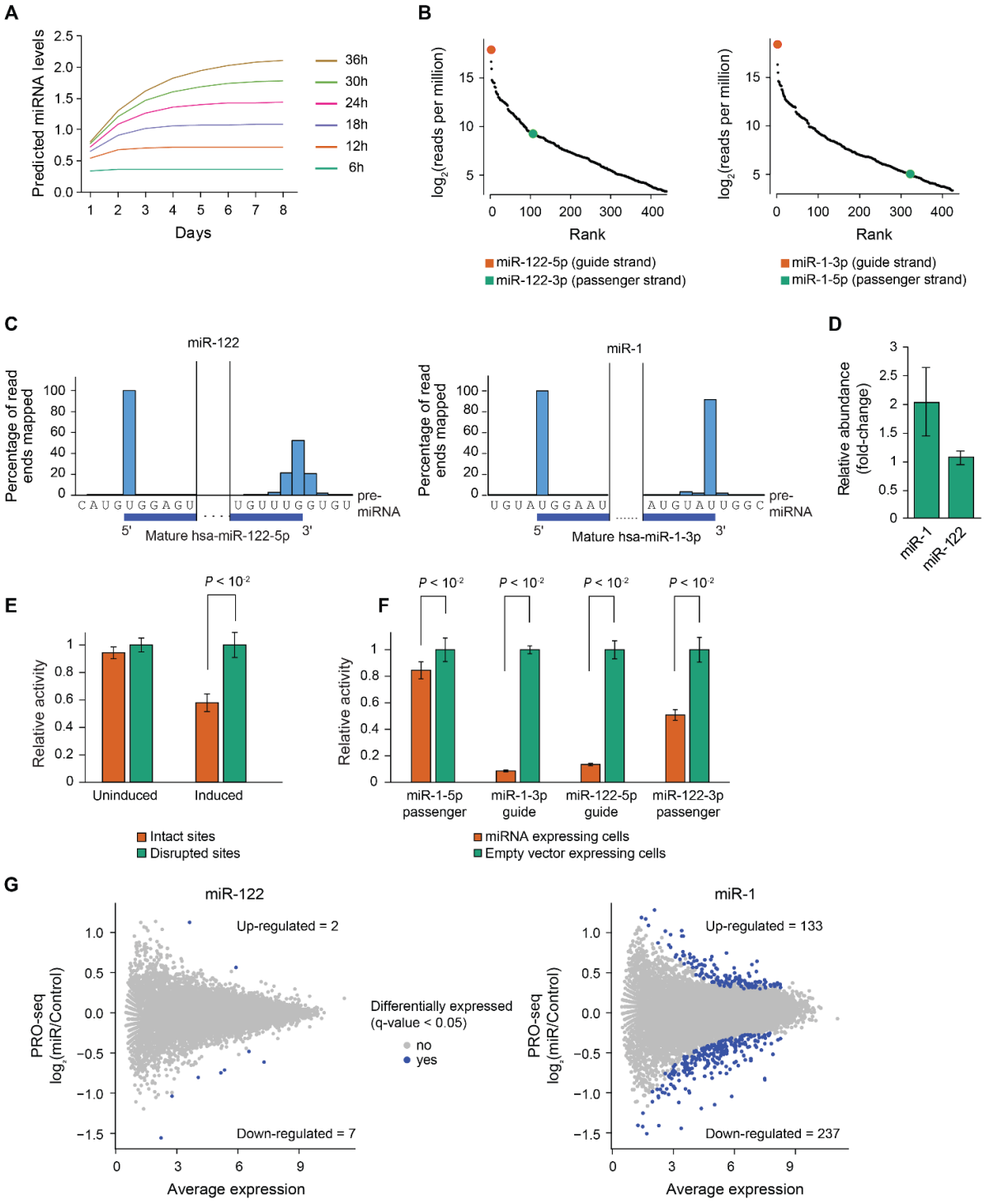

Fig S2.1. Analysis of expression level and processing of induced miRNAs

A. Line graph depicting modeling of steady-state levels for miRNAs with different half-lives.

B. Expression profiling of miRNAs in miR-122 (left) and miR-1 (right) expressing cells.
C. Bar graphs representing relative levels of alternative isoforms of mature miRNAs, miR-122-5p (left) and miR-1-3p (right).
D. qPCR-quantification of relative levels of induced miRNAs compared to the most highly detected endogenous miRNA, miR-10. N=6.
E. Luciferase reporter assays to monitor induced miRNA efficacy. The reporter construct contains a single miR-1 3'UTR target site, or a disrupted target site, and was assayed in cells with and without miR-1 induction. N=6
F. Luciferase reporter assays to evaluate loading of correct miRNA strand. The reporter constructs containing perfect match to either strand of miR-1 or miR-122 were assayed in cells with or without the expression of cognate miRNA. N=3
G. MA plots representing change in transcriptional output upon induction of miR-122 (left) and miR-1 (right).

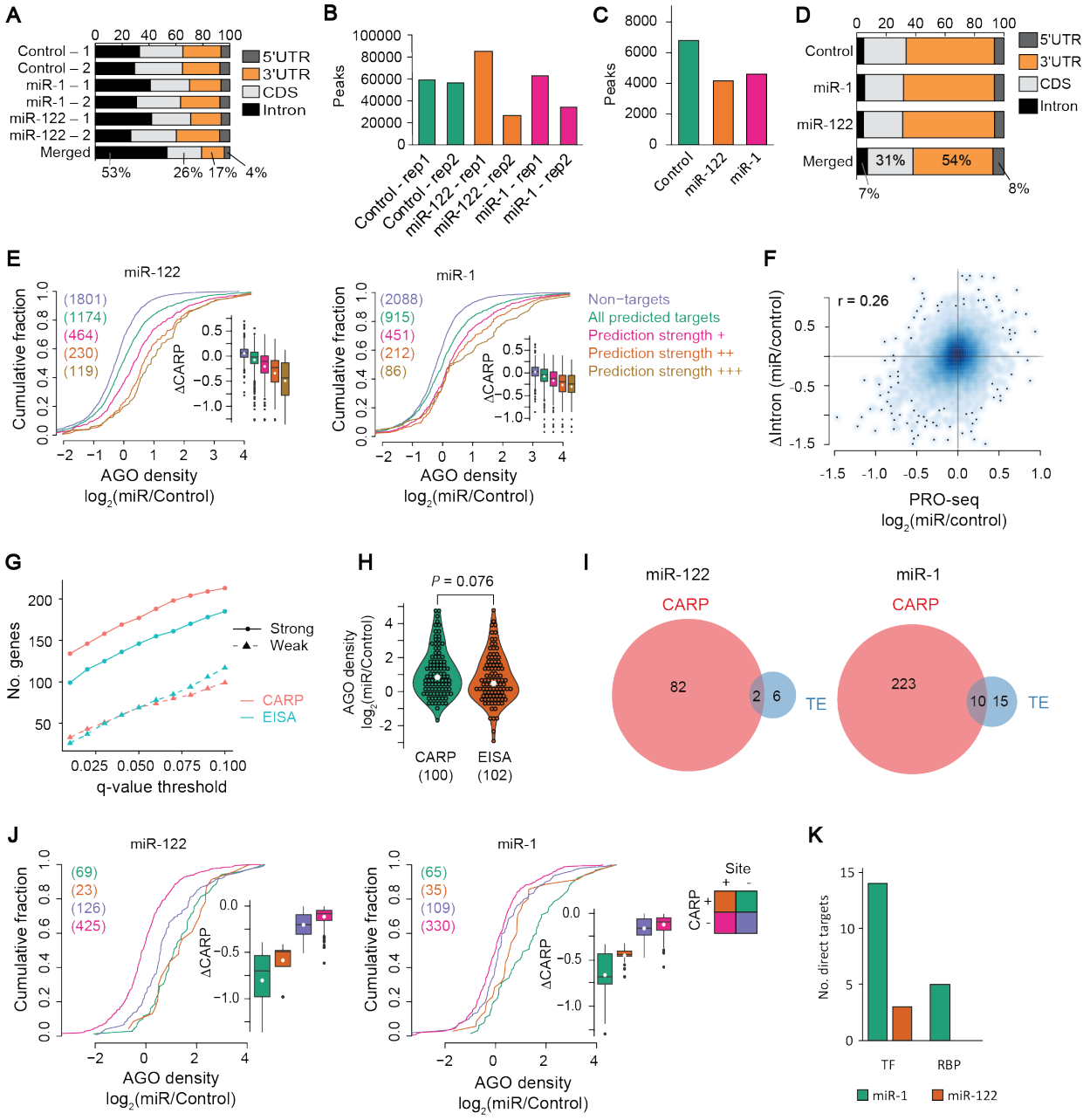

Fig S2.2. Quality filtering of eCLIP data and efficacy of CARP and EISA

A. Distribution of eCLIP peaks in different regions of mRNAs.

B. Variability between replicates in terms of number of eCLIP peaks

C. Number of reproducible eCLIP peaks per sample

D. Distribution of reproducible eCLIP peaks in different regions of mRNAs

E. Cumulative distribution plots depicting increase in AGO enrichment with increasing predicted efficacy of targets. The insets represent decrease in post-transcriptional levels of targets with increasing predicted efficacy of targets.

F. Scatter plot comparing change in transcriptional output measured using PRO-seq with that inferred using intronic reads, with dark blue to light blue colors depicting

high to low two-dimensional density of data. The points represent outliers of density estimation model.

G. The number of CARP- or EISA-positive predicted targets that are repressed at post-transcription.
H. AGO enrichment in 3'UTRs of CARP- or EISA-positive predicted targets that are repressed at post-transcription.
I. Venn diagrams comparing number of genes with significant changes at post-transcription or at translation.
J. Cumulative distributions of AGO enrichment in in 3'UTR of CARP-positive or CARP-negative predicted targets of miR-1 (right) or miR-122 (left). The insets depict fold-change in post-transcriptional levels of the same set of genes.
K. Number of direct targets of miR-1 or miR-122 that have putative activity of transcription factors or RNA binding proteins.

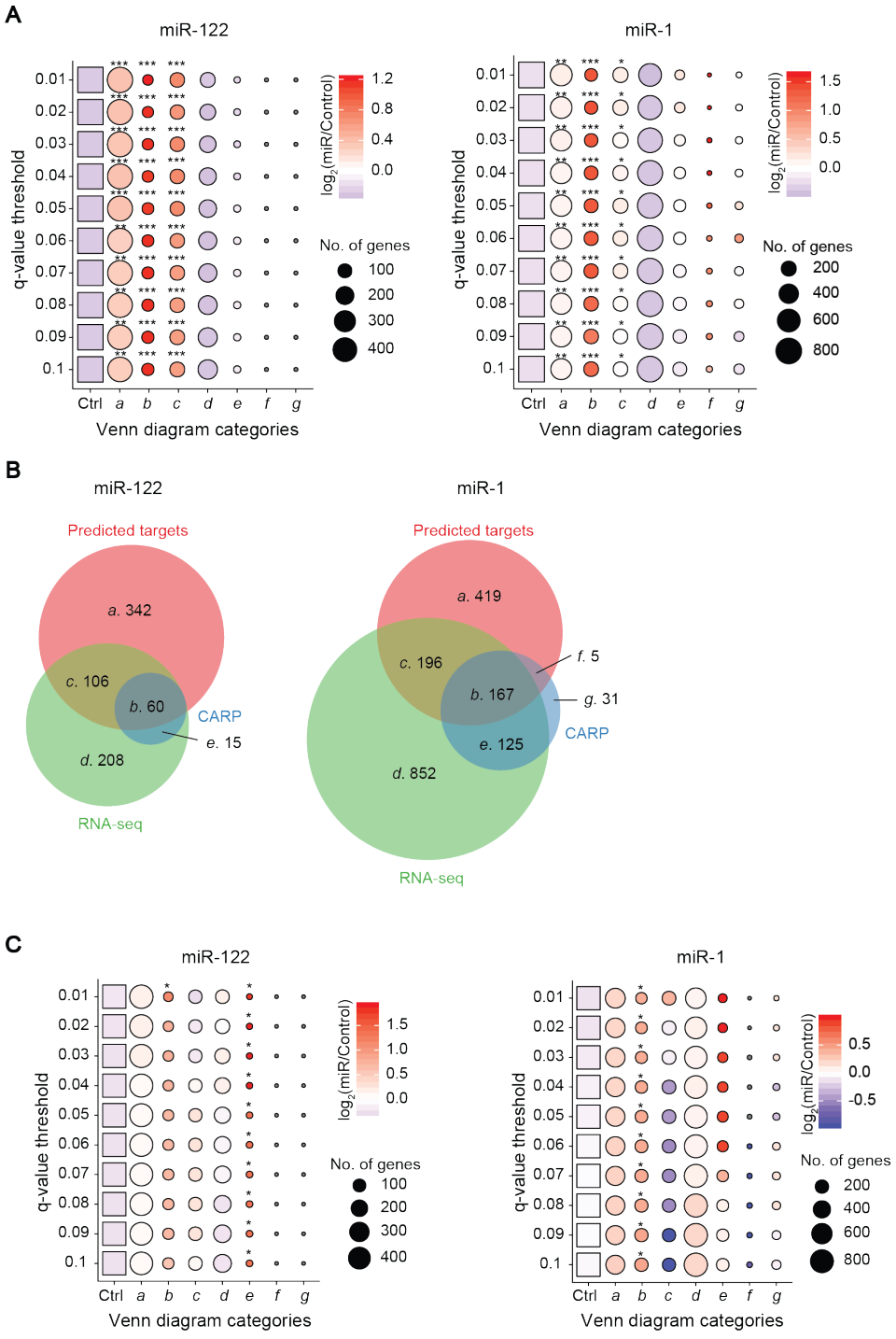

Fig S2.3. Robustness of CARP across different significance thresholds and alternative polyadenylation status

A. Number of mRNAs in a-g sets of Venn diagrams (as in Fig 2.3A) and degree of AGO enrichment in 3'UTRs of each set of mRNAs at different significance cutoffs.

B. Venn diagrams same as in Fig 2.3A, except that the predicted targets that lost predicted target sites due to alternative cleavage and polyadenylation are excluded.
C. Same as A, but representing AGO enrichment in coding region.

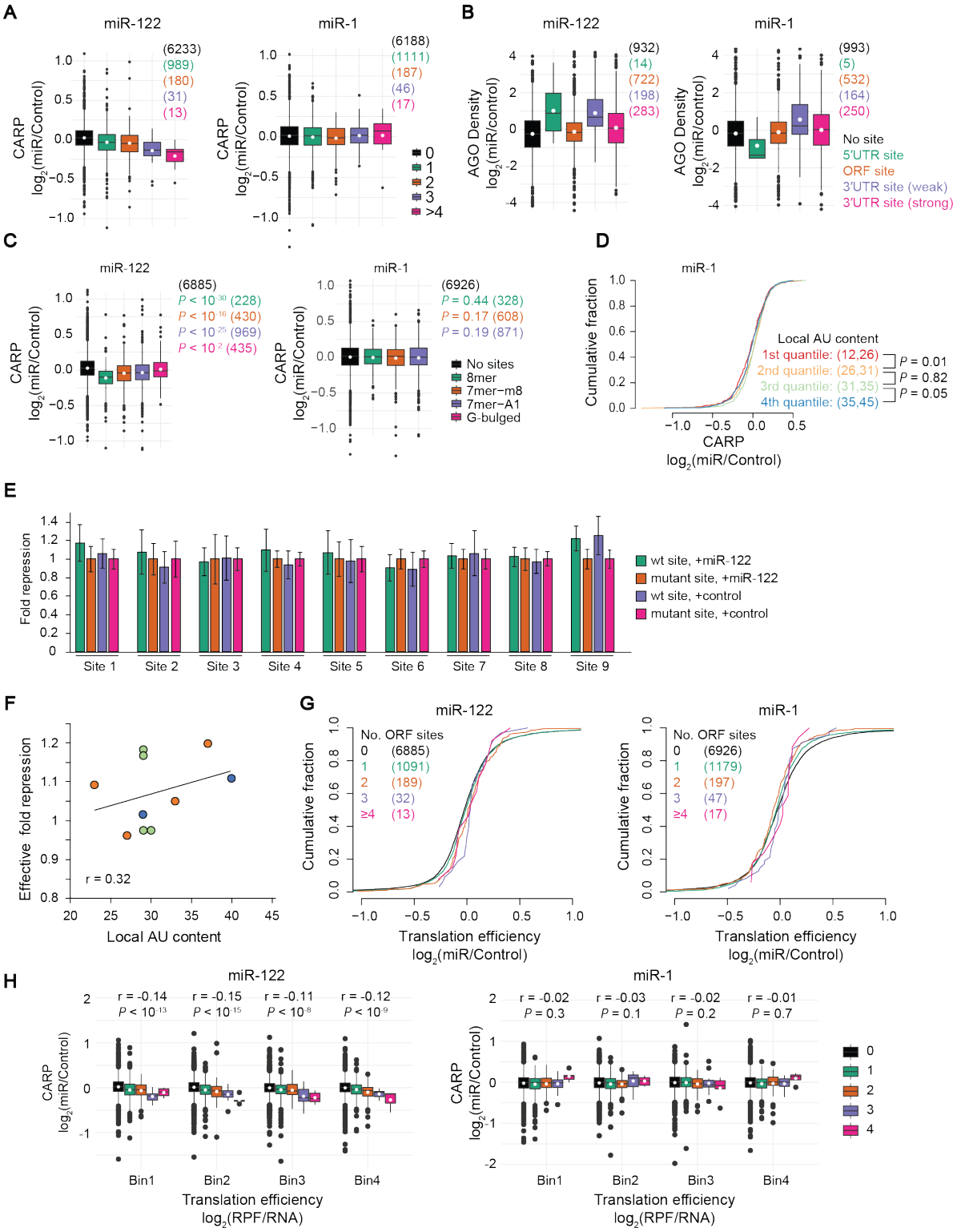

Fig S2.4. Evaluation of ORF site efficacy

A. Change in post-transcriptional repression with increasing number of ORF sites to miR-122 (top) or miR-1 (bottom). Those mRNAs that contain predicted 3'UTR sites are excluded.

B. AGO enrichment in different regions of mRNAs; each group contains mRNAs with predicted target sites exclusively in that region.
C. Post-transcriptional change mediated by different site-types. The mRNAs that contain predicted 3'UTR sites are excluded.
D. Relationship between local AU content around miR-1 ORF sites and post-transcriptional regulation of those mRNAs. The mRNAs that contain predicted 3'UTR sites are excluded.
E. Luciferase reporter assays to assess efficacy of ORF sites. The reporters contain intact or disrupted ORF sites along with native context and are assayed in cells transfected with either the cognate miRNA or a non-specific miRNA mimic.
F. Assessment of relationship between the local AU content and the regulation of ORF sites observed in reporter assays.
G. Cumulative distributions of translation efficiency of mRNA containing zero or increasing number of ORF sites to miR-122 (top) or miR-1 (bottom). The mRNAs that contain predicted 3'UTR sites are excluded.
H. Evaluation of ORF site efficacy for mRNAs with different degree of translation.

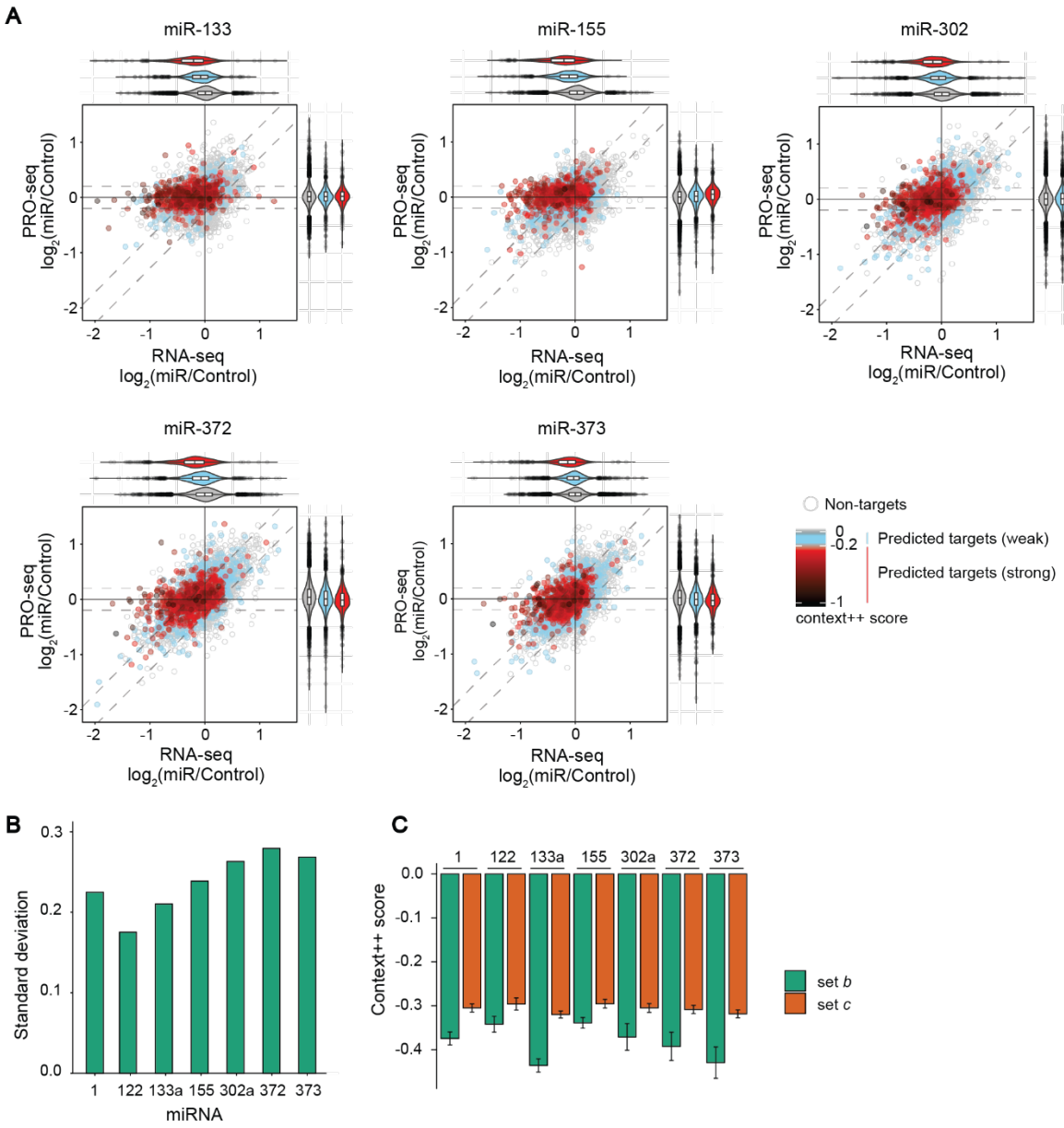

Fig S2.5. Post-transcriptional and transcriptional regulation mediated by different miRNAs used in this study.

A. Dot plots depicting changes in mRNA abundance and transcriptional output for miR-133a, miR-155, miR-302a, miR-372 and miR-373, otherwise same as Fig 2.1C.

B. Degree of transcriptional regulation as measured using standard deviation of PRO-seq $\log_2$ fold-change for different miRNAs.

C. Average prediction strength (context++ scores) for mRNAs that are significantly repressed at post-transcription and those that are not.

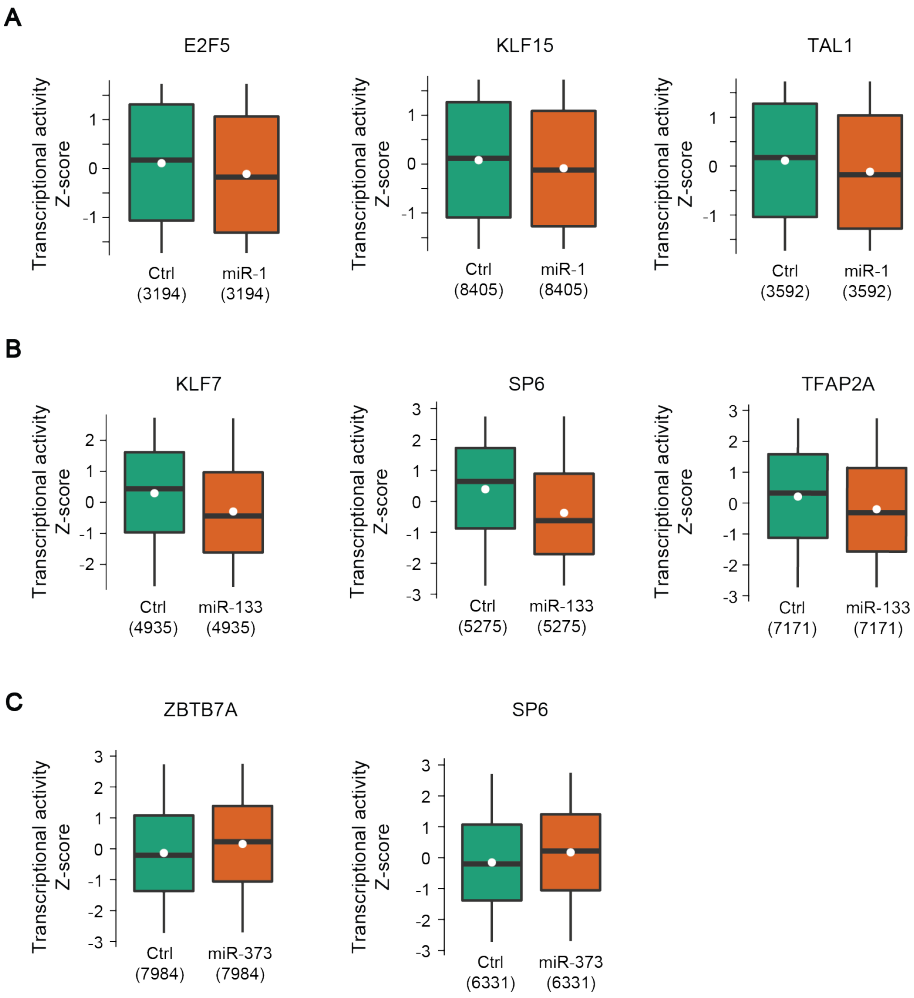

Fig S2.6. Distributions of relative (z-scored) transcriptional activity at putative binding sites of a transcription factor across different samples.

REFERENCES

Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. Elife 4

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136: 215-33

Bartel DP (2018) Metazoan MicroRNAs. Cell 173: 20-51

Bazzini AA, Lee MT, Giraldez AJ (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science 336: 233-7

Bethune J, Artus-Revel CG, Filipowicz W (2012) Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. EMBO Rep 13: 716-23

Booth GT, Parua PK, Sansó M, Fisher RP & Lis JT (2018) Cdk9 regulates a promoter-proximal checkpoint to modulate RNA polymerase II elongation rate in fission yeast. Nat Commun 9: 543

Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of microRNA-target recognition. PLoS Biol 3: e85

Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460: 479-86

Corbett AH (2018) Post-transcriptional regulation of gene expression and human disease. Curr Opin Cell Biol 52: 96-104

Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A (2015) Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods 12: 433-8

Djuranovic S, Nahvi A, Green R (2012) miRNA-mediated gene silencing by

translational repression followed by mRNA deadenylation and decay. Science 336: 237-40

Ecsedi M, Rausch M, Grosshans H (2015) The let-7 microRNA directs vulval development through a single target. Dev Cell 32: 335-44

Fang W, Bartel DP (2015) The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. Mol Cell 60: 131-45

Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science 310: 1817-21

Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19: 92-105

Gaidatzis D, Burger L, Florescu M, Stadler MB (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat Biotechnol 33: 722-9

Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 18: 1139-46

Gosline SJ, Gurtan AM, JnBaptiste CK, Bosson A, Milani P, Dalin S, Matthews BJ, Yap YS, Sharp PA, Fraenkel E (2016) Elucidating MicroRNA Regulatory Networks Using Transcriptional, Post-transcriptional, and Histone Modification Measurements. Cell Rep 14: 310-9

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27: 91-105

Gu S, Jin L, Zhang F, Sarnow P, Kay MA (2009) Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. Nat Struct Mol Biol 16: 144-50

Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466: 835-40

Hausser J, Syed AP, Bilen B, Zavolan M (2013) Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res 23: 604-15

Hobert O (2008) Gene regulation by transcription factors and microRNAs. Science 319: 1785-6

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-23

Kingston ER, Bartel DP (2019) Global analyses of the dynamics of mammalian microRNA metabolism. Genome Res 29: 1777-1790

Kwak H, Fuda NJ, Core LJ, Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339: 950-3

Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P, Kao HL, Gunsalus KC, Pachter L, Piano F, Rajewsky N (2006) A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 16: 460-71

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15-20

Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics 5: 1752-1779

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, Massirer KB, Pratt GA, Black DL, Gray JW, Conboy JG, Yeo GW (2013) Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol 20: 1434-42

Luna JM, Barajas JM, Teng KY, Sun HL, Moore MJ, Rice CM, Darnell RB, Ghoshal K (2017) Argonaute CLIP Defines a Deregulated miR-122-Bound Transcriptome that Correlates with Patient Survival in Human Liver Cancer. Mol Cell 67: 400-410 e7

Lytle JR, Yario TA, Steitz JA (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proc Natl Acad Sci U S A 104: 9667-72

Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ & Lis JT (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc 11: 1455–1476

Mayr C (2017) Regulation by 3'-Untranslated Regions. Annu Rev Genet 51: 171-194

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40: 4288-97

Mishima Y, Fukao A, Kishimoto T, Sakamoto H, Fujiwara T, Inoue K (2012) Translational inhibition by deadenylation-independent mechanisms is central to microRNA-mediated silencing in zebrafish. Proc Natl Acad Sci U S A 109: 1104-9

Moore MJ, Scheel TK, Luna JM, Park CY, Fak JJ, Nishiuchi E, Rice CM, Darnell RB (2015) miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. Nat Commun 6: 8864

Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA 13: 1894-910

Pinzon N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, Seitz H (2017) microRNA target prediction programs predict many false positives. Genome Res 27: 234-245

Riffo-Campos AL, Riquelme I, Brebi-Mieville P (2016) Tools for Sequence-Based miRNA Target Prediction: What to Choose? Int J Mol Sci 17

Roberts JT, Borchert GM (2017) Computational Prediction of MicroRNA Target Genes, Target Prediction Databases, and Web Resources. Methods Mol Biol 1617: 109-122

Schnall-Levin M, Zhao Y, Perrimon N, Berger B (2010) Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3'UTRs. Proc Natl Acad Sci U S A 107: 15751-6

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Harvard FlyBase c, Berkeley Drosophila Genome P, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219-32

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods 13: 508-14

Wang Z, Chu T, Choate LA, Danko CG (2019) Identification of regulatory elements from nascent transcription using dREG. Genome Res 29: 293-303

Wissink EM, Smith NL, Spektor R, Rudd BD, Grimson A (2015) MicroRNAs and Their Targets Are Differentially Regulated in Adult and Neonatal Mouse CD8+ T Cells. Genetics 201: 1017-30

CHAPTER 3[1]

Blurring the line: Neonatal CD8+ T-cells of adaptive immune system elicit innate immune responses

[1]This work is part of a manuscript in preparation. The current authors (¶ denotes equal contributions) on this manuscript are Neva Watson¶, Ravi K Patel¶, Oyebola O Oyesola, Nathaniel Laniewski, Norah L Smith, Kristel J Yee Mon, Cybelle Tabilas, Seth Peng, Samantha P Wesnak, Kristin M Scheible, Elia D. Tait Wojno, Andrew Grimson, and Brian Rudd. Ravi K Patel conceptualized the experiments, prepared samples for RNA-seq and ATAC-seq, analyzed the sequencing data, performed bioinformatic analyses, interpreted data and prepared the manuscript and figures. Neva Watson conceptualized the experiments, maintained the mouse lines, performed mouse work, performed FACS sorting, generated samples for RNA-seq and ATAC-seq, and prepared the manuscript and figures. Jennifer K. Grenier performed ATAC-seq library preparation. Andrew Grimson and Brian Rudd helped to conceptualize experiments, analyze data, and prepare the manuscript.

ABSTRACT

CD8+ T-cells, an important component of the adaptive immune system, protect the host against invading pathogens, like viruses, and cancer, and function in an antigen-dependent manner. Stimulation of these cells via inflammatory cytokines alone, in absence of the cognate antigen, elicits only a weak response. Here, we have demonstrated that the CD8+ T-cells produced early in life undergo a strong response to inflammatory cytokines, in an antigen-independent manner, a process referred to as bystander activation. Analyses of RNA-seq and ATAC-seq revealed that (i) the cells produced up to the age of 5-7 days (neonatal cells) exhibit a transcriptomic profile and chromatin landscape that is distinct from those of adult cells, (ii) the neonatal cells express unusual cytokines that are not previously reported in cells of the adaptive immune system, (iii) the expression profile of neonatal cells resemble that of certain innate immune cells, and (iv) BACH2/JUN transcription factors are responsible for bystander activation of neonatal cells. We validated these findings using *in vitro* and *in vivo* experiments. Ectopic expression of an RNA binding protein (Lin28b) in adults, which is typically upregulated in fetal lymphoid tissues, reprogrammed the adult cells to behave like neonatal cells in bystander activation, as shown by highly similar expression profile and chromatin landscape between neonatal and Lin28b cells and by *in vivo* and *in vitro* experiments. Findings in this chapter reveal the neonatal cells elicit innate-like response, in addition to adaptive immune response in present of an antigen.

INTRODUCTION

Vertebrates have evolved a complex and robust immune system, comprising of innate and adaptive cell types and responses, which protect from the invading pathogens. Once a pathogen is encountered, cells of the innate immune system, such as macrophages and dendritic cells become activated, and secret various inflammatory cytokines, such as IL-1, IL-6, IL-10, IL-12, IL-13, IL-22 etc (Rael & Lockey, 2011; Zenewicz, 2018). Secretion of these cytokines activates other cells of the innate immune system in an antigen-independent manner, resulting in an inflammatory response to the pathogen. While the innate immune system is relatively primitive and nonspecific in nature, and often provides the first line of defense during an initial period of infection, the adaptive immune system is a more complex system that is responsible for clearing of many pathogens with high precision and specificity (Dempsey *et al*, 2003; Janeway, 2001). The CD8+ T-cells are one of the crucial pillars of the adaptive immune system. These cells provide protection against intracellular pathogens, such as viruses, and cancer. Each cell expresses on its surface a unique T-Cell Receptor (TCR) that recognizes a specific antigen. Interaction between the antigen and the TCR is necessary for activation of the naïve CD8+ T-cells in order to mount a robust response to the pathogen. The activation of T-cells via a TCR triggers a signaling cascade and genome-wide alterations in chromatin landscape and the transcriptome (Smith *et al*, 2015), which results in a massive proliferation of antigen-

specific cells, each producing thousands of effector CD8+ T-cells whose function is to kill infected cells (Parish & Kaech, 2009). One of the major players in differentiation of effector cells is the JUN-family of transcription factors, which mediates activation of effector genes, such as interferon-gamma (*Ifng*) upon TCR-stimulation (Rincón & Flavell, 1994; Falvo *et al*, 2000). In naïve CD8+ T-cells, the expression of effector genes is attenuated by binding of a transcriptional repressor, BACH2 at the JUN-binding sites (Oyake *et al*, 1996; Roychoudhuri *et al*, 2016). Upon stimulation by TCR, a serine/threonine-kinase, AKT, phosphorylates BACH2, facilitating inactivation and nuclear export of BACH2, which makes JUN-binding sites accessible (Rincón & Flavell, 1994; Macián *et al*, 2001). This process is followed by transcriptional activation of effector genes by binding of JUN at the regulatory regions of effector genes, resulting in differentiation of effector cells. The effector cells secret cytokines such as interferon-gamma (IFNg) and tumor necrosis factor-alpha (TNFa), and various granzymes (Granzyme A, Granzyme B, Granzyme M), which promotes death of the target cells.

We have shown previously that, in mice, the CD8+ T-cells developed early in life (up to the age of 5-7 days) exhibit a unique expression profile and chromatin landscape and are intrinsically programmed to respond aggressively to TCR-stimulation compared to their adult counterparts (Wissink *et al*, 2015; Appendix). We also showed that these age-related differences are due to the different developmental origins of these cells; CD8+ T-cells are produced by fetal liver-derived hematopoietic stem cells

(HSCs) in early life and by bone marrow-derived HSCs later in life (Smith *et al*, 2014; Wang *et al*, 2016; Smith *et al*, 2018). Our previous and ongoing work have demonstrated that microRNAs are one of the major regulators of these age-related differences (Wissink *et al*, 2015; unpublished). It has been shown that the TCR-mediated stimulation benefits from the simultaneous stimulation by inflammatory cytokine, IL-12 (Cui *et al*, 2009; Curtsinger *et al*, 1999), whereas the stimulation by IL-12 in absence of the relevant antigen (referred to as bystander activation) triggers a minimal response. In this work, we explored the possibility that the naïve CD8+ T-cells from neonates (5-7 days old mice) also respond more aggressively to bystander activation than their adult counterparts. Indeed, we found that the neonatal cells, when stimulated with inflammatory cytokines, IL-12 and IL-18, secret higher amounts of traditional cytokines (e.g. IFNg and GM-CSF) and granzymes compared to adults. Using RNA-seq and ATAC-seq, we show that the neonates demonstrate a unique expression profile and chromatin landscape that resemble that of cells of the innate immune system, and that the expression and epigenetic differences when compared to adults are even more pronounced after bystander activation. We also demonstrate that these age-related differences are driven by higher expression of an RNA-binding protein (Lin28b), which is expressed at high levels in fetal lymphoid tissues and whose well-known role is in inhibiting a microRNA, let-7. The analyses of ATAC-seq data accompanied by experimental validations revealed that the BACH2/JUN axis is involved in the extensive activation of neonatal cells to inflammatory cytokines.

Collectively our results display a novel role of neonatal CD8+ T-cells in innate immune system.

RESULTS

Neonatal CD8+ T-cell exhibit a unique program of bystander activation

To evaluate the response of CD8+ T-cells of different origins to inflammatory cytokines, we isolated adult (2 – 4 months old), neonatal (5 to 7 days old) or adult CD8+ T-cells expressing Lin28b (Lin28), and cultured them overnight in the presence of IL-2 alone (control) or with the addition of IL-12 and IL-18 (inflammatory cytokines), in absence of any antigen. Flow cytometry analysis revealed that the cells treated with IL-12 and IL-18 produced high amounts of cytokines characteristic of T-cell activation, namely, IFNg and GM-CSF (Fig 3.1A). Interestingly, adult cells produced reduced amounts of these cytokines compared to the neonatal and Lin28 cells. As expected, the expression of these proteins was negligible in cells treated with IL-2 alone. These results suggested that neonatal and Lin28 cells, but not adult cells, respond robustly to inflammatory cytokines (bystander activation), a hallmark of innate immune cells.

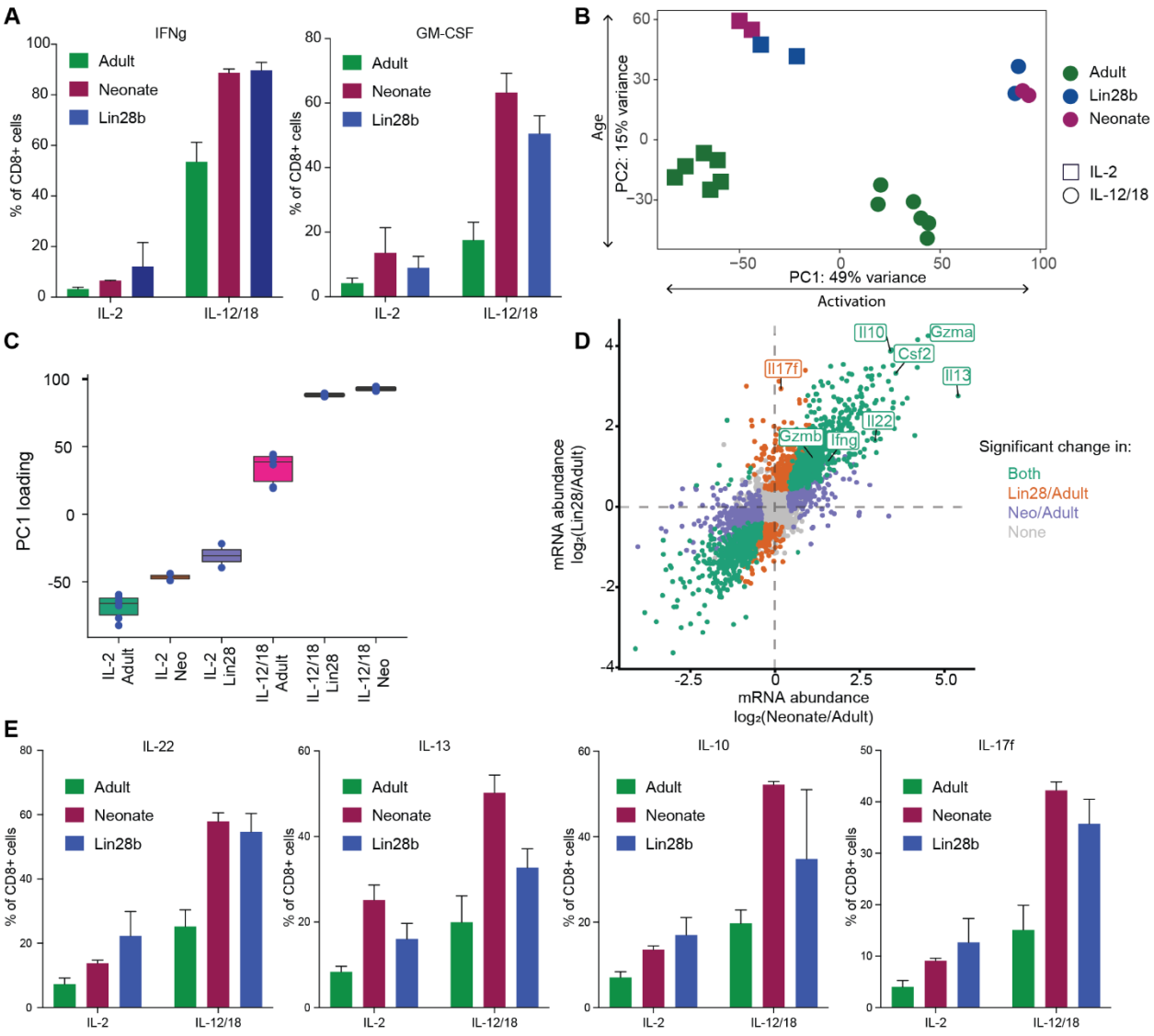

Fig 3.1. Neonatal CD8+ T-cell exhibit unique program of bystander activation (**A**) Flow cytometry analysis of IFNg (left) and GM-CSF (right) expression in adult, neonatal or Lin28 CD8+ T-cells with or without IL-12 and IL-18, the inflammatory cytokines. (**B**) PCA analysis of RNA-seq, with PC1 indicating activation status and PC2 separating age. (**C**) Distributions of PC1 loadings for different samples. (**D**) Comparison of gene expression change in neonatal and in Lin28 cells as compared to adult cells. The traditional (IFNg and GM-CSF) and innate cytokines (IL-10, IL-13, IL-17f and IL-22), and granzymes (Granzyme A and Granzyme B) are labelled. (**E**) Flow cytometry analysis of innate cytokine expression in adult, neonatal or Lin28 CD8+ T-cells with or without IL-12 and IL-18, the inflammatory cytokines.

To investigate the genome-wide response of CD8+ T-cells to bystander activation, we

profiled the transcriptomes of adult, neonatal and Lin28 cells treated with or without IL-12 and IL-18. Principal component analysis (PCA) separated the cells by treatment conditions (PC1) and age (PC2) (Fig 3.1B). Interestingly, the neonatal and Lin28 cells clustered together in both the control condition and stimulation by inflammatory cytokines, whereas the adult cells exhibited a distinct transcriptional response. Additionally, based on the first principal component which separated control and activated cells, the adult cells treated with IL-12 and IL-18 clustered in between the control and activated neonatal cells (Fig 3.1C), suggesting that adults demonstrated a muted genome-wide transcriptional response to innate cytokines, consistent with reduced expression of activation markers, IFNg and GM-CSF, in adults compared to neonates after bystander activation (Fig 3.1A).

To further characterize the response of neonatal and Lin28 cells to bystander activation and to investigate how this response is different from adult cells, we compared the log fold-change of genes differentially expressed in neonatal cells and in Lin28 cells as compared to adult cells. Consistent with the more robust response of neonatal cells to bystander activation compared to their adult counterparts, we observed higher levels of IFNg (*Ifng*), Granzyme A (*Gzma*) and Granzyme B (*Gzmb*) in neonates compared to adult cells (Fig 3.1D). Unexpectedly, we found that neonatal cells also expressed high levels of *innate cytokines*, including IL-13, IL-22, IL-10 and IL-17f, whose secretion is a characteristic of cells of the innate immune system, and which are not expected to be expressed in CD8+ T cells. Using flow cytometry

analyses, we also showed that the increase in mRNA levels were recapitulated in protein levels for granzymes, cytokines characteristic of CD8+ T cells and cytokines not previously reported in these cells (Fig 3.1E). These results indicated that neonatal CD8+ T cells undergo a unique program of bystander T-cell activation (Fig 3.1). Interestingly, adult cells expressing Lin28b displayed a similar program of bystander activation to neonatal cells, suggesting that this program is developmentally regulated by high levels of Lin28b in early life.

Neonatal CD8+ T-cells resemble innate immune cells

Given that neonatal and Lin28 cells express innate cytokines, we wondered if these cells exhibit traits normally exclusive to cells of the innate immune system. To explore this possibility, we performed gene set enrichment analysis (GSEA) for genes that are typically highly expressed in the innate branch of the immune system. A previous study profiled the transcriptome of human cells from innate and adaptive branches of the immune system and characterized marker genes for "innateness" and "adaptiveness" (Gutierrez-Arcelus *et al*, 2019). Using GSEA analysis, we demonstrated that the genes associated with innateness in humans were strongly enriched in CD8+ T-cells from neonatal and Lin28b mice compared to their adult counterparts (Fig 3.2A,B) and this enrichment was even stronger in response to inflammatory cytokines. We found an opposite trend for genes associated with adaptiveness, which were

strongly upregulated in adult compared to neonatal and Lin28b cells (Fig S3.1). These results indicate that the neonatal cells exhibit an expression pattern characteristic of innate immune cells and that Lin28b facilitates this unique phenotype in early life.

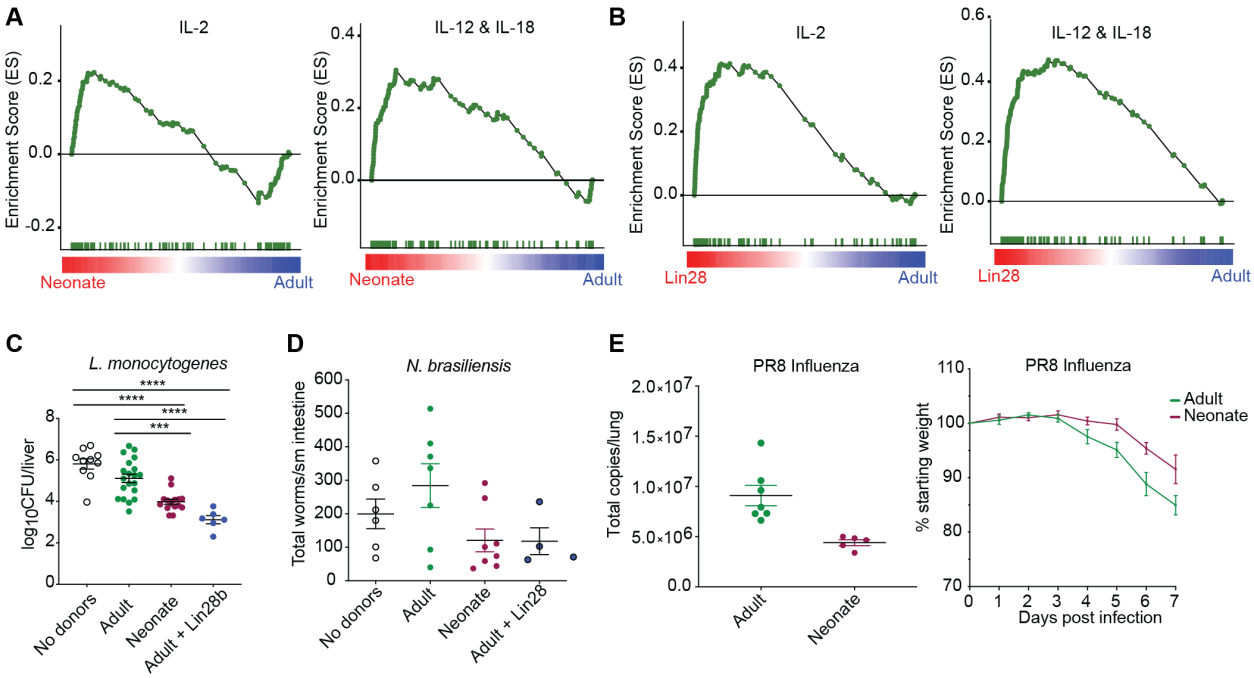

Fig 3.2. Neonatal CD8+ T-cells resemble innate immune cells
(**A & B**) GSEA enrichment score plots for "innateness"-related genes when compared to neonatal (A) or Lin28 (B) cells treated with (right) or without (left) inflammatory cytokines. (**C-E**) Pathogen burden analysis using either *L. monocytogenes* (C), *N. brasiliensis* (D) or PR8 Influenza (E). The weight-loss statistics are plotted for animal that either received adult or neonatal cells and were infected with PR8 Influenza (E, right).

Since neonatal cells exhibit a robust response to *in vitro* stimulation by inflammatory cytokines and exhibit the transcriptome similar to that of innate cells, we next wondered if the neonatal CD8+ T-cells behave like innate cells *in vivo*, in addition to their classical role as part of the adaptive immunity. We hypothesized that if the

neonatal CD8+ T-cells indeed exhibit innate-like response *in vivo*, then, when the mice are infected with pathogens that the neonatal cells cannot recognize, the neonatal cells should still get activated by the inflammatory cytokines secreted by innate immune cells upon infection and should provide protection against the irrelevant pathogens. To test this hypothesis, we used gBT-I CD8+ T-cells, which contain T-cell receptors (TCRs) specific for a gB epitope of Herpes Simplex Virus-1 (HSV-1), and monitored their response to unrelated pathogens. To access the *in vivo* response of these cells, the gBT-I CD8+ T-cells isolated from adult, neonate and Lin28b mice were transferred into TCRa-/- recipients, and the recipients were subsequently infected with an unrelated pathogen, wild type bacteria, *L. monocytogenes*, which is not recognized by the TCR of transferred gBT-I CD8+ T-cells. In accordance with their *in vitro* behaviors, neonate and Lin28 cells displayed reduced bacterial burden in the liver compared to adult cells (Fig 3.2C). We then asked if the unique program of bystander activation allows neonatal cells to protect against other types of unrelated pathogens such as worms and viruses. To address this, we repeated this experiment, infecting recipients with wild type strains of *N. brasiliensis* or Influenza virus. Mice receiving neonatal or Lin28 cells also displayed reduced *N. brasiliensis* burdens in the small intestine (Fig 3.2D). Similarly, the mice receiving neonatal or Lin28 cells exhibited decreased weight loss together with reduced Influenza viral copies in the lung compared to mice receiving adult cells (Fig 3.2E). These findings demonstrate that neonatal and adult cells expressing Lin28b can convey enhanced bystander protection against bacterial,

helminth and viral infections, a characteristic of innate immune cells.

Response to inflammatory cytokines is accompanied by changes in chromatin landscape

We have previously shown that the CD8+ T-cells produced in early life exhibit a distinct chromatin landscape compared to those produced later, and used these data to identify transcription factors that likely underlie the differences between the different developmental layers of CD8+ T cells. To investigate if neonatal CD8+ T cells are epigenetically programmed to behave differently to cytokine stimulation, we employed ATAC-seq. After sequence alignments and quality filtering, we identified peaks of ATAC-seq reads, which correspond to regions of open chromatin. We obtained a set of 40,612 consolidated ATAC-seq peaks across all samples, which comprised of 11,107 promoter-proximal and 29,505 distal peaks, likely representing promoter and enhancer elements, respectively. The PCA of the chromatin accessibility, as approximated by the number of ATAC-seq reads, at these peaks demonstrated that the first principal component distinguishes the activation status

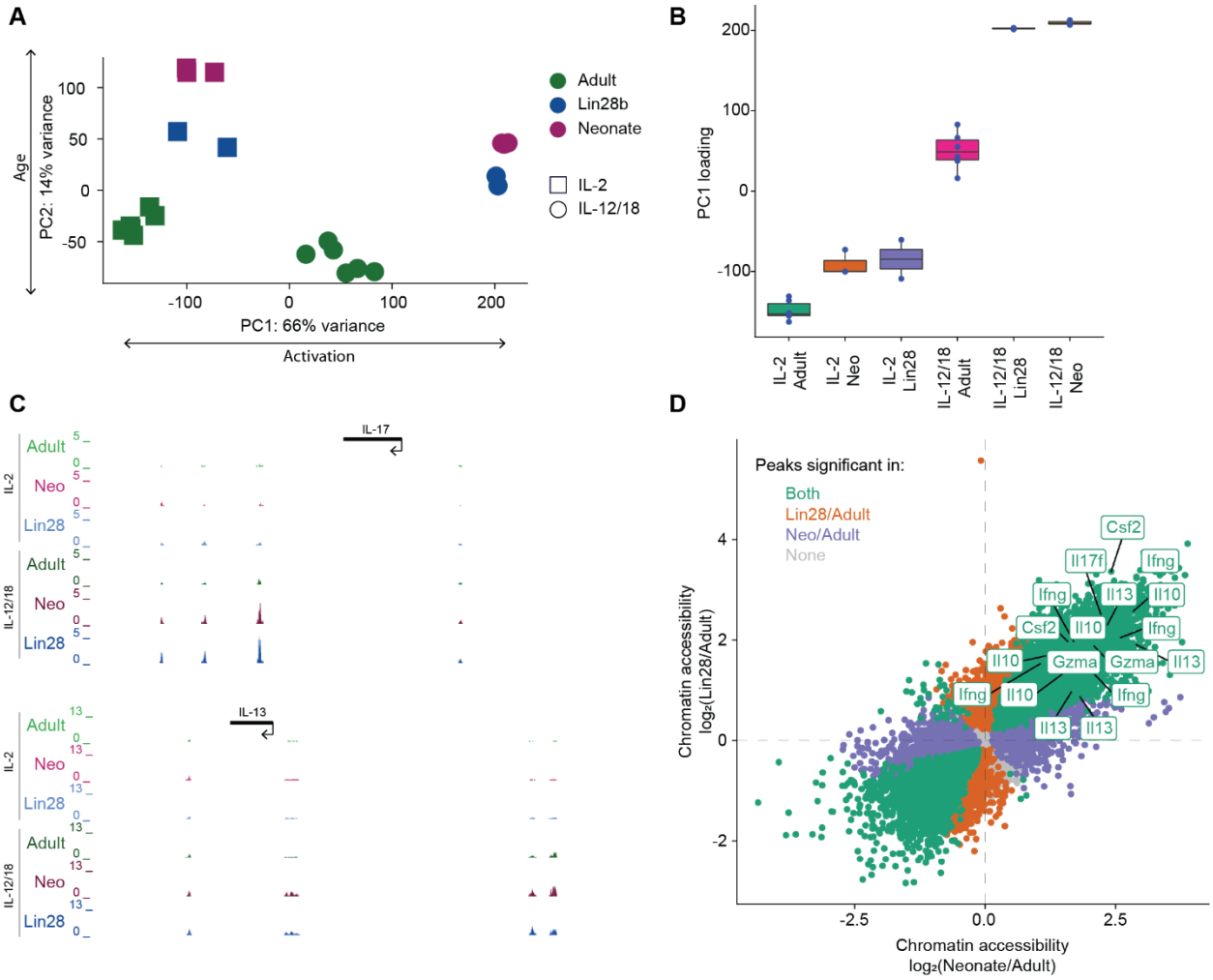

Fig 3.3. Response to inflammatory cytokines is accompanied by changes in chromatin landscape

(**A**) PCA analysis of chromatin accessibility (ATAC-seq), with PC1 indicating activation status and PC2 separating age. (**B**) Distributions of PC1 loadings for different samples. (**C**) Genome browser plot of chromatin accessibility at regulatort regions around IL-17f (top) and IL-13 (bottom) locus. The black box and arrow indicate the gene and its direction, respectively. (**D**) Comparison of change in chromatin accessibility at ATAC-seq peaks in neonatal and in Lin28 cells as compared to adult cells. The traditional (IFNg and GM-CSF) and innate cytokines (IL-10, IL-13, IL-17f and IL-22), and granzymes (Granzyme A and Granzyme B) are labelled.

and the second component separates the age (Fig 3.3A), recapitulating the patterns observed in transcriptome profiling. The neonatal and Lin28 cells exhibited a

chromatin landscape distinct from that of adults in control conditions, with even more pronounced difference after bystander activation. The adult cells displayed weaker response to the inflammatory cytokines, as evident by placement of stimulated adult cells in between control neonates and stimulated neonates (Fig 3.3A,B). In agreement with RNA-seq data, we observed increased accessibility in both neonatal and Lin28 cells at multiple chromatin regions associated with the innate and conventional cytokines and granzymes secreted by activated T-cells (Fig 3.3C,D). These results indicated that the neonatal response to inflammatory cytokines is controlled, in part, by regulating the chromatin landscape, which is eventually reflected in concordant changes in mRNA abundance.

Role of the BACH2/JUN axis in bystander activation

Because bystander activation triggers extensive changes in chromatin landscape and in mRNA abundance, we sought to identify specific transcription factors that underlie these changes. To determine the potential transcription factors that are responsible for the unique program of bystander activation in neonates, we first asked if the putative binding sites for any transcription factor are enriched in regions of open chromatin that are more accessible in neonatal cells compared to adults. We performed this analysis for about 700 different transcription factor binding motifs downloaded from CisBp (Weirauch *et al*, 2014). For neonate-specific regions, we

found enriched motifs for many transcription factors, which were highly similar to those found in Lin28-specific regions (Fig 3.4A). Based on the similarity of binding motifs, we grouped the significantly enriched motifs into three groups: 1) EOMES, 2) T-BET family of factors, and 3) BACH2/JUN-related factors. We have shown previously that the binding motifs for EOMES and T-BET, the factors responsible for effector function, are more accessible in naïve CD8+ T-cells of neonates compared to their adult counterparts (Smith *et al*, 2018). Consistent with that, we also observed enrichment for EOMES and T-BET in bystander activation of neonates, likely suggesting their role in bystander activation. However, putative binding sites of these factors were only marginally enriched. Interestingly, the BACH2/JUN-related factors, which include BACH2, MAFK, NFE2-family, FOS and JUN-family of transcription factors, exhibited strongest enrichment in neonate- and Lin28-specific chromatin regions. While the role of BACH2/JUN-related factors is well studied in TCR-mediated activation, to our knowledge, their role in bystander activation is unexplored.

Next, we hypothesized that if the BACH2/JUN-associated factors are indeed involved in regulating the bystander activation in neonates, we must observe their occupancy through transcription factor footprinting analysis of ATAC-seq data. The

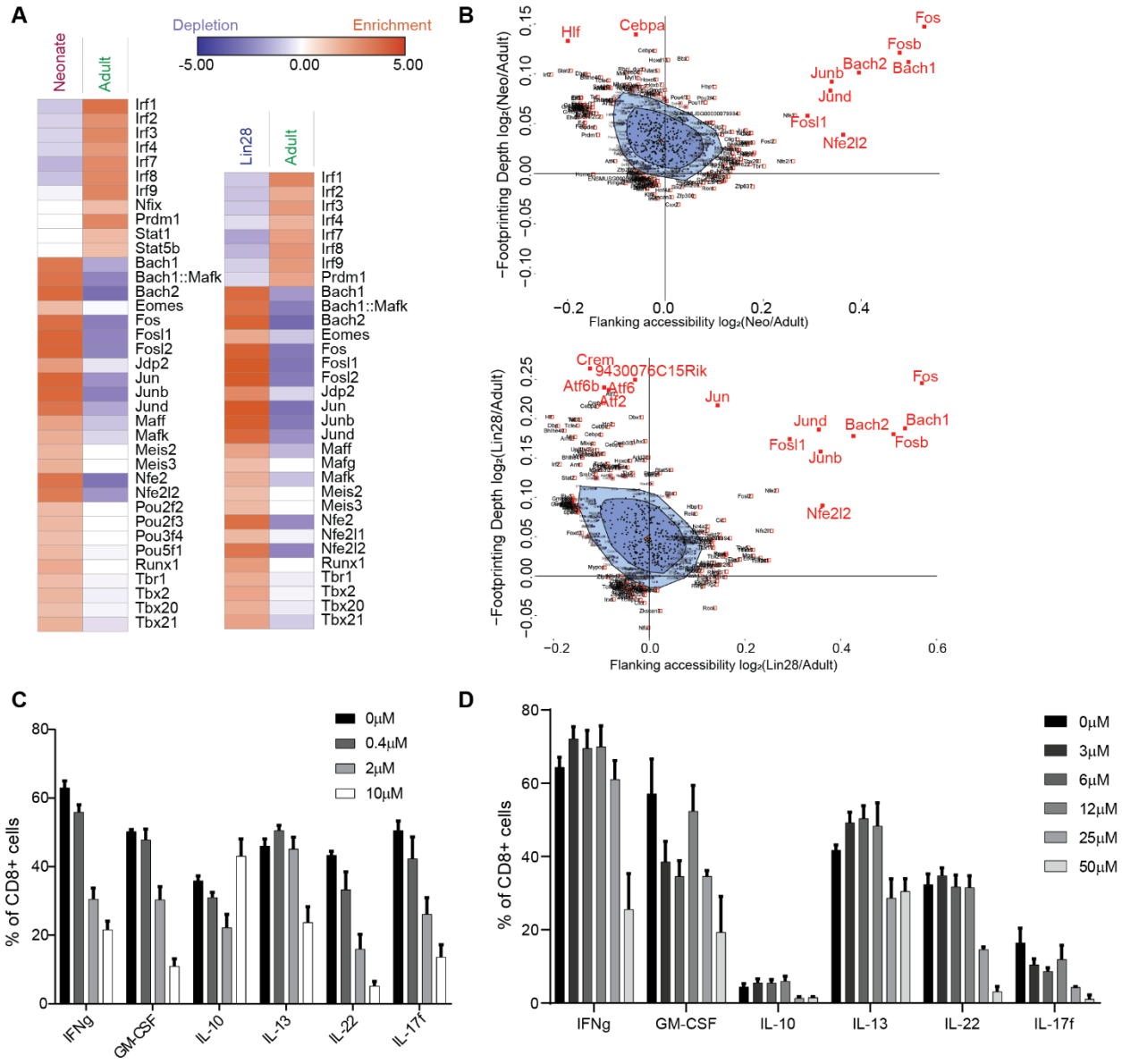

Fig 3.4. Role of BACH2/JUN axis in bystander activation

(**A**) Enrichment (red) or depletion (blue) of transcription factor binding motifs in differentially accessible chromatin regions between either neonate and adult (left) or between Lin28 and adult (right). The significance is color coded and was determined using a binomial test using randomly selected open chromatin regions as background, followed by FDR correction. (**B**) BaGFoot analysis depicting change in transcription factor footprinting and in flanking accessibility between neonate and adult (top) or between Lin28 and adult (bottom). The transcription factor with significance value less than 0.05 are labelled in red; the significance was calculated using Chi-square test. (**C**) Inhibition of AKT using different amounts of AKTi followed by expression measurement for traditional and innate cytokines and granzymes in neonatal cells

either treated with IL-2 (dashed line) alone or with the addition of IL-12 and IL-18 (continuous line). (**D**) Inhibition of JUN-family transcription factors using different amounts of AP-1i followed by expression analysis same as in (C).

rationale behind footprinting analysis is that the binding of transcription factor to DNA precludes transposase-mediated cuts in the binding motif, which can be observed as reduced accessibility at the binding motif in ATAC-seq signal, while the flanking region still maintain higher accessibility. We used an algorithm, Bivariate Genomic Footprinting (BaGFoot) (Baek *et al*, 2017) to determine the transcription factor footprints and flanking accessibility. Consistent with the analysis in Fig 3.4A, we observed strong flanking accessibility at the binding motifs of BACH2/JUN-related factors (Fig 3.4B – X-axis). Interestingly, we also found that the neonatal and Lin28 cells exhibited increased footprints for these transcription factors (Fig 3.4B – Y-axis), likely suggesting a direct role of BACH2/JUN in bystander activation. The BACH2/JUN-related factors share similar binding motifs (Roychoudhuri *et al*, 2016). Therefore, it is challenging to distinguish which transcription factor is occupying the observed footprints and hence is regulating the bystander activation. Nonetheless, we propose the following model for how BACH2/JUN-related transcription factors could be driving the unique program of bystander activation in neonates. Assuming similarities with TCR-mediated activation of CD8+ T-cells, we hypothesized that upon stimulation by inflammatory cytokines, the AKT kinase deactivates BACH2, and JNK kinase activates JUN factors, leading to the

transcriptional activation of the effector genes and perhaps the genes coding for innate cytokines via JUN transcription factors. Based on this model, the robust activation of neonatal and Lin28 cells can be explained if the neonatal and Lin28 cells had reduced expression of *Bach2* and increased expression of *Jun* transcription factors. Consistent with that idea, we observed reduced mRNA abundance of *Bach2* and increased mRNA abundance of *Jun* family transcripts in neonates and Lin28 compared to adults (Fig S3.2). These results indicate that the increased footprints of BACH2/JUN-related factors in neonates after IL-12 and IL-18 treatment likely derive from higher binding of JUN transcription factors at those loci, resulting into higher expression of effector genes, and hence, stronger response to inflammatory cytokines for neonatal and Lin28 cells.

To experimentally validate our model, we modulated the activity of AKT, the kinase that inactivates BACH2, in neonatal cells using a chemical inhibitor (AKTi), followed by bystander activation. Consistent with our model, the inhibition of AKT in neonates muted the bystander activation, as evident by loss of upregulation for traditional and innate cytokines (Fig 3.4C), likely via reduced inactivation of BACH2, preventing JUN-mediated activation of effector genes. Concordantly, when we inhibited the activity of JUN-family factors in neonates using a chemical inhibitor (AP-1i), the neonatal cells lost their ability to respond to inflammatory cytokines (Fig 3.4D). These results strongly indicate that the BACH2/JUN axis plays an important role in bystander activation. Further experiments are being performed to measure

nuclear and cytoplasmic localization of BACH2 during the bystander activation to obtain direct evidence for the role of BACH2 in bystander activation.

DISCUSSION

The central feature of the adaptive immune system is the diverse repertoire of cells in which each lymphocyte carries a unique receptor specific for a particular antigen; for T cells, this receptor is the TCR. During maturation, progenitor cells undergo rearrangement of TCR genes to produce a diverse pool of TCRs, each of which recognizes a unique antigen. While this process of TCR rearrangement is robust in bone marrow-derived T-cells, which reconstitute the CD8+ T-cell pool in adults, this process of rearrangement, however, is less efficient in fetal liver-derived T-cells, which give rise to CD8+ T-cells up to the age of 5-7 days (neonatal age) in mice. This reduced efficiency of TCR rearrangement results into a less diverse repertoire of T-cells, likely reducing the efficacy of the T-cell-mediated immunity in fetus and neonates. This begs a question of how does the T-cell arm of the immune system present in early life handle the rapidly changing environment upon birth. Although, in early life, other branches of the immune system, such as humoral immunity (e.g. maternal antibodies), aids in strengthening the immune system, it is not clear how the fetus and newborns compensate for the deficiency in T-cell-mediated immunity. Our work, for the first time, provides insights into this important question. We

demonstrated that the neonatal CD8+ T-cells exhibit the ability to respond aggressively to inflammatory cytokines in an antigen-independent manner, an unusual response for cells of the adaptive immune system. Perhaps, the adaptive-like and innate-like, the dual nature of the fetal-derived T-cells compensates for the reduced complexity of TCR repertoire in early life.

Our previous work has demonstrated that in the context of TCR-mediated stimulation, neonatal CD8+ T-cells respond more aggressively compared to adult cells although they exhibit reduced TCR complexity (Appendix), perhaps also aiding in strengthening the immune system in early life. Our previous (Appendix) and current data (this chapter) reveals that this age-related effect is not due to the age-dependent differences in the local environment in which the cells reside, because when we transferred neonatal cells into adult recipients, they still exhibited robust response to antigen (Appendix) and to inflammatory cytokines (Fig 3.1 & 3.2). These results suggested that fetal-derived cells are intrinsically programmed to behave in an unconventional way. We previously showed that microRNAs are differentially expressed between neonatal and adult cells, with their targets changing in opposite direction (Wissink *et al*, 2015), suggesting their potential role in choreographing the age-related differences. Indeed, genomic deletion of one of the adult-specific miRNAs, miR-29, changed the phenotype of adult CD8+ T-cells to that of the neonatal cells (unpublished work). Similarly, in the current work, we ectopically expressed an RNA-binding protein (Lin28b) in adults, whose well-known function is

to negatively regulate expression of a miRNA, let-7. As expected, we observed reduced levels of let-7 and upregulation of its targets in Lin28 cells (data not shown). Similar to miR-29 knock-out animals, the cells from adults that express Lin28b displayed phenotypes similar to those of neonatal cells. Together, these results indicate that let-7 and miR-29 are major regulators of the age-related differences in CD8+ T-cell response.

TCR-mediated stimulation of T-cells employs the BACH2/JUN axis to trigger effector response (Roychoudhuri *et al*, 2016). Since the stimulation of neonatal cells via inflammatory cytokines also results into a very similar phenotype as that via the TCR-mediated stimulation, it is highly likely that the cells would repurpose the BACH2/JUN axis for the bystander activation. Consistent with that hypothesis, our results showed that perturbing the BACH2/JUN axis abolishes the response of neonatal and Lin28 cells to inflammatory cytokines. Important outstanding questions relating to this model are as follows. First, how do IL-12 and IL-18 communicate with AKT and JNK, the modulators of BACH2 and JUN activity, respectively? Although TCR-mediated activation utilizes BACH2/JUN axis for T-cell activation, this does not result in expression of innate cytokines, such as IL-10, IL-13, IL-22. Second, how does the BACH2/JUN axis specifically regulate the expression of innate cytokines in response to bystander activation?

MATERIALS AND METHODS

Mice

Neonatal and adult gBT-I animals were described previously (Wang *et al*, 2016) and used at 5 to 7 days old or 2 to 4 months old, respectively. TCRa-/- recipients were purchased from Jackson Laboratories. Lin28 transgenic mice were a generous gift from Leonid A Pobezinsky (University of Massachusetts, Amherst, MA). For B6 bystander activation, adult B6 females and timed pregnant B6 females were ordered from Jackson Laboratories.

Flow cytometry analysis

All antibodies were purchased from Thermo Fisher Scientific, Biolegend or BD Biosciences. For intracellular staining the IC fixation and permeabilization kit from Thermo Fisher Scientific was used according to the manufacturer's instructions. Antibody labeling steps containing 2(+) brilliant fluorochromes were performed using Brilliant stain buffer (BD Biosciences). Data was collected using a FACS Symphony cytometer (BD Biosciences) and analyzed using Flowjo software (Treestar).

Fluorescence activated cell sortin (FACS)

CD8+ T cells were isolated from gBT-I adult, neonate, adult mice expressing Lin28b and were magnetically enriched by positive selection using anti-CD8a microbeads (Miltenyi) according to manufacturer's instructions. Populations were sorted to >95% purity with a FACS Aria III.

In vitro bystander activation

Cells were isolated from spleen for bystander activation. CD8+ T cells were isolated by positive magnetic separation using anti-CD8a microbeads (Miltenyi) according to the manufacturer's instructions. Cells were incubated in RPMI supplemented with 10% fetal bovine serum, L-glutamine, penicillin-streptomycin, 2-mercaptoethanol and IL-2 (10 ng/ml) alone, or IL-2 (10 ng/ml), IL-12 (1 ng/ml unless otherwise specified) and IL-18 (1 ng/ml unless otherwise specified) for 22 hours. Cells were labeled for FACS sorting or Brefeldin A was added to the cells for the final 4 hours prior to antibody labeling for flow cytometry analysis.

Inhibitors

Prior to overnight cytokine stimulation, enriched CD8+ T cells were pretreated for one hour at 37C with DMSO, AP-1i or Akti at indicated concentrations. Following pretreatment, cells were pelleted and resuspended in RP-10 containing IL-2 with or without IL-12 (1ng/ml) and IL-18 (1ng/ml) in the presence of DMSO, AP-1i or Akti.

Adoptive transfers

CD8+ T cells from gBT-I adult, neonate, Lin28 mice were enriched by magnetic selection. $3x10^6$ cells were transferred into TCRa-/- recipients. The next day recipients were infected with WT *Listeria monocytogenes*, WT Influenza or WT *N. brasiliensis*.

Infections and pathogen burdens

Wild type *Listeria monocytogenes* (*L. monocytogenes*) was grown to log phase, and mice were

infected intravenously (i.v.) with $1x10^4$ colony forming units (CFU) *L. monocytogenes* in PBS. At 3 dpi, livers were harvested for bacterial titers and were homogenized in 0.02% NP-40 in water using a GentleMACS dissociator (Miltenyi). Homogenates were serially diluted in water and plated on BHI plates to enumerate CFU per liver.

Wild type PR8 influenza was a generous gift from Jacco Boon (Washington University). Mice were injected intranasally (i.n.) with 100 TCID50 in PBS and weighed daily. At 7 dpi lungs were homogenized in DMEM supplemented with Penicillin-streptomycin using a GentleMACS dissociator (Miltenyi). Homogenates were stored at -80C prior to RNA extraction and qPCR.

For *Nippostrogylus brasiliensis* (*N. brasiliensis*) infections, mice were infected subcutaneously (subQ) with 500 worms in PBS. At 10 dpi small intestines were harvested in PBS and worms were counted using a dissecting microscope.

RNA preparation and sequencing

The total RNA was extracted from the CD8+ T-cells using Trizol extraction, followed by library preparation using NEBNext Ultra II RNA Library Prep Kit for Illumina (New England Biolabs), with initial polyA+ isolation, by the Transcriptional Regulation and Expression Facility at Cornell University. The libraries were sequenced on Illumina NextSeq500.

RNA-seq data analysis

The raw reads were trimmed to remove adapters using cutadapt (Martin, 2011) and were mapped to the mouse genome (mm10) using tophat (Trapnell *et al*, 2009). The

reads mapping to each gene were counted using featureCounts of the Subread package (Liao *et al*, 2014). The specific analysis as described in the text were performed using in-house Perl and R scripts. The differential expression analysis was performed using edgeR.

ATAC-seq library preparation

The FACS sorted cells were permeabilized and nuclei were isolated using a lysis buffer containing 0.1% Igepal and ATAC-seq libraries were prepared following the published protocols (Buenrostro *et al*, 2015; Corces *et al*, 2017) and sequenced on Illumina NextSeq500.

ATAC-seq data analysis and data visualization

The raw reads were trimmed using cutadapt and were aligned to the mouse genome (mm10) using bowtie2. The read alignments were filtered to remove PCR duplicates and low-quality alignments using PICARD and samtools. The peaks of ATAC-seq reads were called using macs2 and the reproducible peaks were obtained using Irreproducibility discovery rate (IDR) analysis. The peaks from all samples were merged using bedtools to generate a set of unified peaks. The accessibility for each peak was estimated by counting number of ATAC-seq reads that map to each peak using featureCounts. The read alignments that fall within the unified set of peaks were normalized by the total mapped reads and were converted to bigwig format for visualization on the genome browser. The differential accessibility analysis was performed using edgeR.

Motif enrichment analysis for transcription factors and BaGFoot analysis

The motif enrichment analysis was performed as described previously (Smith *et al*, 2018). Briefly, a set of open chromatin regions exhibiting increased accessibility in one sample compared to another were searched for the number of putative binding site for a given transcription factor. This number was then compared with a set of open chromatin regions selected randomly for generating the null distribution. The significance of the enrichment or depletion was calculated using binomial test. The BaGFoot analysis was performed according to the published code (Baek *et al*, 2017).

Geneset enrichment analysis

The GSEA analysis was performed using GSEA software (Subramanian *et al*, 2005). The gene sets for "innateness" and "adaptiveness" were downloaded from the supplementary material of (Gutierrez-Arcelus *et al*, 2019).

SUPPLEMENTARY FIGURES

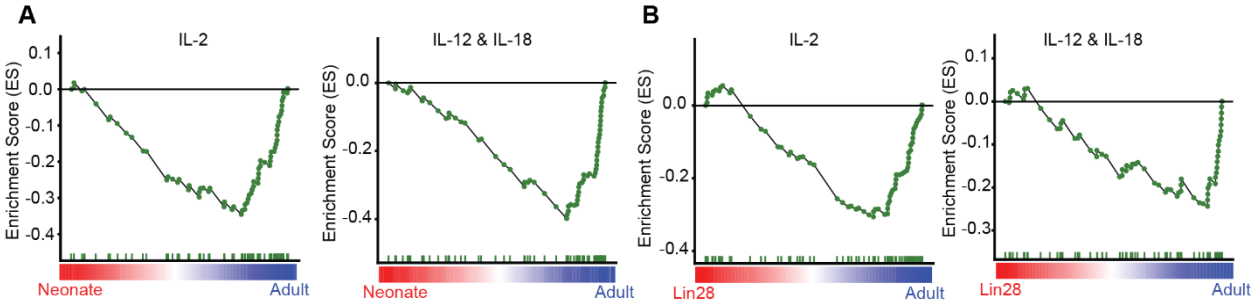

Fig S3.1. Enrichment of "adaptiveness"-genes.
GSEA enrichment score plots for "adaptiveness"-related genes when compared to neonatal (A) or Lin28 (B) cells treated with (right) or without (left) inflammatory cytokines.

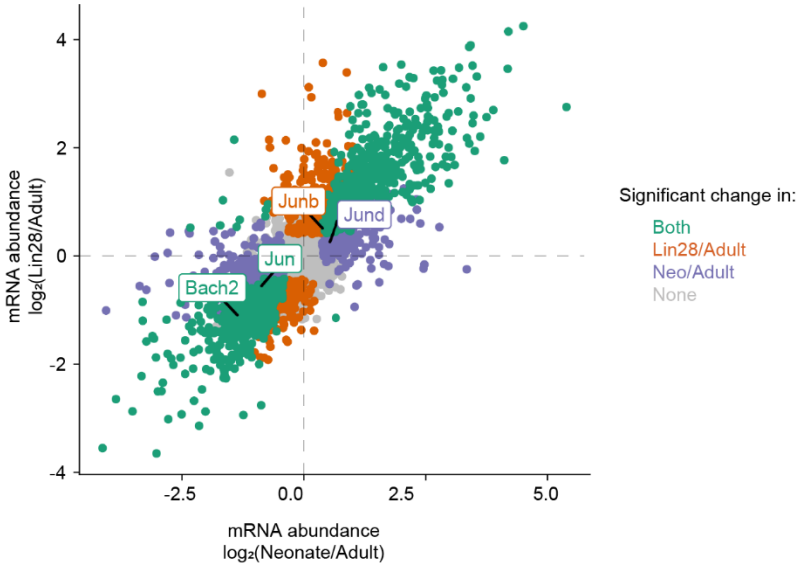

Fig S3.2. *Bach2* and *Jun* expression
Same as Fig 3.1D except that labeling BACH2 and JUN-family transcription factors.

REFERENCES

Baek S, Goldstein I & Hager GL (2017) Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *Cell Rep* **19:** 1710–1722

Buenrostro JD, Wu B, Chang HY & Greenleaf WJ (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109:** 21.29.1-21.29.9

Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, et al (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14:** 959–962

Cui W, Joshi NS, Jiang A & Kaech SM (2009) Effects of Signal 3 during CD8 T cell priming: Bystander production of IL-12 enhances effector T cell expansion but promotes terminal differentiation. *Vaccine* **27:** 2177–2187

Curtsinger JM, Schmidt CS, Mondino A, Lins DC, Kedl RM, Jenkins MK & Mescher MF (1999) Inflammatory cytokines provide a third signal for activation of naive CD4+ and CD8+ T cells. *J. Immunol.* **162:** 3256–3262

Dempsey PW, Vaidya SA & Cheng G (2003) The art of war: Innate and adaptive immune responses. *Cell. Mol. Life Sci.* **60:** 2604–2621

Falvo JV, Uglialoro AM, Brinkman BM, Merika M, Parekh BS, Tsai EY, King HC, Morielli AD, Peralta EG, Maniatis T, Thanos D & Goldfeld AE (2000) Stimulus-specific assembly of enhancer complexes on the tumor necrosis factor alpha gene promoter. *Mol. Cell. Biol.* **20:** 2239–2247

Gutierrez-Arcelus M, Teslovich N, Mola AR, Polidoro RB, Nathan A, Kim H, Hannes S, Slowikowski K, Watts GFM, Korsunsky I, Brenner MB, Raychaudhuri

S & Brennan PJ (2019) Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nat Commun* **10:** 687

Janeway C ed. (2001) Immunobiology: the immune system in health and disease ; [animated CD-ROM inside] 5. ed. New York, NY: Garland Publ. [u.a.]

Liao Y, Smyth GK & Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930

Macián F, López-Rodríguez C & Rao A (2001) Partners in transcription: NFAT and AP-1. *Oncogene* **20:** 2476–2489

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17:** 10

Oyake T, Itoh K, Motohashi H, Hayashi N, Hoshino H, Nishizawa M, Yamamoto M & Igarashi K (1996) Bach proteins belong to a novel family of BTB-basic leucine zipper transcription factors that interact with MafK and regulate transcription through the NF-E2 site. *Mol. Cell. Biol.* **16:** 6083–6095

Parish IA & Kaech SM (2009) Diversity in CD8(+) T cell differentiation. *Curr. Opin. Immunol.* **21:** 291–297

Rael EL & Lockey RF (2011) Interleukin-13 signaling and its role in asthma. *World Allergy Organ J* **4:** 54–64

Rincón M & Flavell RA (1994) AP-1 transcriptional activity requires both T-cell receptor-mediated and co-stimulatory signals in primary T lymphocytes. *EMBO J.* **13:** 4370–4381

Roychoudhuri R, Clever D, Li P, Wakabayashi Y, Quinn KM, Klebanoff CA, Ji Y, Sukumar M, Eil RL, Yu Z, Spolski R, Palmer DC, Pan JH, Patel SJ, Macallan DC, Fabozzi G, Shih H-Y, Kanno Y, Muto A, Zhu J, et al (2016) BACH2 regulates CD8(+) T cell differentiation by controlling access of AP-1 factors to enhancers. *Nat. Immunol.* **17:** 851–860

Smith NL, Patel RK, Reynaldi A, Grenier JK, Wang J, Watson NB, Nzingha K, Yee Mon KJ, Peng SA, Grimson A, Davenport MP & Rudd BD (2018) Developmental Origin Governs CD8+ T Cell Fate Decisions during Infection. *Cell* **174:** 117-130.e14

Smith NL, Wissink E, Wang J, Pinello JF, Davenport MP, Grimson A & Rudd BD (2014) Rapid proliferation and differentiation impairs the development of memory CD8+ T cells in early life. *J. Immunol.* **193:** 177–184

Smith NL, Wissink EM, Grimson A & Rudd BD (2015) miR-150 Regulates Differentiation and Cytolytic Effector Function in CD8+ T cells. *Sci Rep* **5:** 16399

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES & Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102:** 15545–15550

Trapnell C, Pachter L & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111

Wang J, Wissink EM, Watson NB, Smith NL, Grimson A & Rudd BD (2016) Fetal and adult progenitors give rise to unique populations of CD8+ T cells. *Blood* **128:** 3073–3082

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano J-C, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes

JS, et al (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158:** 1431–1443

Wissink EM, Smith NL, Spektor R, Rudd BD & Grimson A (2015) MicroRNAs and Their Targets Are Differentially Regulated in Adult and Neonatal Mouse CD8+ T Cells. *Genetics* **201:** 1017–1030

Zenewicz LA (2018) IL-22: There Is a Gap in Our Knowledge. *IH* **2:** 198–207

Chapter 4

Conclusions and Future directions

Conclusions

In my doctoral work, my primary goal has been to advance our understanding of miRNA-mediated regulation at the systems level. Towards this goal, I developed an experimental approach, named CARP, to robustly identify miRNA targets and applied this approach to disentangle regulatory networks of specific miRNAs.

Work from several labs over more than a decade have revealed important features of miRNA target sites, which has helped significantly in studying roles of miRNAs via identification of their targets. Despite these extensive efforts, I demonstrate that there is still a significant gap in our understanding of how miRNAs determine which mRNAs to target. The prediction algorithms, which are based on an incomplete understanding of miRNA targeting rules, are used extensively in the field. I showed that the false-positive rate of even the most effective algorithms exceeds 50%. One of the major caveats of previous approaches is the use of expression data from cells with transiently introduced, extremely high levels of miRNAs for training the prediction models; this caveat likely contributes to the high rate of false predictions. For my work, I developed a miRNA expression system that mimics the parameters of endogenous miRNAs, facilitating the identification of true miRNA targets. Using the cells that received my miRNA expression system, I showed that the number of true targets for miRNAs used in my studies is much smaller (a couple of hundred, depending on the miRNA) as opposed to the thousands of targets predicted using

algorithms. These results indicated that CARP dramatically improves the identification of miRNA targets.

It is known that miRNAs elicit small changes, yet the loss of specific miRNAs drives major phenotypic consequences. Such large effects on the phenotypes become possible because of miRNA's ability to mediate hundreds of small changes and often to repress transcripts coding for master regulators, such as transcription factors, which eventually results in wide-spread downstream changes. In terms of determining miRNA targets, the latter property of miRNAs poses a major challenge, because it is difficult to separate the direct targets of miRNAs from the downstream regulatory changes using the existing approaches. By combining RNA-seq and PRO-seq, I demonstrated that subtraction of transcriptional changes from the changes in mRNA abundance (CARP) provides robust estimates of the post-transcriptional regulation and that this approach effectively separates the direct targets from transcriptionally regulated indirect targets in response to a perturbed mRNA. I also observed substantial indirect transcriptional changes triggered by individual miRNAs, which, to my knowledge, has not previously been rigorously documented. Most of the miRNA studies in the field, including some from our lab, ignore this transcriptional indirect targeting by miRNAs, likely polluting their conclusions. Furthermore, using motif-enrichment analysis on PRO-seq-identified regulatory regions, I discovered candidate transcription factors responsible for the transcriptional indirect targeting, some of which correspond to the true direct targets of specific miRNAs. Additionally, I

demonstrated that CARP facilitates deconvolution of complex regulatory changes mediated by miRNAs, for example, using CARP, I discovered cases where mRNAs were transcriptionally upregulated in response to a specific miRNA and were also post-transcriptionally downregulated via direct targeting by the same miRNA, resulting in no net change in mRNA abundance. The studies relying only on the changes in mRNA abundance in response to a miRNA readily miss such, perhaps very important, targets. Moreover, our assumption was that majority of indirect targeting occurs at transcription, due to the major role of transcription in gene regulation. However, surprisingly, I found many post-transcriptional changes that were not direct targets, which indicated that these were likely indirect targets regulated at a post-transcriptional level. Possible explanations for such regulation is that miRNAs regulate transcripts coding for RNA-binding proteins via direct targeting, resulting in changes in post-transcriptional levels of the downstream targets of those RNA-binding proteins. Overall, my work demonstrated that CARP robustly distinguishes direct targets from downstream indirect changes and facilitates dissection of complex regulatory changes, thus aiding in a systems-level understanding of miRNA regulatory networks.

The majority of functional miRNA target sites are located in 3'UTRs; the activity of sites located in other regions of the mRNA, such as the 5'UTR and open reading frame (ORFs) have been controversial and are highly debated. Interestingly, I discovered discernible activity of ORF sites, which was restricted to specific miRNAs

and largely absent for other miRNAs. The activity of ORF sites, however, was weaker compared to 3'UTR sites. My data indicated that while ORF sites on their own may be insufficient to trigger substantial repression, they help weak 3'UTR sites to mediate significant repression of target mRNAs. These observations unearthed miRNA-dependent activity of ORF sites, which has not been observed before, in my knowledge.

Collectively, my doctoral work allowed disentangling of miRNA regulatory networks and produced important tools for advancing our understanding of miRNA-mediated regulation at the systems-level.

Future directions

Going forward, I propose two independent directions that demonstrate applications of CARP. First, extend the CARP approach to study biological functions of miRNAs in primary cells. Second, repeat my experiments for a larger cohort of miRNAs in the same cell line, as well as, in different cell lines, and perform comparative analysis to determine novel properties of functional target sites that are responsible for the deviations between the target prediction algorithms and true miRNA targets observed in cells, and to test if target repertoire changes with change in cellular system.

Studying the biological functions of miRNAs in primary cells

While I used a cell-line-based system to establish the efficacy of CARP, the power of CARP lies in investigating the biological roles of miRNAs in primary cells. It is believed that miRNAs and their targets are typically co-expressed, thus, modulating expression of specific miRNAs in the cells where they are usually expressed, and performing CARP in wild-type cells and cells with perturbed miRNAs would provide the most interesting insights about how specific miRNAs control gene regulatory networks and contribute to relevant phenotypes.

In collaboration with the Rudd lab, we are interested in investigating the biological roles of two miRNAs, let-7 and miR-29, in shaping the immune response of CD8+ T-cell to pathogens. As discussed in chapter 3, in Appendix I and in previous work from our labs (Wissink *et al*, 2015), let-7 and miR-29 play crucial roles in dictating the fate of CD8+ T-cells. Briefly, let-7 and miR-29 miRNAs are upregulated in adults compared to neonates, and adults exhibit expression profiles and chromatin landscape that are distinct from those of neonates (Chapter 3 and Appendix I). Loss of either let-7 via ectopic expression of Lin28b (repressor of let-7) or miR-29 by deletion of the locus in adults reprograms adult CD8+ T-cells to mimic neonatal T-cells in terms of their transcriptome profiles and chromatin landscapes, their naïve phenotype, and their response to TCR and bystander stimulations. To mediate such a profound impact on the transcriptome and the phenotype of CD8+ T-cells, these miRNAs must be controlling a complex gene regulatory network. In order to do that, the

miRNAs must be regulating all mRNAs and genes of the network either by directly targeting each of them (acting as a micromanager), by directly targeting only a handful of regulatory proteins, resulting into wide-spread indirect changes in the network (acting as a master regulator), or by combining large numbers of both direct and indirect targets in the network. To test these hypotheses and to understand how these miRNAs individually accomplish the challenging task of shaping the CD8+ T-cell biology, I propose to extend CARP analyses to the miRNA-loss-of-function mouse models. While the target prediction algorithms typically predict several hundreds of targets, I anticipate that the proposed experiments will provide a more manageable list of true miRNA direct targets, which will aid in building realistic hypotheses for further experiments and data analyses. Additionally, by combining CARP and motif-enrichment analyses, I anticipate that we would be able to determine the candidate transcription factors, such as EOMES, T-BET, BACH2, JUN, etc., and to study how those transcription factors are regulated by specific miRNAs, for example, via direct targeting or by indirect regulation of their transcription. Collectively, I foresee that we would be able to disentangle the regulatory networks of let-7 and miR-29 in CD8+ T cells. Gaining detailed understanding of these gene regulatory networks would have a profound impact on the human health, including success of vaccination programs. Although let-7 and miR-29 have distinct seed sequences, indicating different sets of targets for these miRNAs, their loss in adult CD8+ T-cells give rise to a common phenotype, similar to that of neonatal cells. Additionally, neonatal cells exhibit

differential expression of a few other miRNAs compared to adults, which could also be involved in dictating age-related differences. Although each of these miRNAs have distinct sets of targets, it is possible that these sets of targets overlap with each other, and that these miRNAs accomplish the common phenotype by regulating those small set of overlapping targets. In this manner, we can use CARP to determine exclusive and common targets of these miRNAs in order to better understand roles of these miRNAs in age-related differences in CD8+ T-cell response.

Studying the properties of functional miRNA sites

An overarching question in the miRNA field is to completely understand the rules of miRNA targeting. My data clearly show that our understanding of the properties of miRNA target sites is incomplete. Since my miRNA expression system closely mimics the features of endogenous miRNAs, and CARP robustly distinguishes direct targets from indirect regulatory changes, these tools provide a unique opportunity to discover novel features of the functional target sites. I propose to extend the CARP analysis to a larger set of miRNAs by expressing them in HEK293 using my miRNA expression system. I anticipate that by performing a comparative analysis of properties of CARP-identified direct targets across different miRNAs would result in discovery of novel properties of functional miRNA target sites. It would also be of great value to repeat these experiments in different cell lines, because it is highly likely that different cellular

contexts influence miRNA-mediated targeting. For example, since 3'UTRs contain many additional cis-regulatory elements other than miRNA target sites and are bound by multiple RNA binding proteins, regulating the expressing of the target transcripts, it is likely that different levels of these RNA binding proteins in different cellular contexts would affect the accessibility of miRNA target sites located in 3'UTRs. Additionally, the differential binding of RNA binding proteins across different cellular context would also affect the secondary structure of mRNAs, further influencing the target site accessibility. Using CARP analysis for a set of miRNAs across different cell lines would allow answering the important questions like, what is the role of cellular context in miRNA-mediated regulation and how prevalent is such complex regulation of mRNAs.

APPENDIX I

DEVELOPMENTAL ORIGIN GOVERNS CD8+ T CELL FATE DECISIONS DURING INFECTION[1]

ABSTRACT

Heterogeneity is a hallmark feature of the adaptive immune system in vertebrates. Following infection, naïve T cells differentiate into various subsets of effector and memory T cells, which help to eliminate pathogens and maintain long-term immunity. The current model suggests there is a single lineage of naïve T cells that give rise to different populations of effector and memory T cells, depending on the type and amounts of stimulation they encounter during infection. Here, we have discovered that multiple sub-populations of cells exist in the naïve CD8+ T cells pool, which are distinguished by their developmental origin, unique transcriptional profiles, distinct chromatin landscapes and different kinetics and phenotypes after microbial challenge. These data demonstrate that the naïve CD8+ T cell pool is not as homogenous as previously thought and offers a new framework for explaining the remarkable heterogeneity in the effector and memory T cell subsets that arise after infection.

[1]This appendix is adapted from a manuscript of the same name published in 2018 in *Cell*. This work was performed in collaboration with Norah Smith in the Rudd lab. The authors of this manuscript were Norah L. Smith, Ravi K. Patel, Arnold Reynaldi, Jennifer K. Grenier, Jocelyn Wang, Neva B. Watson, Kito Nzingha, Kristel Yee Mon, Seth A. Peng, Andrew Grimson, Miles P. Davenport and Brian D. Rudd. Ravi Patel prepared samples for RNA-seq and ATAC-seq, analyzed the sequencing data, performed bioinformatic analysis, interpreted data and prepared the manuscript and figures. Norah Smith performed most of the experiments including cell sorting experiments, cellular assays and mouse work, analyzed and interpreted data and prepared the manuscript and figures. Detailed description of author contributions can be found below.

INTRODUCTION

The fundamental unit of the adaptive immune response is the lymphocyte. In mice, there are approximately 25 million naïve CD8+ T cells per animal, of which around 80-1,200 cells are specific for a given microbial peptide (Jenkins et al., 2010). Following infection, the peptide-specific cells undergo massive clonal expansion and can produce up to 10 million effector CD8+ T cells in a week, which function to eradicate invading pathogens (Williams and Bevan, 2007). Once the pathogen has been eliminated, the majority (~90-95%) of effector CD8+ T cells undergo apoptosis, but a small percentage (~5-10%) survives and differentiates into long-lived memory cells, which protect the host against reinfection in the upcoming months and years (Kaech et al., 2002; Williams and Bevan, 2007). In the last decade, it has become clear that effector T cell responses are comprised of a wide range of short- and long-lived cell subsets, which in turn give rise to phenotypically and functionally diverse pools of memory cells (Gerlach et al., 2016; Joshi et al., 2007; Olson et al., 2013; Sarkar et al., 2008). Despite intensive research, the underlying source of phenotypic and functional diversity among effector and memory CD8+ T cells remains poorly understood.

The current model suggests that all naïve CD8+ T cells have an equal potential to mature into different effector subsets, but those receiving increased stimulation (in the form of antigen, costimulation or inflammatory cytokines) become biased towards the short-lived effector lineages, whereas those receiving less (but sufficient) stimulation give rise to the longer-lived memory subsets of cells. However, a number of recent studies have demonstrated that some naïve CD8+ T cells possess different cell-intrinsic properties prior to their response to infection, which influence their kinetics and phenotypes after infection (Reynaldi et al., 2016; Smith et al., 2014; Wissink et al., 2015). In particular, neonatal CD8+ T cells derived from fetal liver hematopoietic stem cells (HSCs) are inherently more proliferative and have an enhanced capacity to differentiate, when compared to CD8+ T cells produced from adult bone marrow HSCs later in life (Wang et al., 2016). Similar results have also been described for fetal-derived CD4+ T cells in mice (Adkins et al., 2003; Zens et al., 2017) and humans (Mold et al., 2010). However, there is still the important and unresolved question of whether neonatal-derived CD8+ T cells persist into adulthood and maintain their cell-intrinsic properties during infection.

In this report, we developed a novel strategy to 'timestamp' CD8+ T cells at various stages of development and examined their phenotype and behavior in adulthood. We found that fetal-derived CD8+ T cells preferentially become memory-like CD8+ T cells in adulthood and represent the early effectors during infection. In contrast, the adult-derived CD8+ T cells predominantly give rise to naïve CD8+ T cells with slower kinetics but exhibit an enhanced capacity to form less differentiated memory precursors. Thus, there appears to be a division of labor among CD8+ T cells with different developmental origins, which work in concert to protect the host against invading pathogens. Collectively, our studies demonstrate that the naïve CD8+ T cell pool is not as homogenous as previously thought and offers a new framework for explaining the heterogeneity in the effector and memory T cell subsets that arise after infection.

RESULTS

CD8+ T cells made early in life have an inherent propensity to develop into virtual memory cells

Tracking CD8+ T cells from early life into adulthood required the development of a 'fate-mapping' mouse model. We used a CD4 promoter-driven tamoxifen-inducible cre (CD4cre-ERT2) (Aghajani et al., 2012), which drives expression of the red fluorescent protein TdTomato (RFP) in CD8+ T cells undergoing thymic selection at the CD4+CD8+ double positive (DP) stage of T cell development. This strategy allowed us to permanently 'timestamp' (ts) a wave of CD8+ T cells made in the thymus during tamoxifen exposure (Figures 1A and S1A,B). Our approach was to timestamp CD8+ T cells produced in mice at the following stages: (i) fetal (1d ts); (ii) neonatal (7d ts); and (iii) adulthood (28d ts). Most double-positive (DP) thymocytes at birth are derived from fetal HSCs (Adkins, 1991), whereas the majority of those at day 28 come from adult HSCs (Kim et al., 2007). The DP thymocytes in neonates (7d ts) likely represent a mixed population of cells derived from both fetal and adult HSCs.

We first asked whether CD8+ T cells from different developmental origins exhibit unique phenotypes in the periphery of adult mice. Previous studies (Akue et al., 2012; Wang et al., 2016), indicated that uninfected neonatal mice contain a large fraction of "virtual memory" (VM) cells, which are naïve CD8+ T cells that acquire phenotypic markers (CD44 and CD122) and functional properties (rapid response to stimulation) akin to memory CD8+ T cells. However, VM cells are produced by homeostatic mechanisms and arise in the absence of foreign antigen. When we examined timestamped cells in adults, we found over 5 times more $CD44^{hi}CD122^{hi}$ cells in 1d ts CD8+ T cells compared to those from 28d ts (50% vs 9%) (Figure 1B). It is possible that some $CD44^{hi}CD122^{hi}$ cells are actually "true memory" cells and the phenotypic differences among cells with disparate origins is due to differences in exposure to foreign antigen. However, multiple lines of evidence argued against this possibility. First, most 1d ts cells do not express CD49d (Figure S1C), which is exclusively expressed on true memory cells (Haluszczak et al., 2009). Second, 1d and 7d ts cells express significantly more Ly6C, CXCR3, and T-bet than their adult counterparts (Figure S1D) and expression of these markers is consistent with their VM phenotype (Figure S1E) (White et al., 2017). Third, 1d ts CD8+ cells were enriched in the liver (Figure S1F), which was recently shown to be a reservoir for virtual memory cells (White et al., 2016). Finally, in adults, the majority of TCR transgenic (gBT-I) 1d ts CD8+ T cells exhibited VM phenotype without exposure to cognate antigen (Figure S1G). Collectively, these data demonstrate that the phenotype and localization of CD8+ T cells in adult animals is linked to the stage of life at which they were originally produced.

Next, we investigated why the 1d ts CD8+ T cell pool contains more VM cells than 28d ts CD8+ T cell pool in adult mice. One possibility is that the 1d ts cells are much older than 28d ts cells and have therefore had more time to undergo homeostatic proliferation and convert to VM CD8+ T cells in the periphery. To test this, we compared 1d and 28d ts CD8+ T cells at the same time post-marking. 1d ts CD8+ cells acquired a VM phenotype more rapidly than 28-day marked cells (61.7% vs 13.4%), when compared at the same time point after labeling (4 wks). Moreover, differences are maintained for 3 months (Figure 1C). These results indicate that it is not simply the post-thymic age of the cell that dictates

their propensity to become a VM cell, but rather the age of the animal when the cells were produced.

We also considered that the peripheral environment is different in neonatal and adult animals. Indeed, neonatal mice are largely devoid of T cells, which may support lymphopenia-induced proliferation (Min et al., 2003) and explain why 1d ts cells more rapidly become VM cells. To examine this possibility, we transplanted a newborn timestamp thymus into an adult timestamp mouse and marked waves of newborn (RFP) and adult (YFP) thymocytes that we could track in the same peripheral environment (Figure 1D). Newborn-derived CD8+ T cells underwent more proliferation (Figure S1H) and preferentially developed into VM cells, whereas the majority of adult-derived CD8+ T cells do not (Figure 1E,F). Indeed, four weeks after labeling, cells from newborn thymii that matured in adult hosts had a similar proportion of VM cells to that seen in their natural environment (58 vs. 62% respectively), suggesting lymphopenia-induced proliferation in the neonatal environment was not required. Together, these data indicate CD8+ cells of early developmental origin have a cell-intrinsic propensity to become VM cells and that this fate bias occurs independently of environmental factors experienced in the periphery.

CD8+ T cells made during different developmental windows exhibit distinct gene regulatory programs

To examine whether CD8+ T cells derived from different developmental layers display global differences in gene expression that persist in adulthood, we performed RNA-seq on 1d and 28d ts CD8+ T cells from uninfected adults. Principal component analysis revealed that 1d ts and 28d ts cells have distinct transcriptomes (Figure 2A). Overall, we found 222 significantly upregulated genes and 73 genes that were downregulated in 1d ts cells (Figure 2B, Table S1). To understand differences in gene expression, we investigated enrichment (Subramanian et al., 2005) of the transcripts that are preferentially found in 1d or 28d ts cells relative to lists of genes that are known to typify CD8+ T cells at various stages of their responses, as defined by the ImmGen consortium (Best et al., 2013). Interestingly, we found that 1d ts cells express a significantly higher proportion of genes typically found in short-term effector, late effector and memory cells (clusters VI, VIII, IX, X) (Figures 2B,C and S2, Table S2). Cluster X genes, which typically show the highest level of expression at the peak of the response and after the pathogen has been cleared, exhibited the most enrichment in genes upregulated in naïve 1d ts cells (Figure 2C). In contrast, 28d ts cells expressed more genes that characterize naïve and late memory cells (cluster IV) (Figure S2 and Table S2). Among the genes upregulated in 1d ts cells, we observed higher levels of genes encoding phenotypic markers (*Ly6c, Cx3cr1*) (Figure 2D), effector functions (*Gzma, Ifng, Prf1*) (Figure 2E) and transcription factors (*Tbx21, Zeb2, Prdm1*) characteristic of effector CD8+ T cells (Figure 2F). CD8+ T cells made early in life also expressed elevated transcripts typically found in NK cells (*klrk1, klrb1, klra7*) (Figure 2G), raising the possibility that they possess innate-like functions. These data suggest that, prior to infection, 1d ts represent a unique and more effector-like population of CD8+ T cells in the periphery of uninfected adult mice.

CD8+ T cells with early developmental origins respond more rapidly to antigen and

cytokine signals than their adult counterparts

The large differences in phenotype and gene expression profiles in CD8+ T cells with different origins prompted us to examine whether such cells behave differently upon activation. Our previous work demonstrated neonatal CD8+ T cells undergo more cell division than adults after *in vitro* stimulation (Reynaldi et al., 2016; Smith et al., 2014; Wang et al., 2016), We performed similar assays with timestamped cells and found nearly 3 times as many 1d ts cells had undergone cell division compared to 28d ts cells (Figure 3A). To assess the underlying dynamics, we used a previously developed mathematical modeling approach (Gett and Hodgkin, 2000; Hawkins et al., 2007; Reynaldi et al., 2016) and found that 1d ts cells entered division earlier and divided faster (Figure 3B). We also determined that death rate prior to first division is not significantly different (Figure 3C). However, on each division cycle, we found 1d ts cells have better survival (Figure 3D). 1d ts cells lose more CD62L per division, indicating they more rapidly differentiate (Figure 3E). These analyses are summarized in Table S3.

We also asked if fetal-derived cells showed differences in activation markers or effector molecules. We found that prior to their first round of division 1d and 7d ts cells have increased production of the effector proteins granzyme B and IFNγ (Figure 3F,G) and elevated amounts of granzyme B are maintained after proliferation starts (Figure 3H). To assess whether stamped cells produced at different times display any innate immune functions, we stimulated with pro-inflammatory cytokines (IL-12 and IL-18). After an 18-hour incubation period, we observed rapid IFNγ production by 1d and 7d ts cells that expressed CD44 (Figure 3I). Collectively, these findings demonstrate that neonatal cells persisting into adulthood retain their capacity to rapidly respond to stimulation and are capable of participating in both innate and adaptive immune responses.

CD8+ T cells made during early stages of development are the first to respond to infection in adults

We next asked if developmental origin of CD8+ T cells alters their behavior during infection *in vivo*. Factors that potentially influence the fate of developmentally disparate cells include a less diverse repertoire (Rudd et al., 2011) and smaller number of precursors (Nelson et al., 2015), which could potentially alter immune function (Badovinac et al., 2007; Messaoudi et al., 2002). To directly evaluate cell intrinsic differences and control for these variables, we adoptively co-transferred cells from gBT-I TCR transgenic timestamp and Thy1.1 mice into congenically marked recipients. Consistent with our *in vitro* stimulation results, we found nearly three times more 1d ts cells at 5dpi (Figure 4B). However, this early burst is not sustained and 28d ts cells were found to make up a larger portion of the transferred pool of cells at later time points (Figure 4B). More rapid proliferation of 1d ts cells was accompanied by a notable difference in their phenotype. Whereas 1d ts cells were preferentially KLRG1^{hi} (a marker of terminal differentiation) their 28d ts counterparts had a more balanced population of KLRG1^{hi} and KLRG1^{lo} cells (Figure 4C). These data indicate that, even after controlling for developmental differences in TCR repertoire and precursor frequency, cells produced early in life proliferate and differentiate more quickly than those produced later in life.

We next asked whether such differences were evident in 1d, 7d and 28d ts non-TCR

transgenic timestamp mice (Figure 4D). To determine whether cells made early in life proliferate more early in infection, we compared the percentage of ts cells during infection to that observed prior to infection in each animal. Indeed, we found a larger increase in the percentage of 1d and 7d ts CD8+ T cells at 5 days post-infection (Figure 4E), suggesting polyclonal cells from early life are more proliferative. We confirmed that 1d ts cells have a greater proportion of terminally differentiated short-lived effector cells and make more IFNγ during early phases of the infection (Figure 4F,G). Collectively, these data indicate that the developmental origin plays a deterministic role in the fate of CD8+ T cells during infection.

We also compared the impact of inflammation on the response of CD8+ T cells with different developmental origins. Using a dendritic cell (DC)-vaccination approach in the presence or absence of IL-12 stimulation, we could determine whether differences in IL-12 sensitivity alone is sufficient to recapitulate our infection results. We found that DC-immunized 1d ts mice, receiving IL-12 (i.p.) generated a higher percentage of KLRG1^{hi} cells than mice marked in adulthood (Figure S3). These findings raise the possibility that 1d ts cells undergo more robust effector cell differentiation because of enhanced reactivity to innate cytokines.

CD8+ T cells produced in early life are pre-programmed to mount a rapid effector response

CD8+ T cells with early developmental origins exhibit distinct phenotypes and display more rapid effector responses later in adulthood. However, it remained unclear whether these differences were intrinsic or due to differences in amount of post-thymic maturation or initial phenotype prior to infection (1d ts cells having a substaintially higher proportion of VM cells). We therefore designed an experiment to disentangle the developmental origin from both the time spent in the periphery before analysis, as well as the phenotype of cells at the time of infection. Four distinct gBT-I donor populations (1d ts bulk, 1d ts VM, 28d ts bulk, 28 ts VM), all 4 weeks post-mark, were adoptively transferred into separate recipients and responses to LM-gB were compared (Figure 5A). Regardless of initial phenotype, 1d ts cells preferentially differentiated into KLRG1^{hi} terminal effectors, whereas 28d ts cells gave rise significantly fewer KLRG1^{hi} terminal effector cells (Figure 5B). RNAseq from 1d and 28d ts VM cells at 5 dpi showed that 58 genes were differentially expressed between 1d and 28d ts VM cells (Figures 5C and S4, Table S1). Enrichment analysis using the ImmGen clusters revealed 1d ts cells upregulated genes typically observed in short-term and late-effector CD8+ T cells (clusters VI, VIII, IX, X) (Figure S4, Table S2). Together, these data indicate that 1d and VM cells give rise to more terminally differentiated effectors than 28d ts VM cells.

We also compared the phenotype of 1d and 28d ts VM cells that transition into the resting memory pool at later stages of infection. Given that the differentiation state of effector cells predicts the type of memory cell that can be formed (Gerlach et al., 2016; Mackay et al., 2013), we hypothesized that cells made early in life give rise to more activated and effector-like populations of memory cells, which express higher levels of KLRG1 and lower levels of CD27 and CD43 (Hikono et al., 2007; Olson et al., 2013). Indeed, at 41 dpi, the 1d ts VM population has approximately twice as many KLRG1^{hi}CD62L^{lo} (Figure 5D) and CD27^{lo}CD43^{lo} cells (Figure 5E). Collectively, our results suggest that more rapid effector

cell differentiation in 1d ts cells is not due to age-related differences in post-thymic maturation or phenotype, but rather that cells with disparate developmental origins adopt different effector and memory fates after infection because they are programmed differently.

CD8+ T cells made early in life have an effector-like chromatin landscape

Recent studies have mapped chromatin accessibility in naïve, effector and memory CD8+ T cells and showed that significant changes in the epigenetic landscape are required to progress towards differentiation pathways (Gray et al., 2017; Yu et al., 2017). However, comparisons of the regulatory landscape among different subsets of naïve CD8+ T cells in the same host have yet to be performed. Based on our cumulative data, we hypothesized that CD8+ T cells with different developmental origins adopt distinct and predictable fates during infection because they have different chromatin landscapes prior to infection. To test our hypothesis, we mapped regions of open chromatin in 1d ts VM, 28d ts VM cells from gBT-I mice using ATACseq. Additionally, we included CD44lo true naïve (TN) cells from the 28d ts group as an additional reference population. Principal component analysis of ATACseq profiles clearly segregated the 1d VM subset from the 28d VM and 28d TN groups, suggesting cells produced in early life do indeed exhibit a unique chromatin landscape (Figure 6A). To systematically compare open chromatin regions in CD8+ T cells of different developmental origin, we clustered all 46,140 ATACseq peaks into six distinct groups with concordant behavior (Figure 6B, Table S4). Each group contained peaks associated with genes involved in effector and memory cell differentiation (Figure 6B). However, 1d ts VM cells showed increased accessibility to genes (group 4) that favor effector cell differentiation (Tbx21, Id2, IL2ra, IL15ra) and decreased accessibility to genes (groups 1 and 6) promoting naïve and memory cell development (foxo1, foxo3, IL6st, TGFbr) (Figure 6B, Table S4).

To obtain an unbiased perspective, we next compared genes associated with peaks in each of the six groups with gene-clusters described by the ImmGen consortium (Figure 6C). Our enrichment analysis revealed that effector genes are enriched in group 4 and depleted in group 1, indicating higher accessibility of these genes in 1d ts VM cells than the 28d ts VM cells. We confirmed these patterns using independent annotated gene sets that differentiate naïve, short-lived effector (SLEC), memory-precursor effector (MPEC), and memory CD8+ T-cells (Joshi et al., 2007; Luckey et al., 2006; Subramanian et al., 2005) (Figure S5). Consistent with these trends, ATACseq peaks corresponding to genes encoding certain effector functions (*GzmA, IFNg, GzmM*) showed strong signatures of open chromatin in the 1d VM cells compared to either 28d ts VM or 28d ts TN cells. In contrast, enhancers close to the gene for *IL2*, a key cytokine known to promote the formation of memory cells (Williams et al., 2006), were more accessible in 28d ts VM cells (Figure 6D). Collectively, these data suggest that naïve CD8+ T cells with different developmental origins exhibit unique chromatin architectures and that cells made early in life are poised to become effectors prior to infection.

CD8+ T cells with early developmental origins exhibit effector-like transcription factor signatures

Transcription factors (TFs) regulate differentiation of CD8+ T cells (Kaech and Cui,

2012; White et al., 2017); they do so by binding regulatory motifs in open enhancers and promoters (He et al., 2016; Scharer et al., 2017; Yu et al., 2017). To identify TFs that determine the fates of CD8+ T cells produced at different ages, we calculated the enrichment of TF binding motifs in loci that exhibited differential chromatin accessibility among 1d ts VM, 28d ts VM, or 28d ts TN cells. Regions with increased accessibility in 1d ts VM cells compared to 28d ts VM cells contained enriched motifs for a number of TFs (Tbx21, Eomes and Runx1) associated with driving effector cell differentiation (Figure 7A) (Best et al., 2013; Kaech and Cui, 2012; Zhang et al., 2008a). Interestingly, motifs for Tbx21 and Eomes were also associated with open chromatin in 28d ts VM cells compared to 28d ts TN cells, indicating their ability to promote differentiation of naïve CD8+ T cells depends on both the developmental origin and the initial phenotypic status of the cell (Figure 7B). In contrast, TFs with roles in repressing effector cell differentiation (Sp1, Egr2, NF-kB) (Miao et al., 2017; Moskowitz et al., 2017; Teixeiro et al., 2009) showed enriched binding in chromatin more accessible in 28d ts VM cells and depleted binding in chromatin more open in 1d ts VM cells (Figure 7A).

To confirm differential accessibility of regulatory regions, we reexamined the ATACseq data specifically for the TFs highlighted above, this time examining the complete spectrum of ATACseq peaks corresponding to each TF, rather than the minority that differ (Figure 7C). Consistent with our initial analysis, chromatin signatures for Eomes and Tbx21 are graded from most open in 1d ts VM cells to least open in 28d ts TN. Runx1 binding motifs are preferentially open only in 1d ts VM cells. In contrast, Sp1, Rela and Egr2 all exhibited signatures of reduced activity in specifically 1d VM cells. Additional examples are included in Figure S6. Collectively, these analyses indicate CD8+ T cells produced during different stages of life have distinct regulomes that influence their responsiveness to key TFs.

Lastly, we prepared RNA-seq libraries from the same samples used to generate our ATACseq libraries to directly compare the transcriptomes and chromatin landscapes of CD8+ T cells with disparate developmental origins. The transcriptomes were congruent with ATACseq data and again confirmed that differential gene expression profiles are linked to developmental origin (Figure S7). However, we identified 569 and 970 poised genes in 1d ts VM and 28d ts VM, respectively, with pronounced differences in ATACseq signals that were not differentially expressed concordantly in the transcriptome. Gene ontology (GO) enrichment analysis indicated genes poised in 1d ts samples specifically overlap with genes characteristic of cell death and secretion, while poised genes in 28d samples, on the other hand, exhibited enrichment of more general GO terms, although one notably enriched term was 'negative regulation of cell death' (Figure 7D,E; Table S5). This data suggests that cells with different developmental origins adopt distinct fates during infection because they contain different regulatory circuitry prior to stimulation.

DISCUSSION

During early development of the immune system, the periphery is colonized by waves of hematopoietic stem cells differing in their capacity to proliferate and undergo self-renewal (Mold and McCune, 2011). As a result, unique populations of lymphocytes (e.g., B1a-B cells and DETCs) are created and populate the host in a sequential manner, resulting in a stratified immune system based on developmental layers (Hardy and Hayakawa, 1991; Ikuta et al., 1990; Kantor et al., 1992). In this report, we used a novel approach to fate map CD8+ T cells produced during different stages of life, allowing the identification of developmental layers in the CD8+ T cell compartment that play unique roles during infection in adulthood. Our data indicate the host response to intracellullar pathogens in adulthood is linked to how the CD8+ T cell compartment is constructed during immune ontogeny.

Perhaps one of the most striking findings from our study is that the fate of effector CD8+ T cells is 'pre-programmed', at least in part, by the developmental stage at which precursor cells undergo maturation in the thymus. Previous work showed individual CD8+ T cells from TCR transgenic OT-1 mice can differentiate into all effector subsets typically observed in endogenous recipient populations (Stemberger et al., 2007). As a consequence, the prevailing notion in the field is that all naïve CD8+ T cells have the same potential to mature into memory cells and that phenotypic diversification primarily occurs during priming. However, these studies also show that individual cells adopt quite different fates, producing very different proportions of effector and memory subsets (Plumlee et al., 2013). Importantly, earlier studies were performed with naïve CD8+ T cells from adults. In light of our data, it seems likely that a substantial portion of variation in the fates and 'burst sizes' that have been observed in individual cells during infection stems from recruiting cells produced during different stages of life. These findings highlight the importance of closely examining the origins of cells present in the starting population, rather than focusing solely on the phenotype and function of cells that are participating in the response.

Our results also provide a different perspective on how CD8+ T cells mediate immune protection against intracellular pathogens. Currently, the field focuses on how different stimuli drives individual precursor pools of naïve CD8+ T cells to differentiate into various subsets of effector and memory cells. However, 'timestamping' offers a view through a developmental lens, suggesting that different developmental layers in the CD8+ T cell compartment react differently to stimuli and work together to control intracellular pathogens. Cells made during infancy exhibit innate-like functions in adulthood and rapidly deploy effector molecules in response to pro-inflammatory cytokines. In contrast, the most adult-derived CD8+ T cells respond only to their cognate antigen but are far more efficient at transitioning into the long-lived memory pool after infection than early origin counterparts. This holistic view of the CD8+ T cell response to infection is reminiscent of B cell subpopulations (B1a, B1b, B2 cells)(Herzenberg and Herzenberg, 1989) that co-exist in adult mice and suggests individual CD8+ T cells contribute specific functional roles to the immune system during development. In the future, it will be important to determine whether the more peptide-promiscuous fetal-derived CD8+ T cells (Gavin and Bevan, 1995) exhibit different antigen recognition patterns in adulthood, as has been described for B1-B cells (Berland and Wortis, 2002).

Developmental layers present during immune ontogeny are consistent with

progressive evolution of the immune system in the animal kingdom (Herzenberg and Herzenberg, 1989). In mammals, slower adaptive immune cells are layered on top of more primitive fast effector cells that arise in fetal life. Presumably, the primitive lineages were retained during evolution to provide a measure of innate immune protection until a fully mature adaptive immune system is established, although they clearly serve specialized functions later in life as well. How do these primitive lineages of neonatal-derived immune cells retain their cell-intrinsic properties in adulthood? Our data suggest that neonatal and adult-derived cells possess markedly distinct chromatin landscapes and gene expression programs. Naïve cells produced during early stages of development contain a regulome more typically seen in effector cells, suggesting they are poised to respond prior to stimulation. This data is similar to observations in ILCs and CD4+ T cells, where the ILCs contain a more 'pre-primed' regulome and the CD4+ T cells required additional chromatin remodeling upon activation (Koues et al., 2016; Shih et al., 2016). In the case of CD8+ T cells, it seems likely that the fast and innate T cell response represents the basic foundation of CD8+ T cell response, which was later elaborated on during evolution to generate CD8+ T cells exhibiting more flexibility and specificity to different stimuli and develop into memory cells.

It is also interesting to consider how the layering of CD8+ T cells may vary in the human population. Indeed, our data demonstrating that the heterogeneity in the effector and naïve pools are linked raises the possibility that we may one day use measures of developmental layering to better predict outcomes to vaccination, infections and cancer immunotherapy. It will be important to determine how the developmental architecture in the CD8+ T cell compartment is shaped by intrinsic and environmental factors with progressing age. For example, previous work indicated that the elderly respond poorly to vaccines because their CD8+ T cells become senescent (Miller, 1996). Our data offers an alternative explanation: that aged individuals are no longer able to respond vigorously to infection because they have lost most of the fetal and neonatal layers of the CD8+ T cell compartment. Similarly, early thymectomies result in a CD8+ T cell compartment lacking the adult layer. Whether these individuals exhibit an imbalance in the generation of effector and memory CD8+ T cells remains an open question.

In conclusion, our work suggests that 'early effectors' in the primary CD8+ T cell response to infection represent cells with early developmental origins and altered chromatin landscapes. Although our work is confined to CD8+ T cells, it seems likely the same phenomenon may be seen in other subsets of T cells, revealing a general paradigm of how immune ontogeny contributes to diversity. In particular, there is extensive work demonstrating neonatal CD4+ T cells are inherently biased towards Th2 differentiation (Adkins et al., 2004; Zaghouani et al., 2009), and that fetal naïve CD4+ T cells are predisposed to Treg differentiation after stimulation (Mold et al., 2010). Thus, it will be of particular interest to determine whether the outgrowth of CD4+ T cell subsets (Th1, Th2, Tregs) in adults is linked to when the naïve precursors are initially made. Knowledge gained from all of these studies is expected to broaden our fundamental understanding of immune ontogeny and the differentiation of effector and memory T cells after infection.

STAR Methods

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Brian D. Rudd (bdr54@cornell.edu).

Experimental Models and Subject Details

Mice

B6, Thy1.1, Ai9, and eYFP mice were purchased from Jackson Laboratory. The CD4cre-ERT2 were obtained from Fotini Gounari (The University of Chicago) (Aghajani et al., 2012) and are now commerically available at Jackson Laboratory. C57Bl/6 and Ly5.2 mice were purchased from the National Cancer Institute. gBT-I mice TCR transgenic mice (TCRαβ specific for the HSV-1 glycoprotein $B_{498-505}$ peptide SSIEFARL) were provided by Dr. Janko Nikolich-Zugich (University of Arizona, AZ)(Mueller et al., 2002). Large experimental cohorts were generated by setting up timed pregnancies and dividing litters into experimental groups (> 2 litters in all cases). At the time of experimentation, mice were 8-9 weeks of age (unless otherwise noted), sex-matched within the experiment and were maintained under pathogen-free conditions at Cornell University College of Veterinary Medicine, accredited by the American Association of Accreditation of Laboratory Animal Care. No association of response with sex was observed or expected. The Institutional Animal Care and Use Committee at Cornell University reviewed and approved all protocols for animal usage.

Cell culture

B16-Flt3 tumor cell line was provided by Dr. Stephen Jameson (University of Minnesota, MN) (Fulton et al., 2015; Shih et al., 2016)). Cells were cultured in RP-10 (RPMI supplemented with 10% FBS) and conditioned media from culture was saved at -20°C in 15 ml aliquots for future experimental use.

Method Details

Time-stamping method

We generated timestamp reporter mice by crossing Ai9 (RFP reporter) or eYFP reporter mice to CD4cre-ERT2 mice. We administered tamoxifen by i.p. or oral gavage to induce RFP expression. To mark the cells of newborns (1 d), 2.5 mg tamoxifen was administered to dams by oral gavage on days 0 and 1 (2.5 mg/mouse 2-3 times in a 24 hr period) and pups received tamoxifen through lactation. To mark 7 d and 28 d cells, mice were given 0.25 mg (single dose) and 1-2 mg tamoxifen (1-2 doses in a 24 hr period), respectively, at the appropriate age, also by oral gavage. Mice were then allowed to age to 8 weeks before use in experiments. The same procedure was use to generate TCR transgenic timestamp mice, except Ai9 mice were crossed to gBT-I mice to generate a TdTomato reporter mouse with a clonal TCR.

Flow Cytometry

All antibodies were purchased from Thermo Fisher Scientific or Biolegend. To directly assess antigen-specific cells, we used APC-labeled gBI tetramer obtained from the NIH

tetramer core facility. When fixation was required, we used the IC fixation and permeabilization kit from Thermo Fisher Scientific according to manufacturer's instructions. For intracellular staining we use the BD fix/perm buffer set with BD perm plus. Data was collected using an LSR II flow cytometer (BD Biosciences) and analyzed with Flowjo software (Treestar).

Tissue distribution

Blood samples were obtained by retro-orbital bleed. Mice were then euthanized, perfused and spleen, lymph nodes (cervical, mesenteric and inguinal), lung, liver and one femur were removed. Single cells suspensions of spleen, lymph node and bone marrow were made by manual dissociation and filtration through a 40 μM filter. Liver tissue was manually dissociated and enriched for lymphocytes using a percoll gradient. Lung tissue was manually dissociated, digested with 0.5mg/ml collagenase in RP-10 for 20 minutes at 37°C and layered on a percoll gradient to enrich for lymphocytes. Following preparation of single cell suspensions, cells were processed for flow cytometry.

In vitro proliferation

CD8+ T cells were isolated by positive magnetic selection using anti-CD8a microbeads (Miltenyi) according to manufacturer's instructions. Following isolation, cells were incubated with CFSE and, after labeling, placed in RPMI supplemented with 10% Fetal bovine serum, L-glutamine, Penicillin-streptomycin, 2-mercaptoethanol and 100 U/ml IL-2. Cells were plated on flat-bottomed 96-well plates that had been coated with anti-CD3 (clone 2Cll, 5 μg/ml) and anti-CD28 (37.51, 5 μg/ml) and cultured for times indicated. At analysis time points, cells were harvested, stained for additional surface markers and effector molecules and analyzed by flow cytometry.

In vitro bystander activation

CD8+ T cells were isolated by positive magnetic selection using anti-CD8a microbeads (Miltenyi) according to manufacturer's instructions. Following isolation, cells were incubated in RP-10 with IL-2 alone (100 U/ml) or IL-2 (100 U/ml), IL-12 (10 ng/ml) and IL-18 (10 ng/ml) for 18 hours. At that time, 1.5 μg/ml Brefeldin A was added to the cells and they were incubated an additional 4 hours. Cells were then harvested and stained for both surface and intracellular antibodies for flow cytometry as indicated above.

Dendritic cell immunizations

The B16-Flt3 tumor cell line was provided by Dr. Stephen Jameson (University of Minnesota, MN). Cells were cultured in RP-10 and conditioned media from culture was saved and frozen at -20°C in 15 ml aliquots. Dendritic cell donor mice were injected with 3-5x10^6 B16-Flt3 tumor cells subcutaneously. Tumors developed by 10-14 days post-injection. The day before harvest, these donor mice were injected with 2 ug LPS i.v. to mature dendritic cells. Spleens were collected the next day and digested in RP-10 containing 0.5 mg/ml collagenase-I for 20 minutes at 30°C, followed by addition of EDTA at a final concentration of 10 mM. After centrifugation, dendritic cells were resuspended in a 2:1

mixture RP-10 and B16-Flt3 conditioned media with 2 uM gB peptide and incubated for 2-3 hours at 37°C. Dendritic cells were then isolated using anti-CDllc microbeads (Miltenyi) according to manufacturer's instructions. Purity of cells was determined by flow cytometry and $1x10^6$ dendritic cells were injected i.v. into recipient mice. Recipient mice were given daily injections of either 200 ng IL-12 or PBS on days 0-3. On day° 5, splenocytes were harvested from recipients and prepared for flow cytometry as indicated above.

Infections

Dr. Sing Sing Way (Cincinnati Children's Hospital, OH) provided Wild-type *Listeria monocytogenes* expressing epitope HSV-1 $gB_{498-505}$ (Lm-gB)(Orr et al., 2007). Bacteria were grown to log phase and mice were injected intravenously (i.v.) with 5×10^3 colony forming units (CFU) in 100 ul of PBS, as previously described (Wang et al., 2016). At indicated timepoints, splenocytes were recovered. To enrich samples for CD8+ T cells, we depleted single cell suspensions of splenocytes of CD4+, CD19+, MHC II+ and Ter119+ cells by magnetic separation. Remaining cells were analyzed by flow cytometry.

Adoptive transfers and co-transfers

RFP expressing CD8+ cells from gBT-I Ai9 x CD4cre-ERT2 mice and CD8+ T cells from gBT-I x Thy1.1 mice were enriched by magnetic selection and then sorted to >90-95% purity using a FACS Aria III. For virtual memory experiments, CD4-CD8+$CD44^{hi}$ RFP+ cells, were sorted on the FACS Aria III and $5x10^3$ cells were injected (i.v.) into recipients. For co-transfers, RFP+CD8+ and CD8+Thy1.1+ were mixed at a 1:1 ratio and i.v. injected $1x10^4$ ($5x10^3$ of each) cells into each recipient. The next day, recipients were infected with Lm-gB. At indicated days post-infection, blood sample were taken by retro-orbital bleed and processed for flow cytometry.

Thymic transplants

Thymic transplants were performed using a previously described protocol (Morillon et al., 2015). Briefly, thymi were isolated from 0-1 day old RFP timestamp reporter mice. Thymi were separated into individual lobes and one lobe was inserted under the kidney capsule of eYFP timestamp reporter mice that were 6 weeks old. To mark cells coming out of the donor and recipient thymi simultaneously, 5 mg of tamoxifen was given on day 0-2 by oral gavage (three administrations in total). At weeks 2,3 and 4 post-surgery, recipients were bled and the phenotype of marked cells was assessed by flow cytometry

RNA preparation and sequencing

To isolate RNA from cells of interest, we magnetically enriched cells by positive selection with CD8a microbeads (Miltenyi) and sorted populations to > 90% purity with a FACS Aria III. For bulk CD8+ populations of time-stamped cells we sorted on CD4- CD8+ TdTomato+ cells. To look at virtual memory or true naïve cells we selected those cells that were CD4-CD8+$CD44^{hi}$ TdTomato+ or CD4-CD8+$CD44^{lo}$ TdTomato+, respectively. Cells were placed in Trizol and RNA was extracted according to manufacturer's instructions, with the addition of a chloroform extraction and Glycoblue (ThermoFisher) carrier prior to precipitation (1 hr on ice) and a second wash of the pelleted RNA in 70% ethanol. RNA

integrity was confirmed on a Fragment Analyzer (AATI). For total CD8+ samples (Figure 2), RNAseq libraries were generated with the NEBNext Ultra RNA Library Prep Kit (New England Biolabs) using 50ng total RNA. For the timestamp experiments in gBT-I mice, RNAseq libraries were generated with the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs) using RNA isolated from 10,000 cells (naïve experiment, ~15ng) or 25ng total RNA (5dpi experiment). All RNAseq libraries were sequenced with 81-85nt single-end reads on the NextSeq500 (Illumina). Raw reads were trimmed and filtered with cutadapt (Martin, 2011) (parameters -m 50 -q 20 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTC --match-read-wildcards). Reads were mapped to the mm10 genome with tophat2 (Kim et al., 2013) (parameters –G <UCSC:genes.gtf> --no-novel-juncs [--library-type fr-firststrand for directional libraries]); normalized FPKM values and differential gene expression analysis were generated with cuffdiff2 (Trapnell et al., 2013) (default parameters, UCSC:genes.gtf) using FDR cutoff 0.05. Additional criteria for the final list of differentially-expressed genes included a minimum FPKM value of 2 and minimum CPM value of 8 in at least one timestamp (expressed genes) and minimum 2-fold change between timestamps. JMP Pro 11 was used for PCA analysis of expressed and variable genes (expressed: at least 2 individual samples with FPKM ≥ 2 and CPM ≥ 5; variable: at least 2-fold range from lowest to highest FPKM value of individual samples). GSEA v2 (Subramanian et al., 2005) was used to test for gene enrichment of clusters defined by the Immunological Genome Project Consortium (expressed timestamp genes preranked by log2[d1/d28]; classic enrichment statistic) (Best et al., 2013)

ATACseq library preparation and sequencing

We magnetically enriched cells by positive selection with CD8a microbeads (Miltenyi) and sorted populations to > 90% purity with a FACS Aria III. VM cells were CD4-CD8+CD44 hi RFP+ and TN cells were CD4-CD8+CD44 lo RFP+. The FACS sorted cells were permeabilized in lysis buffer (10mM Tris-Cl pH 8.0, 300mM sucrose, 10mM NaCl, 2mM MgAc2, 3mM CaCl2, 0.1% NP-40(Igepal), 0.5mM DTT, 1x Pierce protease inhibitor (Thermo Scientific), and 40 units of RiboLock RNase inhibitor (Thermo Scientific)) for 10mins, washed with buffer W (10mM Tris-Cl pH 8.0, 300mM sucrose, 10mM NaCl, 2mM MgAc2, 0.008% Tween20, 0.5mM DTT, 1x protease inhibitor, and 40 units of RNase inhibitor), and resuspended in storage buffer (50mM Tris-Cl pH 8.3, 40% glycerol, 5mM MgCl2, 0.1mM EDTA, 0.5mM DTT, and 40 units of RNase inhibitor). The cells were pelleted at 1000xg for 10 minutes for each step of buffer change. The permeabilized cells were counted, snap-frozen in liquid nitrogen, and stored at -80°C.. ATACseq libraries were prepared from 30,000 permeabilized cells primarily as described in (Buenrostro et al., 2015) and (Corces et al., 2017). Frozen permeabilized cells were thawed, washed twice in ATAC-RSBbuffer containing 0.1% Tween20, resuspended in Transposition Mix (Nextera DNA Sample Preparation Kit, Illumina) containing 0.1% Tween20 and PBS, and incubated for 30 minutes at 37°C with shaking (1000rpm). After column cleanup, libraries were amplified using Nextera PCR oligos (Nextera Index Kit, Illumina) and Ultra II Q5 Master Mix (NEB) for a total of 12-14 PCR cycles based on qPCR quantification after the initial 5 cycles. The libraries were cleaned up twice (in series) with a 2:1 ratio of SPRIselect beads (Beckman Coulter). ATACseq libraries were sequenced with 61nt single-end reads on the HiSeq2500

(Illumina).

ATACseq read processing and peak calling

Raw sequencing reads were trimmed and filtered using cutadapt v1.12 (Martin, 2011), with parameters '-a CTGTCTCTTATACACATCT -e 0.06 -m 15'. The resulting highquality reads were aligned to the mouse genome (mm10) using bowtie2 v.2.2.8 (Langmead and Salzberg, 2012), with default parameters. The read alignments were filtered for PCR duplicates and lowquality alignments (MAPQ < 30) using samtools v1.4.1(Li et al., 2009). Peaks were called using macs2 v2.1.1.20160309 (Zhang et al., 2008b), with parameters '-f BAM -g mm -B -q 0.05'. To obtain confidently reproducible peaks, we used our two replicates of each sample and performed Irreproducibility Discovery Rate (IDR) analysis (Li et al., 2011), as previously described by the ENCODE Consortium (Gerstein et al., 2012), with some modifications. Briefly, we calculated IDR for each peak (command: batchconsistencyanalysis.r; parameters: peak.half.width=-1, min.overlap.ratio=0, ranking.measure=p.value) and filtered using IDR threshold of 0.1 to obtain consistent peaks between replicates. Despite being reproducible between replicates, boundaries of many peaks were slightly shifted between replicates. To generate one representative peak for each replicatepair, we called peaks separately for replicatepooled data and retained only those peaks that overlapped with reproducible peaks from the IDR analysis, resulting into a single set of peaks for each sample. To generate a unified set of peaks from all samples, we merged reproducible peaks from each celltype using bedtools v2.26.0 (Quinlan and Hall, 2010). Lastly, we filtered out all peaks that matched blacklist of artefactual regions in mm10 (https://www.encodeproject.org/annotations/ENCSR636HFF/) and further filtered peaks that contained reads with mapping quality of at least 38, resulting in total of 46,140 peaks. We used FeatureCounts from Subread v1.5.1 (Liao et al., 2014) to calculate raw read counts, with parameters '-F SAF -s 0 -Q 38', which we used as a measure of chromatin accessibility for further analysis. We used bedtools and custom Perl, Shell and R scripts and commands for further ATACseq data analysis.

ATACseq visualization

The read alignments that map to unified set of peaks were extracted. The read coverage at each genomic position was normalized (RPM) using *genomeCoverageBed* (Quinlan and Hall, 2010), converted to bigwig format using *bedGraphToBigWig* from USCS toolkit, and visualized with the USCS genome browser.

ATACseq data clustering and gene association

Reads were first counted for each peak in replicatepooled data, and then counts were converted to zscores across samples (rowwise zscores) prior to clustering analysis. Morpheus (Broad Institute) was used for kmeans clustering, and visualization using heatmaps. Each peak was assigned to a nearest gene based on the shortest distance between the peak and gene's promoter, on either strand; promoters were defined as 1 kb upstream and 500 bp downstream of annotated transcription start site (TSS) in Gencode vM12 annotations. Each gene was assigned to a peakcluster that had the most number of geneassociated peaks; for

ties, genes were randomly assigned to one of the tied clusters.

Motif enrichment analysis for transcription factor binding sites

Putative transcription factor binding sites (TFBS) were identified in ATACseq peaks by first obtaining the 1117 motifs corresponding to binding sites for 583 distinct transcription factors (TF) from JASPAR v2018 (Khan et al., 2018) and CISBP (Weirauch et al., 2014). For CISBP motifs, we collected only those motifs found in TRANSFAC (Matys et al., 2006) or described previously (Jolma et al., 2013). These TF binding motifs were then searched in 150bp of sequence centered on the summit of each peak to identify putative TFBS, using FIMO with pvalue cut-off of 10e5 (Grant et al., 2011). Enrichment of TF binding motif was computed over different subsets of peaks, as described previously (Yu et al., 2017) with some changes. Briefly, we first defined subsets of peaks, denoted by *S*, that are significantly (qvalue < 0.1)(Storey and Tibshirani, 2003) more accessible in one sample compared to another in different pair-wise comparisons, using a pipeline built upon the edgeR v3.18.1 (Robinson et al., 2010) framework for differential gene expression analysis. We then computed the fraction of peaks in a subset *s*, and fraction of peaks in a set of 10,000 randomly selected peaks that contain at least one binding motif for a TF *t*, denoted by *fs* and *fr*, respectively. The ratio of *fs/fr* was used to determine enrichment (ratio > 1) or depletion (ratio < 1) of binding motif for TF *t* in a subset *s*. The significance P-value of enrichment or depletion was computed using a binomial test, where the set of 10,000 randomly selected peaks was used as the null distribution, followed by FDR correction. We applied stringent thresholds of 10e5 for FDR corrected binomial P-value to obtain confidently enriched or depleted TF binding motifs.

Poised gene analysis

Poised genes were defined as genes that have at least one significantly (qvalue < 0.05) more accessible region in one sample compared to another, using ATACseq data, and, without evidence for transcript upregulation, assessed using RNAseq data. Gene ontology (GO) and GO-slim enrichment analysis on poised genes was performed using GOrilla (Eden et al., 2009) and BiNGO (Maere et al., 2005), respectively; a list of all expressed genes was used as a background set for the analysis. The P-values were FDR corrected, and the significant (P-value < 0.05) terms were visualized using Cytoscape (Shannon et al., 2003).

Mathematical modeling

The rate of division and death was estimated using the precursor cohort method as previously described (Gett and Hodgkin, 2000; Hawkins et al., 2007). Briefly, the number of cell divisions was calculated based on their CFSE dilution. The average division number (based on the corrected cell number) was then used to estimate time-to first division, division rate and death rate of cells. The same method was also recently used to quantify the rate of division of neonatal and adult $CD8^+$ T cell *in vitro* (Reynaldi et al., 2016).The rate of cellular differentiation was quantified using the assumption of simple division-linked differentiation. That is, assuming that on each division a cell will have a probability of differentiation (ie: to gain or lose a specific surface marker) (Reynaldi et al., 2016; Schlub et

al., 2010). For example, on each division, cell might have a probability of losing the CD62L^{hi} phenotype and gaining a CD62L^{low} phenotype. Thus, the proportion of cells that are CD62L^{hi} on each division can be written as

$$H(n) = H_0(1 - c)^n,$$

where H_0 is the initial proportion of CD62L^{hi} on division zero, and c is the loss rate on each division. We used GraphPad Prism (version 6, GraphPad Software, California) to test whether each parameter is equal between the three groups (F test).

Quantification and Statistical Analysis

Statistical analyses (except for RNA and ATAC sequencing, please see above) were performed using Prism software (Graphpad). Error bars represent SEM or SD, as indicated in figure legends. Significance was determined by 1-way or 2-way ANOVA followed by an appropriate post test, as indicated in the figure legends. Significance is denoted as follows: * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$.

Data and Software Availability

The RNAseq and ATACseq data reported in this paper have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO SuperSeries accession number GSE97802 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE97802).

ACKNOWLEDGEMENTS

We thank the NIH Tetramer Core Facility for Kb:gB$_{489-505}$ tetramers. Denny Totman from Cornell Center for Animal Resource and Education (CARE) provided expert mouse breeding assistance. Cell sorting was done at Cornell University's Biotechnology Resource Center with the assistance of Carol Bayles and Adam Wojno. This work was supported by National Institute of Health awards R01AI105265 and R01AI110613 (to B.D.R., from the National Institute of Allergy and Infectious Disease) and U01AI131348 (to B.D.R and A.G., from the National Institute of Allergy and Infectious Disease), and Core B of P50HD076210 (to A.G. and J.G., from National Institute of Child Health and Human Development). MD is supported by an NHMRC (Australia) Senior Research Fellowship (1080001).

AUTHOR CONTRIBUTIONS

N.S. planned and performed experiments, analyzed and interpreted data, and wrote the manuscript. R.P., A.R., J.G. and A.G. generated, analyzed and interpreted genomics data and wrote the manuscript. J.W., N.W., K.N., K.Y., and S.P. performed experiments. M.D. and B.R. conceptualized the study, planned experiments and wrote the manuscript.

FIGURES

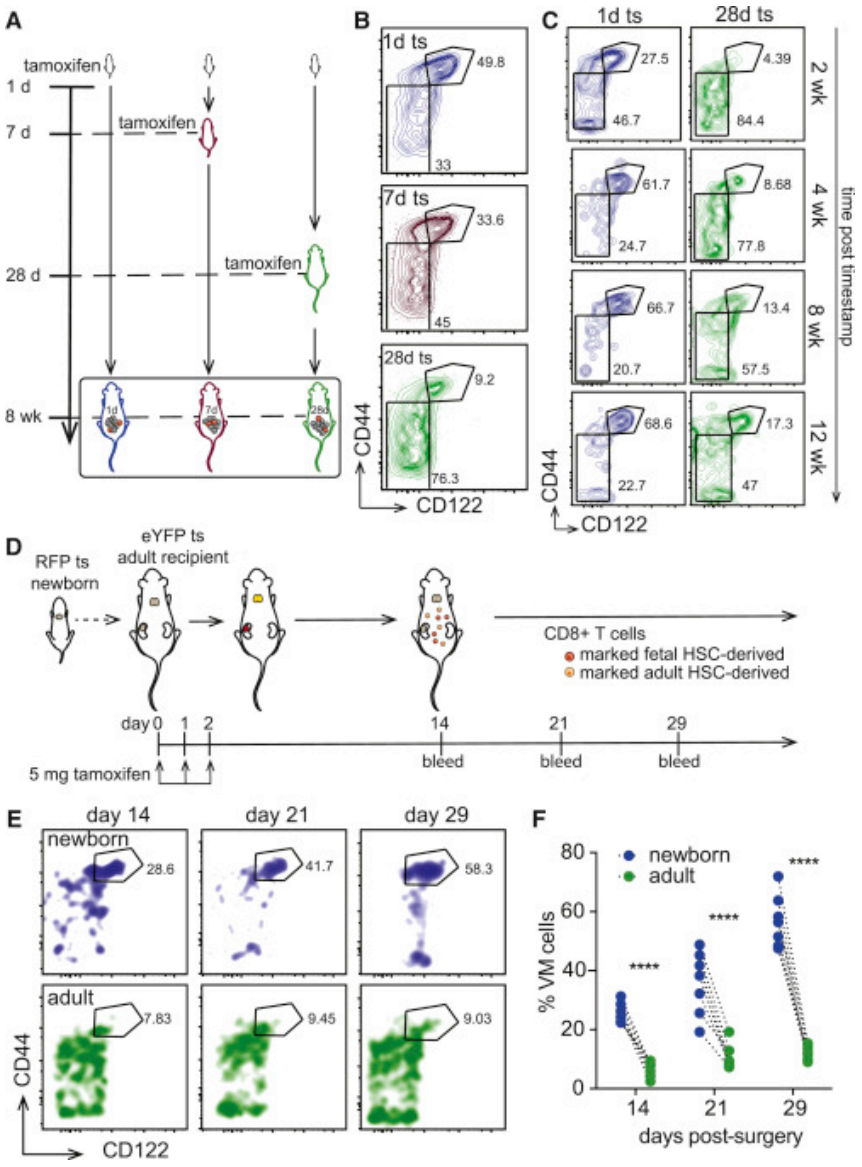

Figure 1. Cells made early in life have an inherent propensity to form memory phenotype cells. (A) Timestamping set up (B) expression of CD44 and CD122 in 8 wk old timestamp mice exposed to tamoxifen at 1d (blue), 7d (maroon) and 28d (green). (C) CD44 and CD122 expression in stamped CD8+ T cells where comparisons across rows are cells that are the same time interval post-marking. (D) Thymic transplantation schematic (E) CD44 and CD122 expression post-transplant. (F) %VM in cells of adult and neonatal origin in transplant mice. Statistical significance by 2-way ANOVA and Sidak's multiple comparison correction. For all parts, data representative of 2-4 independent experiments, n=3-8 mice.

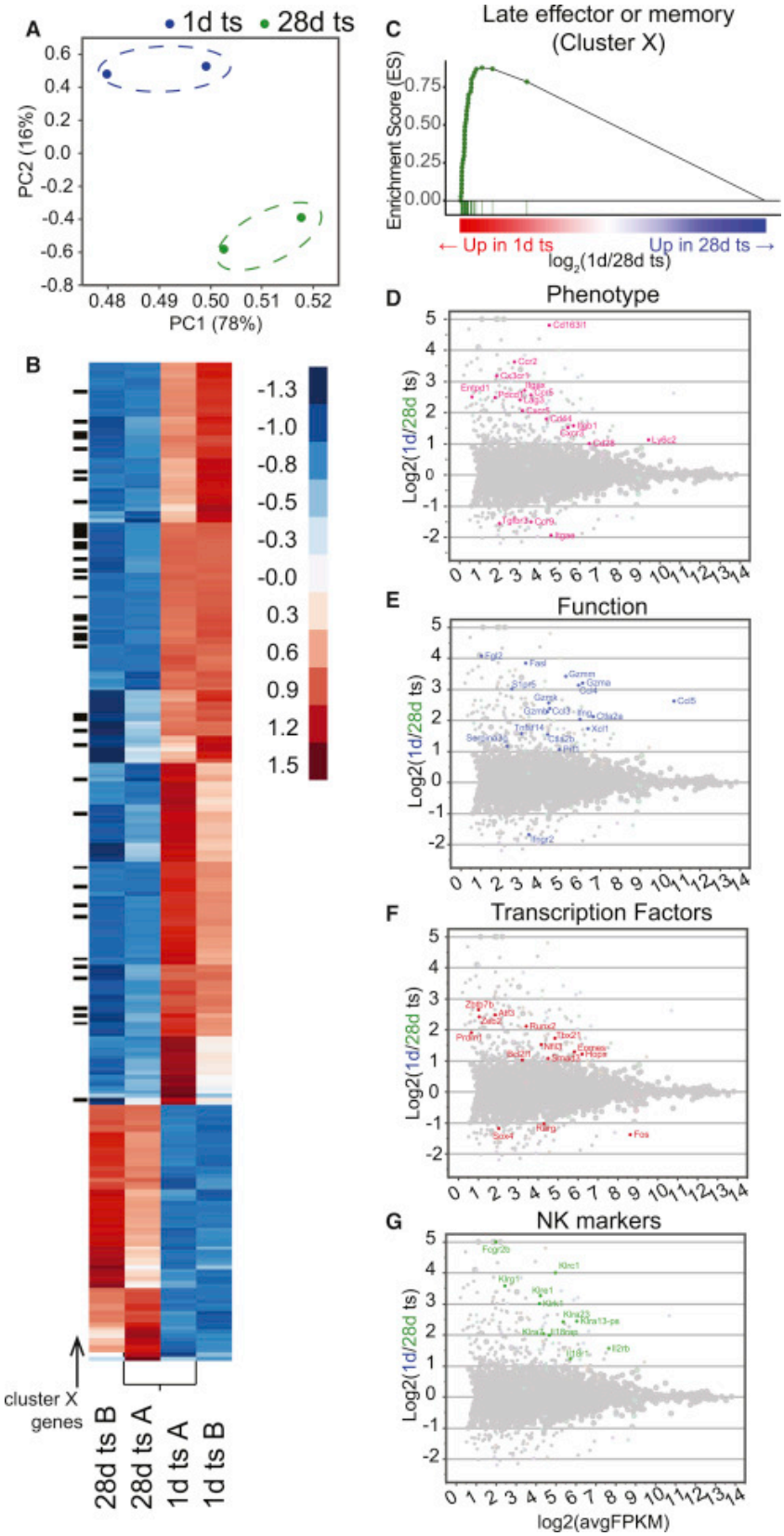

Figure 2. CD8+ T cells with different developmental origins have differential gene expression. (A) Principle component analysis of genome-wide variation in 1d ts and 28d ts

CD8+ T cells (B) Gene expression profiles of RNAseq samples, differentially-expressed genes in ImmGen cluster X (black bars); heatmap of row-normalized gene expression. (C) Gene set enrichment score plot for ImmGen cluster X. (D-G) differential expression of markers for phenotype, function, natural killer cells and transcription factors, respectively.

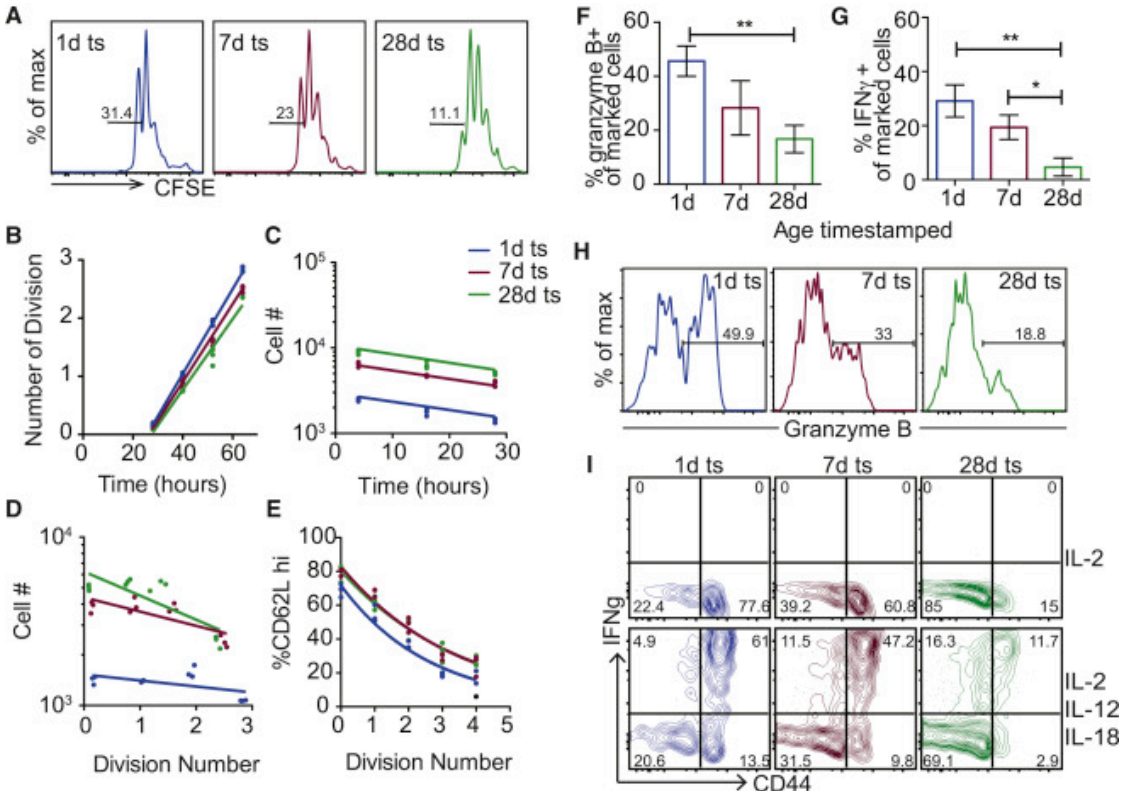

Figure 3. CD8+ T cells made early in life preferentially respond to antigen and inflammatory cytokines. (A) CFSE dilution after 52 hrs with anti-CD3/CD28 (B) Corrected mean division number (C) Corrected cell number over stimulation time (D) Corrected cell number per round of division and (E) At 52 hrs post-stimulation, % $CD62L^{hi}$ in each division. (F-G) Granzyme B and IFNγ after 18 hr with anti-CD3/CD28 stimulation. Statistical significance by 1-way ANOVA and Tukey's multiple comparisons. Error bars represent SD. (H) Granzyme B production 48 hours after antigen stimulation. (I) IFNγ and CD44 expression in timestamped cells stimulated with Il-12 and Il-18. For all parts, 2 experiments, n= 3.

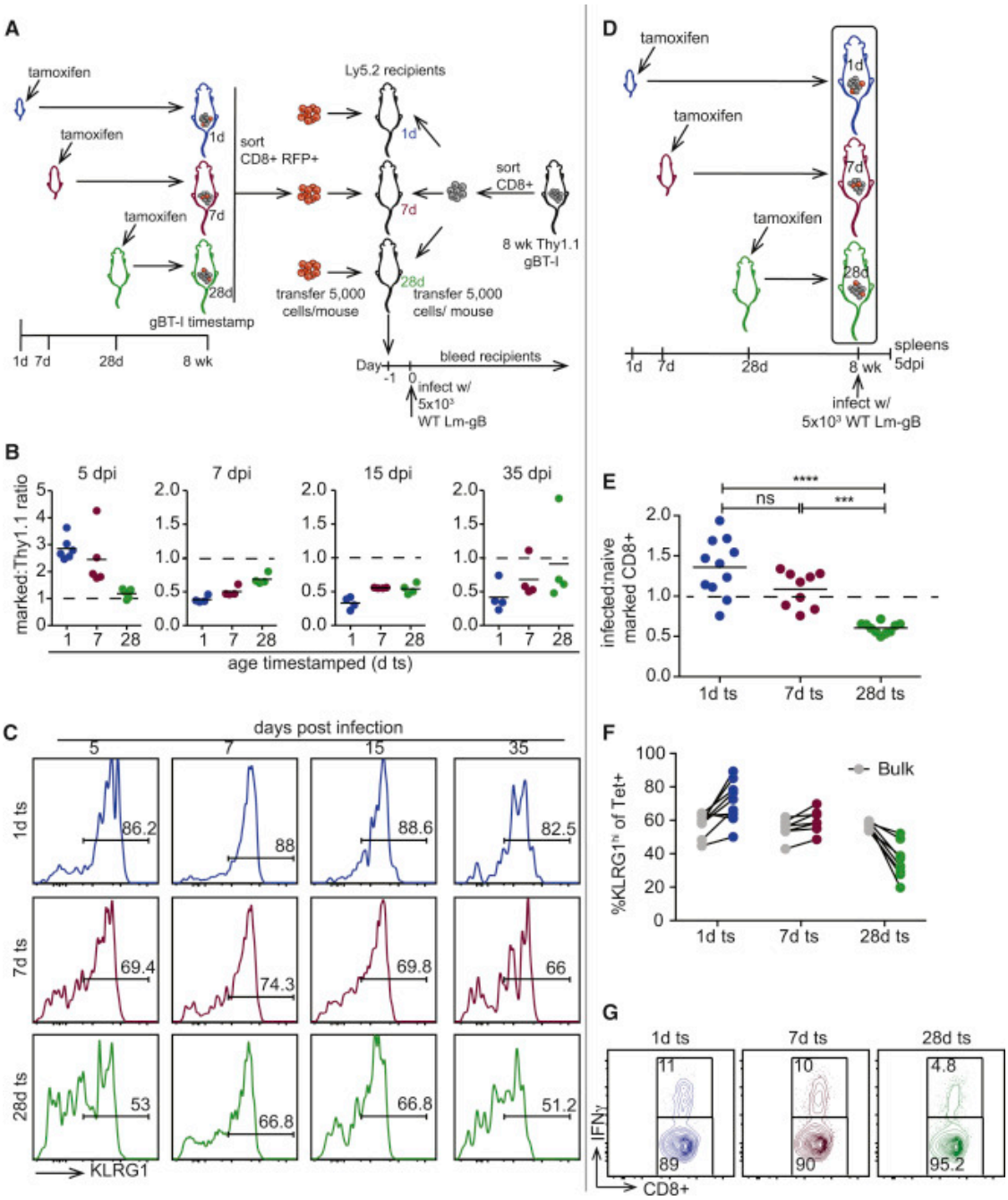

Figure 4. Cells made early in life constitute the early effector response during infection in adulthood. (A) Experimental schematic for B,C. (B) Ratio of timestamped to bulk gBT-I (WT) at days indicated. (C) KLRG1 expression at days indicated. B-C, data are representative of 2 experiments, n= 4-6 mice. (D) Experimental schematic for E-G. (E)Post- to pre-infection ratio of ts cells. Statistical significance determined by 1-way ANOVA and Tukey's multiple comparisons. (F) percentage of ts (color) or total (gray) tetramer+ KLRG1^{hi} cells in each animal. (G) IFNγ expression by CD8+ timestamped T cells. (E-G) Data representative of 2 experiments, n=8-11 mice.

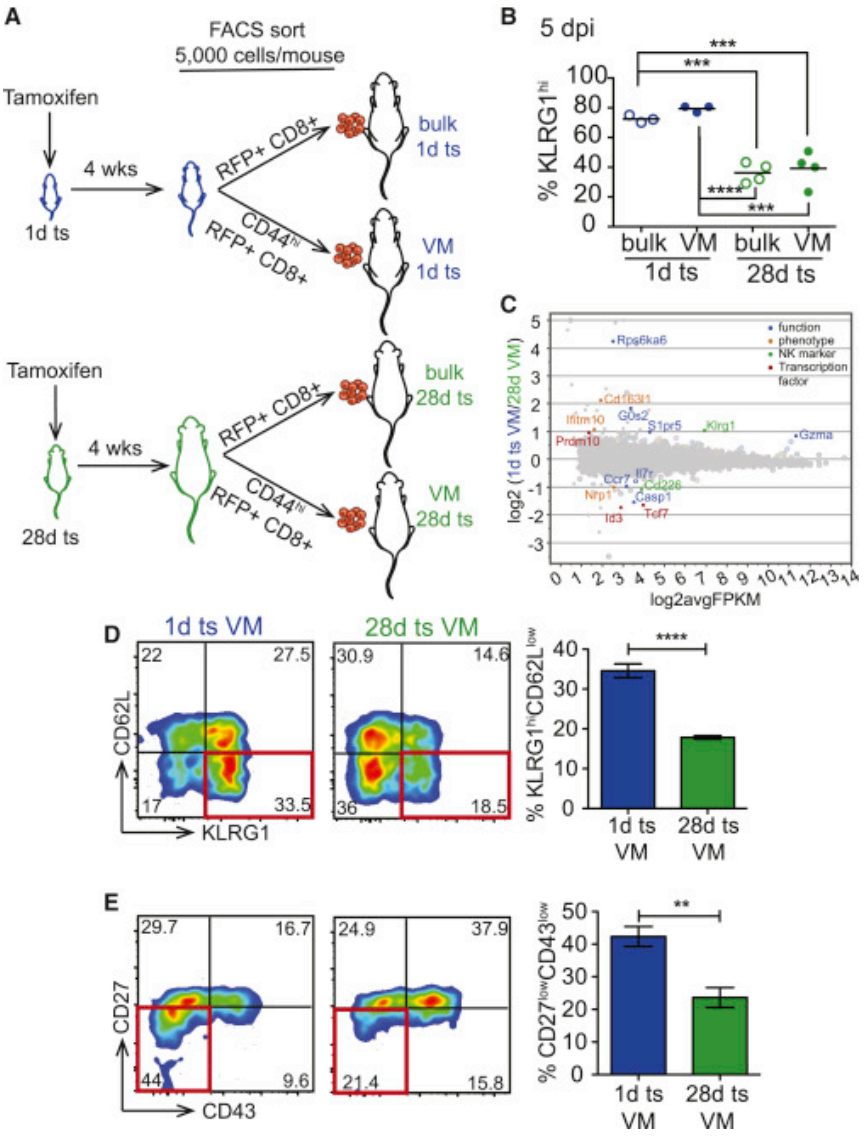

Figure 5. The long-term fates of cells following infection are linked to their developmental origins. (A) Experimental schematic. (B) KLRG1^{hi} cells in the blood on 5dpi; Two experiments, n=3-4 mice. Statistical significance determined by 1-way ANOVA followed by Tukey's multiple comparison correction. (C) differential gene expression of selected genes of interest from RNA-seq of 1d ts VM and 28d ts VM at 5 dpi. (D) CD62L and KLRG1 surface expression at 41 dpi. (E) CD27 and CD43 surface expression at 41 dpi. For E-F, two experiments, n=4 mice. Statistical significance determined by student T-test.

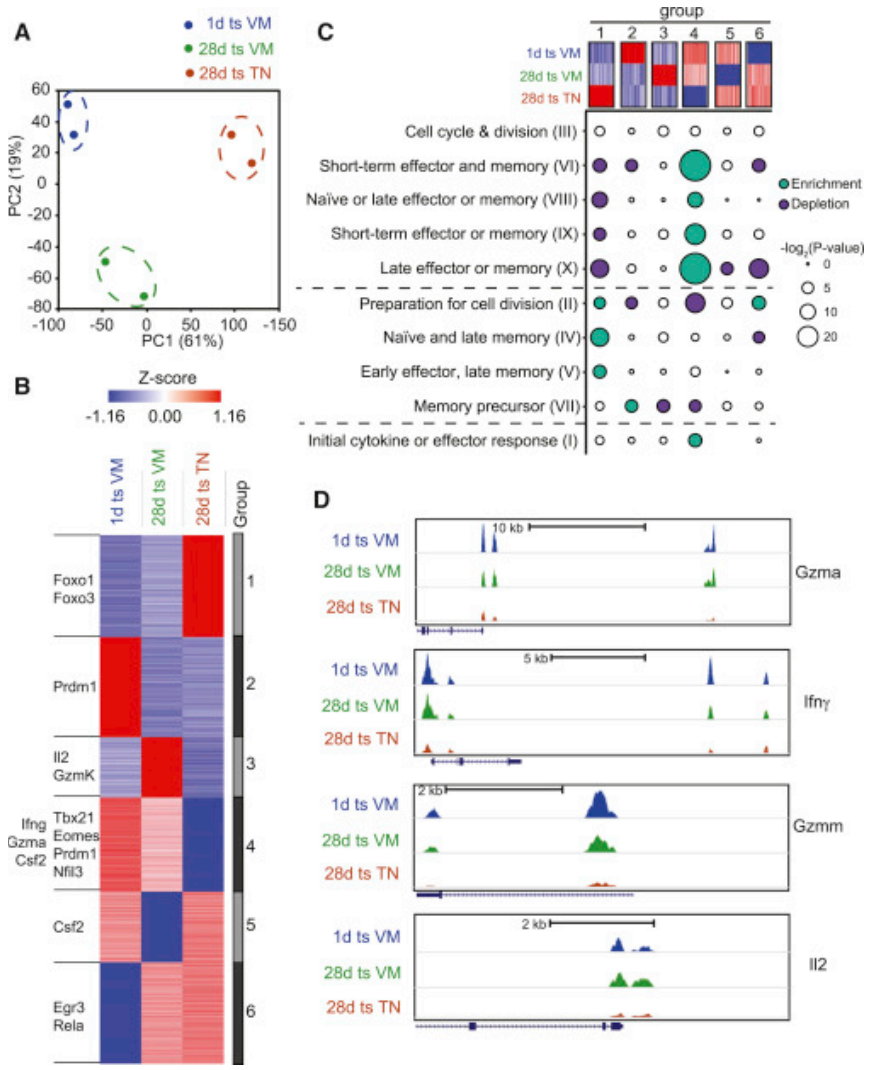

Figure 6. CD8+ T cells with early developmental origin exhibit effector-like chromatin landscape. (A) Principle component analysis for global chromatin accessibility measured using ATACseq in 1d ts VM, 28 ts VM and 28d ts TN CD8+ T cells. (B) Clustering (k-means analysis; k=6) of ATACseq intensities in total peaks. Intensities determined by row-wise z-scores of RPM (reads per million). (C) Enrichment analysis for ImmGen clusters across different groups (1-6) defined in panel B. A ratio of observed number of genes over expected number of genes of an ImmGen cluster in a given group was calculated to determine enrichment (green circles; ratio > 1 & P-value < 0.05) or depletion (purple circles; ratio < 1 & P-value < 0.05). The significance P-values were calculated using Fisher's exact test, followed by FDR correction. (D) UCSC browser views of variably accessible regions between different samples.

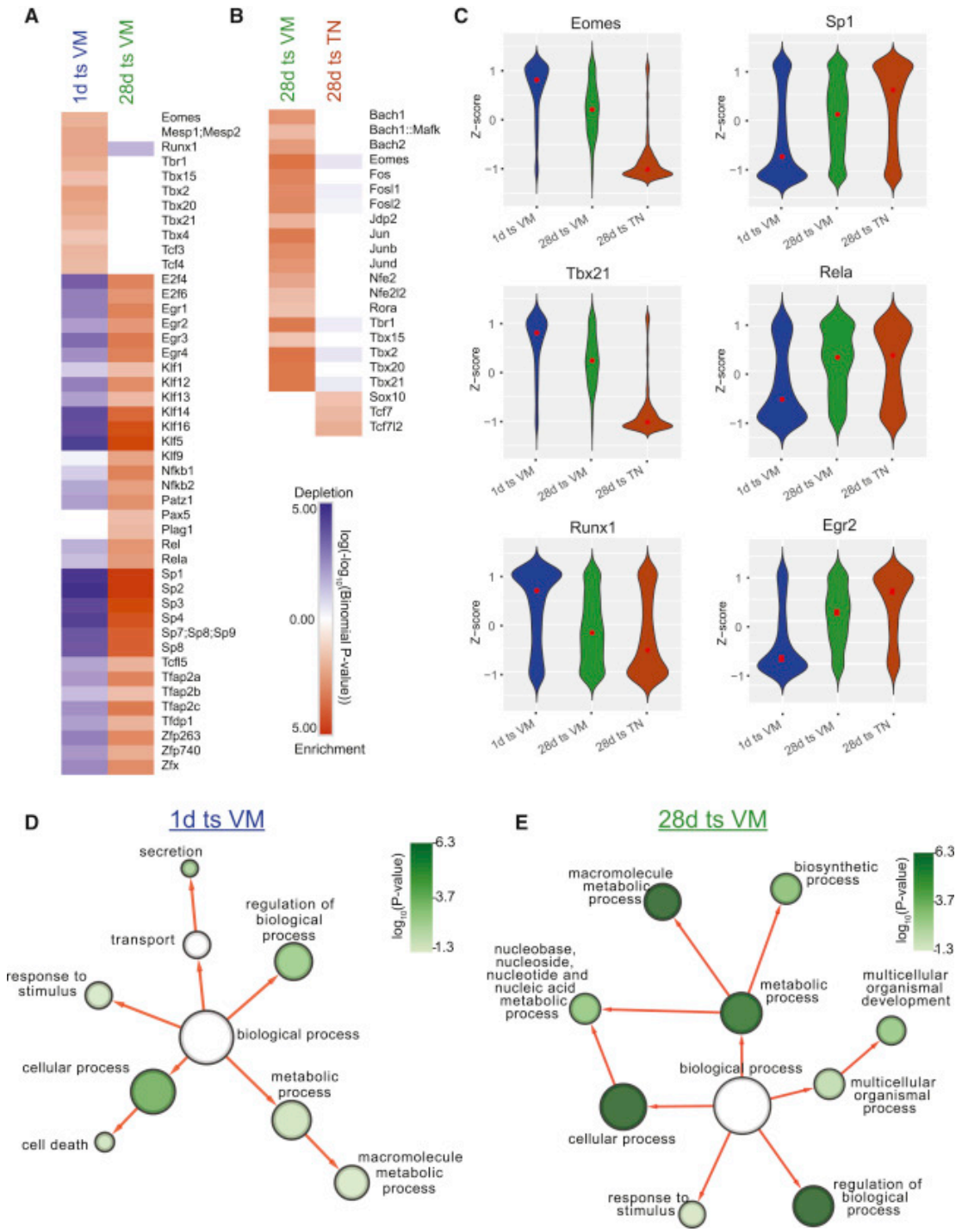

Figure 7. The regulatory landscape of CD8+ T cells made early in life is more responsive to effector-like TFs. (A,B) Enrichment (red) or depletion (blue) of TF binding motifs in differentially accessible chromatin regions between 1d ts and 28d ts VM CD8+ T cells (A) or between 28d-marked VM and TN CD8+ T cells (B). The significance is color-coded and was determined with a binomial test using randomly selected open chromatin regions as background, followed by FDR correction. (C) Distribution of relative chromatin accessibility at regions with binding motifs for respective TFs. Relative chromatin accessibility is

represented by z-score as in Fig 6B. (D,E) Enrichment analysis for GO-slim 'biological process' terms in poised genes in 1d ts VM and 28d ts VM, respectively. The significant (P-value < 0.05) terms are presented as networks; the size of the circle indicates enrichment score, and FDR corrected enrichment P-value is color-coded.

SUPPLEMENTAL FIGURES

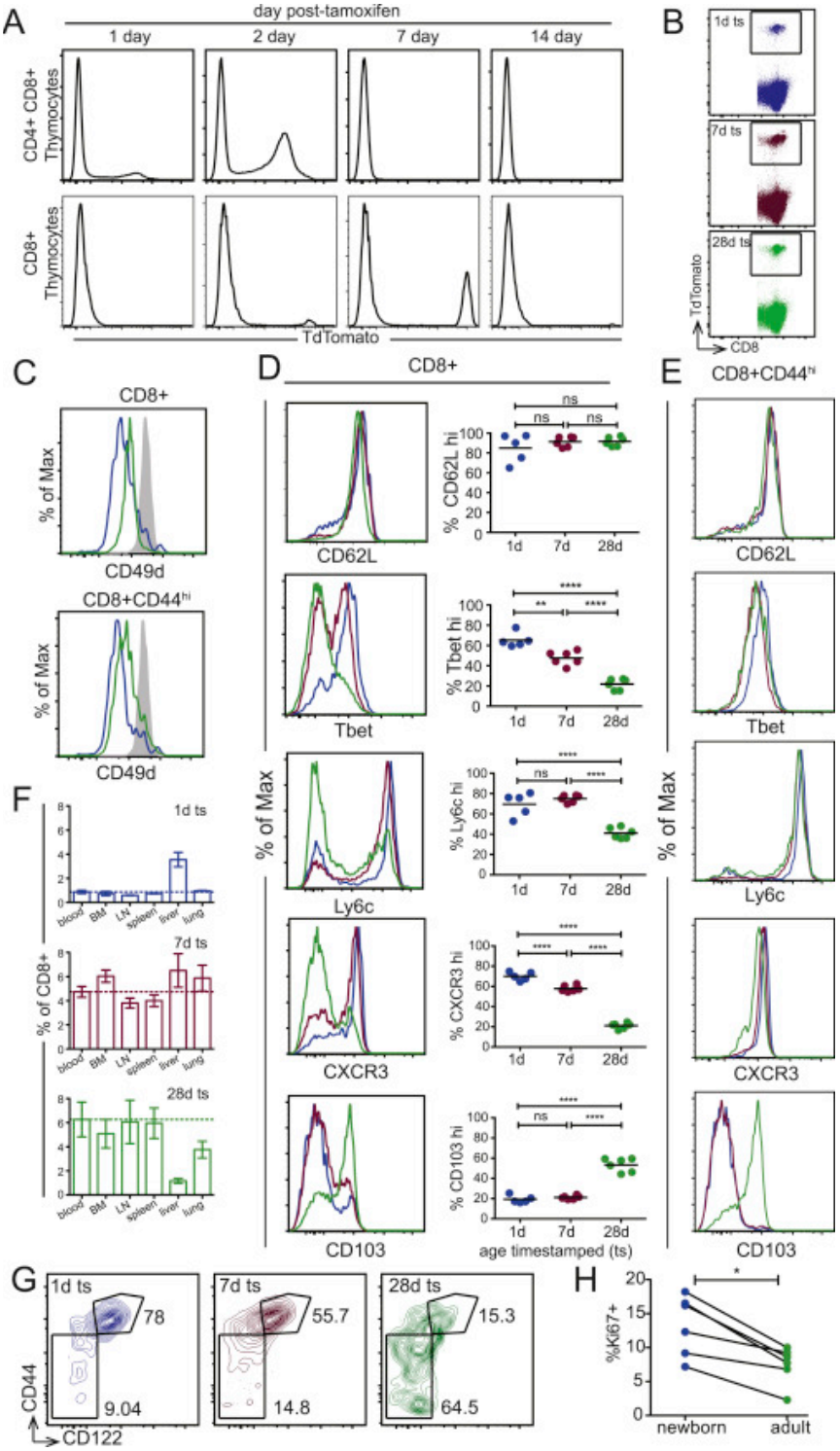

Figure S1. (Related to Figure 1) (A) RFP expression in double positive thymocytes and single CD8+ thymocytes (B) RFP expression on CD8+ splenocytes in 8-wk-old marked mice. (C) CD49d protein expression in CD8+ and CD8+CD44^{hi} cells. Grey histogram is true memory control. (D) histograms and statistical analysis of surface markers in timestamped adults. (E)

surface marker expression on VM timestamped CD8+ T cells. (F) % of CD8+ T cells in tissues throughout the body at 8 weeks of age (error bars represent SEM). Statistical significance determined by 1-way ANOVA and Tukey's multiple comparisons. (G) expression of CD44 and CD122 in RFP expressing populations of CD8+ T cells in the blood of 8-week-old gBT-I timestamp. (H) Percent Ki67+ stamped cells 2 weeks post-thymic transplant. Statistical significance determined by paired student T test. Data representative 2 experiments, n=3-8 mice.

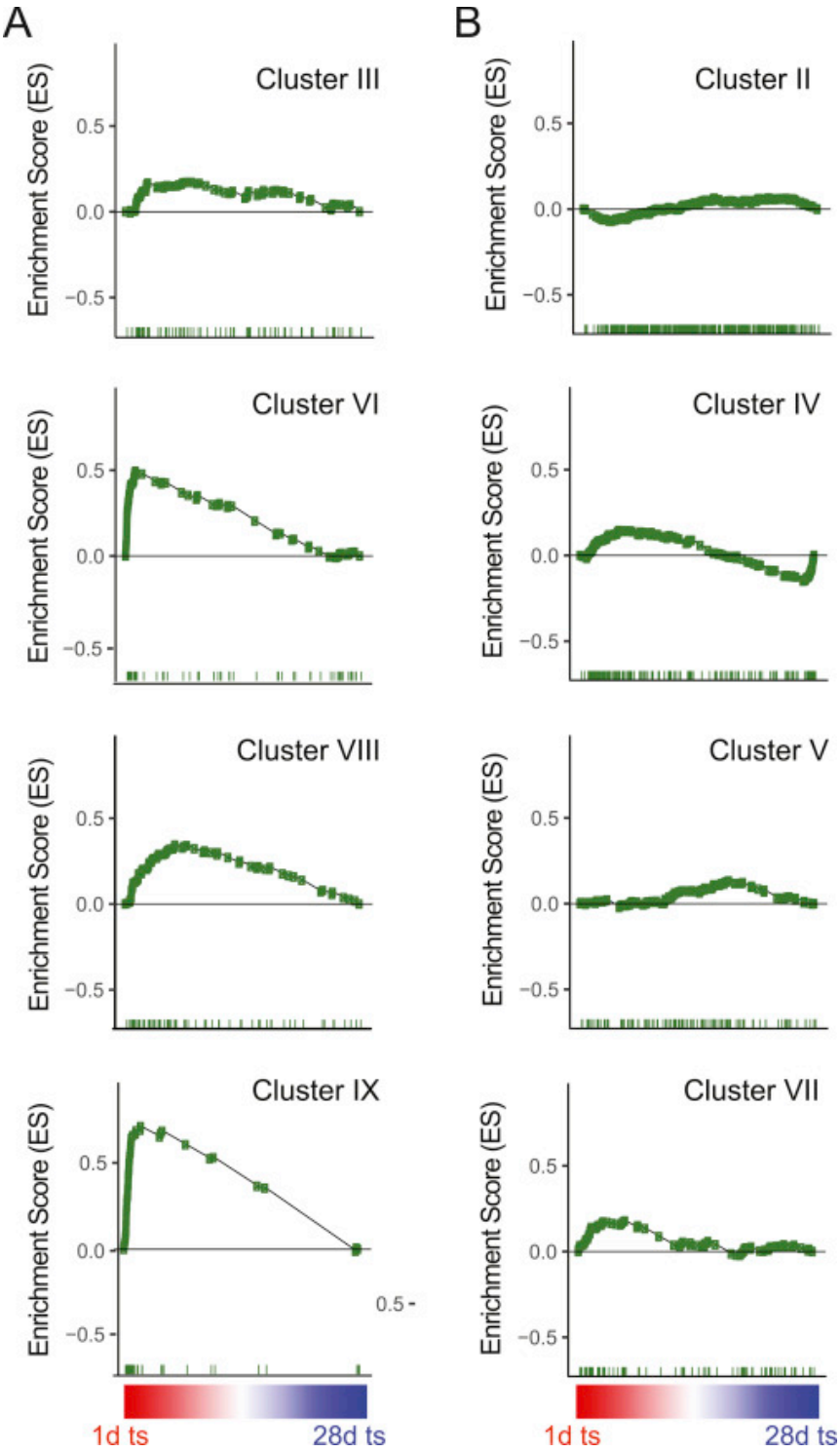

Figure S2. (Related to Figure 2). GSEA enrichment score plots for ImmGen gene clusters containing (A) effector-like genes and (B) naïve/memory-like genes.

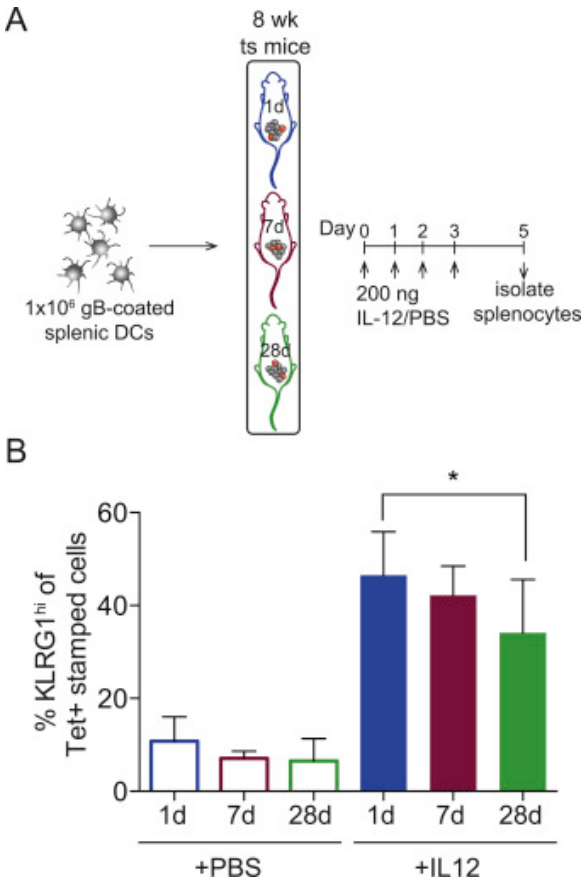

Figure S3. (Related to Figure 4) (A) Dendritic cell immunization model. (B) % KLRG1^{hi} cells on 5 days post-immunization. Data representative of 2 experiments n=4-6 mice. Statistical significance determined by 1-way ANOVA and Tukey's multiple comparisons.

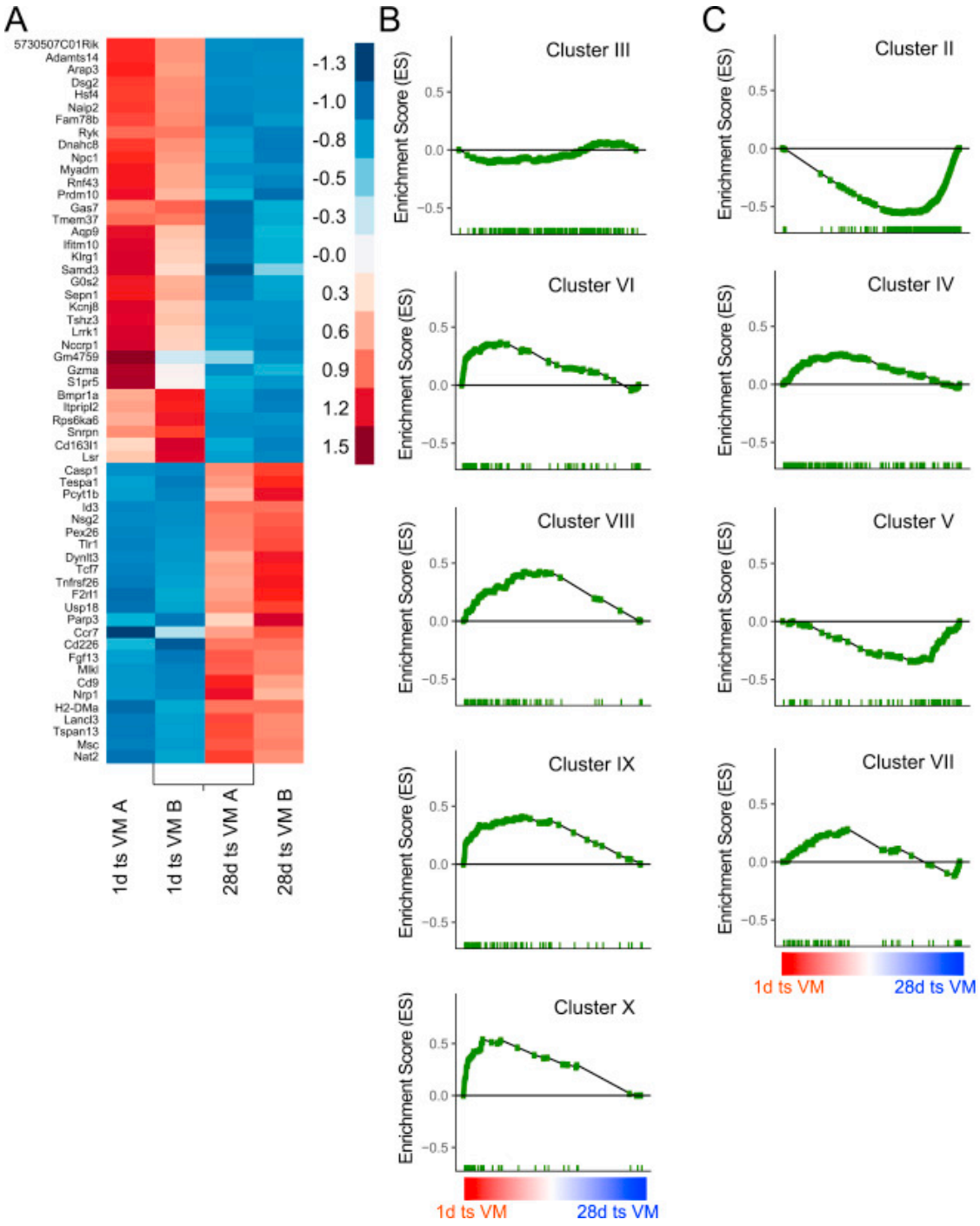

Figure S4. (Related to Figure 5). Analysis of gene expression comparing 1d ts and 28d ts RNAseq from virtual memory phenotype cells at 5dpi with *L. monocytogenes*. (A) Differential gene expression; only genes meeting stringent criteria (see STAR Methods) are shown. (B,C) GSEA enrichment score plots for ImmGen gene clusters containing (B) effector-like genes and (C) naïve/memory-like genes.

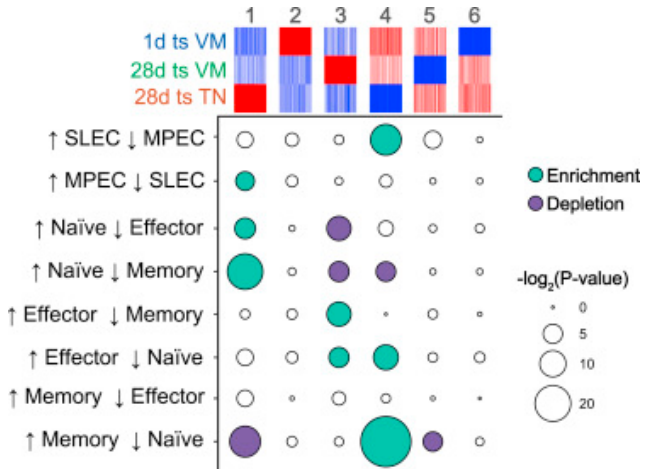

Figure S5. (Related to Figure 6.) Enrichment analysis for gene-sets from two independent studies across different groups (1-6), otherwise as described in Figure 6B. The P-values were calculated using Fisher's exact test with FDR correction.

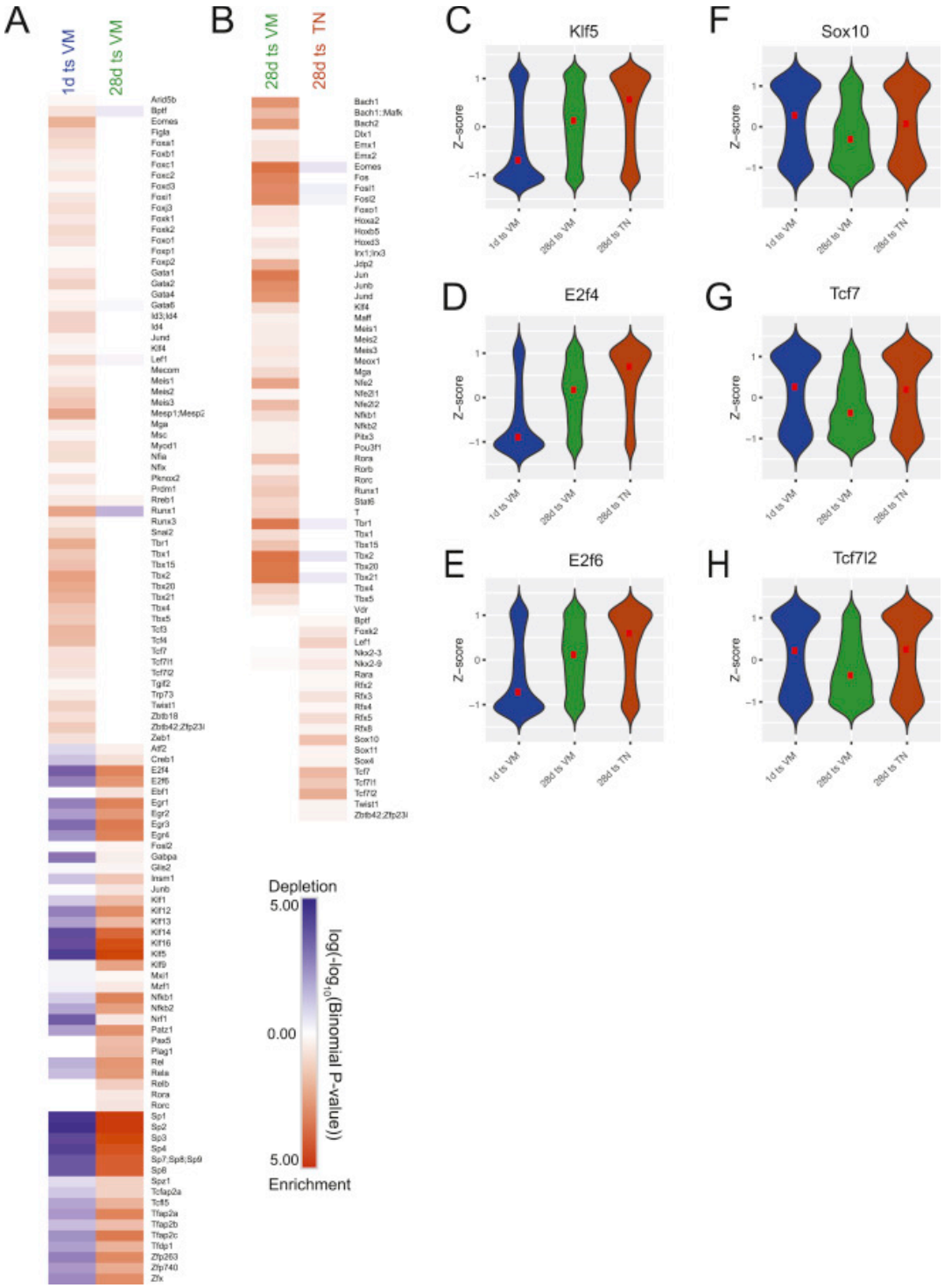

Figure S6. (Related to Figure 7.) (A-B) Enrichment (red) or depletion (blue) of TF binding motifs in differentially accessible chromatin regions between 1d ts and 28d ts VM CD8+ T cells (A), or 28d ts VM and TN CD8+ T cells (B), as described in Figure 7; includes all binding motifs with FDR corrected binomial P-values of 0.05 or smaller. (C-H) Distribution of relative chromatin accessibility across different samples at chromatin regions that contain matches to indicated TF binding motif.

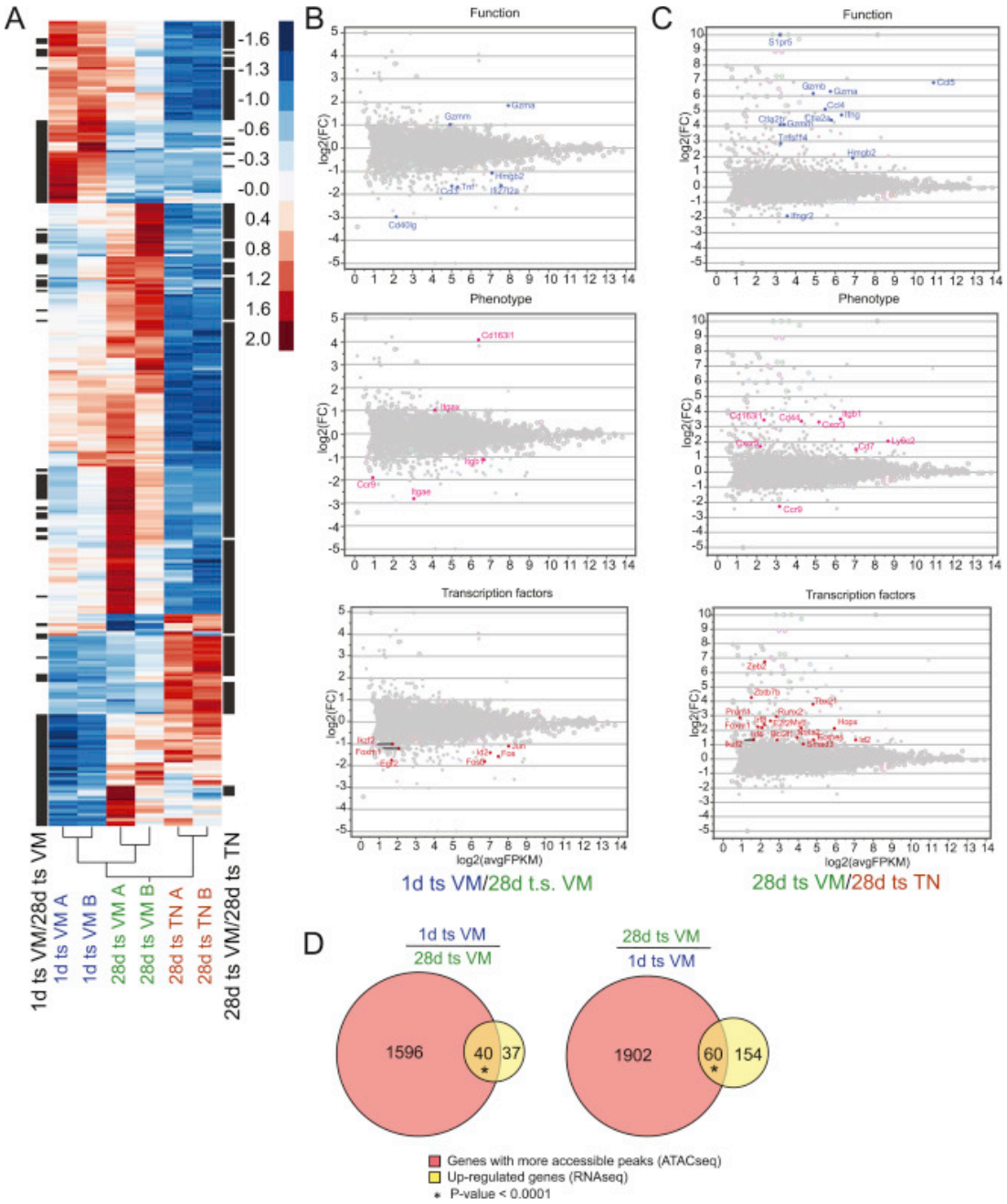

Figure S7. (Related to Figure 7.) Analysis of gene expression comparing 1d ts VM, 28d ts VM cells and 28d ts TN cells. (A) Heatmaps of differentially expressed genes between 1d ts VM versus 28d ts VM cells (black bars on left) or between 28d ts VM versus 28d ts TN cells (black bars on right). Only genes meeting stringent criteria for differential expression are shown (see Methods). (B,C) MA plots highlight differential expression of markers for phenotype, function, and transcription factors between (B) 1d ts VM and 28d ts VM cells or (C) 28d ts VM and 28d ts TN cells. (D) Venn-diagrams representing concordance between ATACseq and RNAseq data. Differentially expressed genes (RNAseq) compared with the genes associated with differentially accessible peaks (ATACseq) in two comparisons: 1d ts VM versus 28d ts VM or 28d ts VM versus 1d ts VM.

REFERENCES

Adkins, B. (1991). Developmental regulation of the intrathymic T cell precursor population. J Immunol *146*, 1387-1393.

Adkins, B., Leclerc, C., and Marshall-Clarke, S. (2004). Neonatal adaptive immunity comes of age. Nat Rev Immunol *4*, 553-564.

Adkins, B., Williamson, T., Guevara, P., and Bu, Y. (2003). Murine neonatal lymphocytes show rapid early cell cycle entry and cell division. J Immunol *170*, 4548-4556.

Aghajani, K., Keerthivasan, S., Yu, Y., and Gounari, F. (2012). Generation of CD4CreER(T(2)) transgenic mice to study development of peripheral CD4-T-cells. Genesis *50*, 908-913.

Akue, A.D., Lee, J.Y., and Jameson, S.C. (2012). Derivation and maintenance of virtual memory CD8 T cells. J Immunol *188*, 2516-2523.

Badovinac, V.P., Haring, J.S., and Harty, J.T. (2007). Initial T cell receptor transgenic cell precursor frequency dictates critical aspects of the CD8(+) T cell response to infection. Immunity *26*, 827-841.

Berland, R., and Wortis, H.H. (2002). Origins and functions of B-1 cells with notes on the role of CD5. Annu Rev Immunol *20*, 253-300.

Best, J.A., Blair, D.A., Knell, J., Yang, E., Mayya, V., Doedens, A., Dustin, M.L., Goldrath, A.W., and Immunological Genome Project, C. (2013). Transcriptional insights into the CD8(+) T cell response to infection and memory T cell formation. Nat Immunol *14*, 404-412.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol *109*, 21 29 21-29.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B.*, et al.* (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods *14*, 959-962.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Fulton, R.B., Hamilton, S.E., Xing, Y., Best, J.A., Goldrath, A.W., Hogquist, K.A., and Jameson, S.C. (2015). The TCR's sensitivity to self peptide-MHC dictates the ability of naive CD8(+) T cells to respond to foreign antigens. Nat Immunol *16*, 107-117.

Gavin, M.A., and Bevan, M.J. (1995). Increased peptide promiscuity provides a rationale for the lack of N regions in the neonatal T cell repertoire. Immunity *3*, 793-800.

Gerlach, C., Moseman, E.A., Loughhead, S.M., Alvarez, D., Zwijnenburg, A.J., Waanders, L., Garg, R., de la Torre, J.C., and von Andrian, U.H. (2016). The Chemokine Receptor CX3CR1 Defines Three Antigen-Experienced CD8 T Cell Subsets with Distinct Roles in Immune Surveillance and Homeostasis. Immunity *45*, 1270-1284.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R.*, et al.* (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91-100.

Gett, A.V., and Hodgkin, P.D. (2000). A cellular calculus for signal integration by T cells. Nature immunology *1*, 239-244.
Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.
Gray, S.M., Amezquita, R.A., Guan, T., Kleinstein, S.H., and Kaech, S.M. (2017). Polycomb Repressive Complex 2-Mediated Chromatin Repression Guides Effector CD8(+) T Cell Terminal Differentiation and Loss of Multipotency. Immunity *46*, 596-608.
Haluszczak, C., Akue, A.D., Hamilton, S.E., Johnson, L.D., Pujanauski, L., Teodorovic, L., Jameson, S.C., and Kedl, R.M. (2009). The antigen-specific CD8+ T cell repertoire in unimmunized mice includes memory phenotype cells bearing markers of homeostatic expansion. J Exp Med *206*, 435-448.
Hardy, R.R., and Hayakawa, K. (1991). A developmental switch in B lymphopoiesis. Proc Natl Acad Sci U S A *88*, 11550-11554.
Hawkins, E.D., Hommel, M., Turner, M.L., Battye, F.L., Markham, J.F., and Hodgkin, P.D. (2007). Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. Nat Protoc *2*, 2057-2067.
He, B., Xing, S., Chen, C., Gao, P., Teng, L., Shan, Q., Gullicksrud, J.A., Martin, M.D., Yu, S., Harty, J.T.*, et al.* (2016). CD8(+) T Cells Utilize Highly Dynamic Enhancer Repertoires and Regulatory Circuitry in Response to Infections. Immunity *45*, 1341-1354.
Herzenberg, L.A., and Herzenberg, L.A. (1989). Toward a layered immune system. Cell *59*, 953-954.
Hikono, H., Kohlmeier, J.E., Takamura, S., Wittmer, S.T., Roberts, A.D., and Woodland, D.L. (2007). Activation phenotype, rather than central- or effector-memory phenotype, predicts the recall efficacy of memory CD8+ T cells. J Exp Med *204*, 1625-1636.
Ikuta, K., Kina, T., MacNeil, I., Uchida, N., Peault, B., Chien, Y.H., and Weissman, I.L. (1990). A developmental switch in thymic lymphocyte maturation potential occurs at the level of hematopoietic stem cells. Cell *62*, 863-874.
Jenkins, M.K., Chu, H.H., McLachlan, J.B., and Moon, J.J. (2010). On the composition of the preimmune repertoire of T cells specific for Peptide-major histocompatibility complex ligands. Annu Rev Immunol *28*, 275-294.
Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G.*, et al.* (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327-339.
Joshi, N.S., Cui, W., Chandele, A., Lee, H.K., Urso, D.R., Hagman, J., Gapin, L., and Kaech, S.M. (2007). Inflammation directs memory precursor and short-lived effector CD8(+) T cell fates via the graded expression of T-bet transcription factor. Immunity *27*, 281-295.
Kaech, S.M., and Cui, W. (2012). Transcriptional control of effector and memory CD8+ T cell differentiation. Nat Rev Immunol *12*, 749-761.
Kaech, S.M., Wherry, E.J., and Ahmed, R. (2002). Effector and memory T-cell differentiation: implications for vaccine development. Nat Rev Immunol *2*, 251-262.
Kantor, A.B., Stall, A.M., Adams, S., Herzenberg, L.A., and Herzenberg, L.A. (1992). Differential development of progenitor activity for three B-cell lineages. Proc Natl Acad Sci U S A *89*, 3320-3324.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G.*, et al.* (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res *46*, D260-D266.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Kim, I., Saunders, T.L., and Morrison, S.J. (2007). Sox17 dependence distinguishes the transcriptional regulation of fetal from adult hematopoietic stem cells. Cell *130*, 470-483.

Koues, O.I., Collins, P.L., Cella, M., Robinette, M.L., Porter, S.I., Pyfrom, S.C., Payton, J.E., Colonna, M., and Oltz, E.M. (2016). Distinct Gene Regulatory Pathways for Human Innate versus Adaptive Lymphoid Cells. Cell *165*, 1134-1146.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann Appl Stat *5*, 1752-1779.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Luckey, C.J., Bhattacharya, D., Goldrath, A.W., Weissman, I.L., Benoist, C., and Mathis, D. (2006). Memory T and memory B cells share a transcriptional program of self-renewal with long-term hematopoietic stem cells. Proc Natl Acad Sci U S A *103*, 3304-3309.

Mackay, L.K., Rahimpour, A., Ma, J.Z., Collins, N., Stock, A.T., Hafon, M.L., Vega-Ramos, J., Lauzurica, P., Mueller, S.N., Stefanovic, T.*, et al.* (2013). The developmental pathway for CD103(+)CD8+ tissue-resident memory T cells of skin. Nat Immunol *14*, 1294-1301.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448-3449.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 *17*.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-110.

Messaoudi, I., Guevara Patino, J.A., Dyall, R., LeMaoult, J., and Nikolich-Zugich, J. (2002). Direct link between mhc polymorphism, T cell avidity, and diversity in immune defense. Science *298*, 1797-1800.

Miao, T., Symonds, A.L.J., Singh, R., Symonds, J.D., Ogbe, A., Omodho, B., Zhu, B., Li, S., and Wang, P. (2017). Egr2 and 3 control adaptive immune responses by temporally uncoupling expansion from T cell differentiation. J Exp Med *214*, 1787-1808.

Miller, R.A. (1996). The aging immune system: primer and prospectus. Science *273*, 70-74.

Min, B., McHugh, R., Sempowski, G.D., Mackall, C., Foucras, G., and Paul, W.E. (2003). Neonates support lymphopenia-induced proliferation. Immunity *18*, 131-140.
Mold, J.E., and McCune, J.M. (2011). At the crossroads between tolerance and aggression: Revisiting the "layered immune system" hypothesis. Chimerism *2*, 35-41.
Mold, J.E., Venkatasubrahmanyam, S., Burt, T.D., Michaelsson, J., Rivera, J.M., Galkina, S.A., Weinberg, K., Stoddart, C.A., and McCune, J.M. (2010). Fetal and adult hematopoietic stem cells give rise to distinct T cell lineages in humans. Science *330*, 1695-1699.
Moskowitz, D.M., Zhang, D.W., Hu, B., Le Saux, S., Yanes, R.E., Ye, Z., Buenrostro, J.D., Weyand, C.M., Greenleaf, W.J., and Goronzy, J.J. (2017). Epigenomics of human CD8 T cell differentiation and aging. Sci Immunol *2*.
Mueller, S.N., Heath, W., McLain, J.D., Carbone, F.R., and Jones, C.M. (2002). Characterization of two TCR transgenic mouse lines specific for herpes simplex virus. Immunol Cell Biol *80*, 156-163.
Nelson, R.W., Rajpal, M.N., and Jenkins, M.K. (2015). The Neonatal CD4+ T Cell Response to a Single Epitope Varies in Genetically Identical Mice. J Immunol *195*, 2115-2121.
Olson, J.A., McDonald-Hyman, C., Jameson, S.C., and Hamilton, S.E. (2013). Effector-like CD8(+) T cells in the memory population mediate potent protective immunity. Immunity *38*, 1250-1260.
Orr, M.T., Orgun, N.N., Wilson, C.B., and Way, S.S. (2007). Cutting edge: recombinant Listeria monocytogenes expressing a single immune-dominant peptide confers protective immunity to herpes simplex virus-1 infection. J Immunol *178*, 4731-4735.
Plumlee, C.R., Sheridan, B.S., Cicek, B.B., and Lefrancois, L. (2013). Environmental cues dictate the fate of individual CD8+ T cells responding to infection. Immunity *39*, 347-356.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
Reynaldi, A., Smith, N.L., Schlub, T.E., Venturi, V., Rudd, B.D., and Davenport, M.P. (2016). Modeling the dynamics of neonatal CD8(+) T-cell responses. Immunol Cell Biol *94*, 838-848.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Rudd, B.D., Venturi, V., Davenport, M.P., and Nikolich-Zugich, J. (2011). Evolution of the antigen-specific CD8+ TCR repertoire across the life span: evidence for clonal homogenization of the old TCR repertoire. J Immunol *186*, 2056-2064.
Sarkar, S., Kalia, V., Haining, W.N., Konieczny, B.T., Subramaniam, S., and Ahmed, R. (2008). Functional and genomic profiling of effector CD8 T cell subsets with distinct memory fates. J Exp Med *205*, 625-640.
Scharer, C.D., Bally, A.P., Gandham, B., and Boss, J.M. (2017). Cutting Edge: Chromatin Accessibility Programs CD8 T Cell Memory. J Immunol *198*, 2238-2243.
Schlub, T.E., Badovinac, V.P., Sabel, J.T., Harty, J.T., and Davenport, M.P. (2010). Predicting CD62L expression during the CD8+ T-cell response in vivo. Immunol Cell Biol *88*, 157-164.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Shih, H.Y., Sciume, G., Mikami, Y., Guo, L., Sun, H.W., Brooks, S.R., Urban, J.F., Jr., Davis, F.P., Kanno, Y., and O'Shea, J.J. (2016). Developmental Acquisition of Regulomes Underlies Innate Lymphoid Cell Functionality. Cell *165*, 1120-1133.

Smith, N.L., Wissink, E., Wang, J., Pinello, J.F., Davenport, M.P., Grimson, A., and Rudd, B.D. (2014). Rapid proliferation and differentiation impairs the development of memory CD8+ T cells in early life. J Immunol *193*, 177-184.

Stemberger, C., Huster, K.M., Koffler, M., Anderl, F., Schiemann, M., Wagner, H., and Busch, D.H. (2007). A single naive CD8+ T cell precursor can develop into diverse effector and memory subsets. Immunity *27*, 985-997.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A *100*, 9440-9445.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Teixeiro, E., Daniels, M.A., Hamilton, S.E., Schrum, A.G., Bragado, R., Jameson, S.C., and Palmer, E. (2009). Different T cell receptor signals determine CD8+ memory versus effector development. Science *323*, 502-505.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.

Wang, J., Wissink, E.M., Watson, N.B., Smith, N.L., Grimson, A., and Rudd, B.D. (2016). Fetal and adult progenitors give rise to unique populations of CD8+ T cells. Blood *128*, 3073-3082.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K.*, et al.* (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431-1443.

White, J.T., Cross, E.W., Burchill, M.A., Danhorn, T., McCarter, M.D., Rosen, H.R., O'Connor, B., and Kedl, R.M. (2016). Virtual memory T cells develop and mediate bystander protective immunity in an IL-15-dependent manner. Nature communications *7*, 11291.

White, J.T., Cross, E.W., and Kedl, R.M. (2017). Antigen-inexperienced memory CD8(+) T cells: where they come from and why we need them. Nat Rev Immunol *17*, 391-400.

Williams, M.A., and Bevan, M.J. (2007). Effector and memory CTL differentiation. Annu Rev Immunol *25*, 171-192.

Williams, M.A., Tyznik, A.J., and Bevan, M.J. (2006). Interleukin-2 signals during priming are required for secondary expansion of CD8+ memory T cells. Nature *441*, 890-893.

Wissink, E.M., Smith, N.L., Spektor, R., Rudd, B.D., and Grimson, A. (2015). MicroRNAs and Their Targets Are Differentially Regulated in Adult and Neonatal Mouse CD8+ T Cells. Genetics *201*, 1017-1030.

Yu, B., Zhang, K., Milner, J.J., Toma, C., Chen, R., Scott-Browne, J.P., Pereira, R.M., Crotty, S., Chang, J.T., Pipkin, M.E.*, et al.* (2017). Epigenetic landscapes reveal transcription factors that regulate CD8(+) T cell differentiation. Nat Immunol *18*, 573-582.

Zaghouani, H., Hoeman, C.M., and Adkins, B. (2009). Neonatal immunity: faulty T-helpers and the shortcomings of dendritic cells. Trends Immunol *30*, 585-591.

Zens, K.D., Chen, J.K., Guyer, R.S., Wu, F.L., Cvetkovski, F., Miron, M., and Farber, D.L. (2017). Reduced generation of lung tissue-resident memory T cells during infancy. J Exp Med *214*, 2915-2932.

Zhang, F., Meng, G., and Strober, W. (2008a). Interactions among the transcription factors Runx1, RORgammat and Foxp3 regulate the differentiation of interleukin 17-producing T cells. Nat Immunol *9*, 1297-1306.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008b). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

APPENDIX II

CDK2 KINASE ACTIVITY IS A REGULATOR OF MALE GERM CELL FATE.

ABSTRACT

The ability of men to remain fertile throughout their lives depends upon establishment of a spermatogonial stem cell (SSC) pool from gonocyte progenitors, and thereafter balancing SSC renewal vs terminal differentiation. Here, we report that precise regulation of the cell cycle is crucial for this balance. Whereas cyclin-dependent kinase 2 (Cdk2) is unnecessary for mouse viability or gametogenesis stages prior to meiotic prophase I, mice bearing a deregulated allele (Cdk2Y15S) are severely deficient in spermatogonial differentiation. This allele disrupts an inhibitory phosphorylation site (Tyr15) for the kinase WEE1. Remarkably, Cdk2Y15S/Y15S mice possess abnormal clusters of mitotically active SSC-like cells, but they are eventually removed by apoptosis after failing to differentiate properly. Analyses of lineage markers, germ cell proliferation over time, and single cell RNA-seq data revealed delayed and defective differentiation of gonocytes into SSCs. Biochemical and genetic data demonstrated that Cdk2Y15S is a gain-of-function allele causing elevated kinase activity, which underlies these differentiation defects. Our results demonstrate that precise regulation of CDK2 kinase activity in male germ cell development is critical for the gonocyte-to-spermatogonial transition and long-term spermatogenic homeostasis.

Key Words: spermatogonia; gonocytes; mouse; cell cycle

---

[1]This appendix is adapted from a manuscript of the same name published in 2019 in *Development* (https://dev.biologists.org/content/146/21/dev180273). This work was performed in collaboration with Priti Singh in the Schimenti lab. The authors of this manuscript were Priti Singh, Ravi K. Patel, Nathan Palmer, Jennifer K. Grenier, Darius Paduch, Philipp Kaldis, Andrew Grimson and John C. Schimenti. Ravi K Patel analyzed the single-cell RNA sequencing data, performed bioinformatic analysis related to scRNA-seq, interpreted data and prepared figures. Priti Singh performed most of the experiments including microscopy, cell sorting experiments and mouse work, analyzed and interpreted data and prepared the manuscript and figures. Detailed description of author contributions can be found below.

INTRODUCTION

Human and mouse males are capable of reproduction throughout much of their lives due to a continuously regenerating pool of spermatogonial stem cells (SSCs) in adult testes. In mice, the progenitors of SSCs, called primordial germ cells (PGCs), arise as a group of ~45 cells in the epiblast of 6-6.5 day old embryos (Ginsburg et al., 1990). The PGCs migrate to the genital ridges by ~embryonic day (E) 10.5, and proliferate rapidly as the gonads differentiate into either primitive ovaries or testes. These germ cells, now called gonocytes (also called "prospermatogonia" in males), reach a population of ~25,000 by E13.5. Then, female gonocytes ("oocytes") directly enter meiosis, while the male gonocytes largely cease proliferation for the remainder of gestation (Sasaki and Matsui, 2008; Tam and Snow, 1981).

About one day after birth at postnatal (P) day 1.5, the gonocytes resume proliferation (Nagano et al., 2000) and complete a process known as the gonocyte-to-spermatogonia transition (GST). The GST is not precisely delineated, as there is no clear-cut distinction between gonocytes/prospermatogonia and spermatogonia, although marker analysis and single cell sequencing indicate that this transition begins in late gestation (Law et al., 2019; Pui and Saga, 2017). During the GST, some gonocytes establish the permanent pool of SSCs that will seed waves of spermatogenesis throughout adult life, while others initiate a distinct prepubertal round of spermatogenesis involving several mitotic spermatogonial divisions, meiosis, and postmeiotic development (Bellve et al., 1977). This first round of spermatogonial differentiation is unique because it originates from neurogenin 3 (NGN3)-expressing gonocytes rather than SSC populations (Yoshida et al., 2004; Yoshida et al., 2006).

During adult life, new coordinated waves of spermatogenesis derive from divisions of an individual A_{single} (A_s) type SSC and its progeny to form A_{paired}, (A_{pr}), and then $A_{aligned}$ (A_{al}) type SSCs. A_{al} SSCs divide without cytokinesis to form clonal chains (A_{al4} to A_{al32}) linked via their cytoplasm. Further divisions result in A2, A3, A4, Intermediate (In) and B type spermatogonia before ultimately differentiating into preleptotene spermatocytes that enter meiosis. The "stemness" potential of these cells decreases as spermatogonia chain length increases [reviewed in (de Rooij, 2017)]. The balance between SSC renewal versus differentiation is crucial, and failure to do so can cause either insufficient sperm production or exhaustion of the stem cell pool. Genetic or environmental events that lead to SSC exhaustion, or which compromise the generation or viability of SSCs or their progenitors, can cause SCOS (Sertoli Cell Only Syndrome), a histological phenotype categorizing a subset of patients with non-obstructive azoospermia (NOA).

Normal proliferation of cells is dependent on cell cycle regulation. Key players in this process are cyclin-dependent kinases (CDKs) and their activating partner proteins, cyclins, several of which exist in mammals. CDK activity is controlled during the cell cycle in part by the association of CDKs with cyclins. In their activated state, these complexes propel cells through successive stages of the cell cycle, including entry into and through the S and M phases (Satyanarayana and Kaldis, 2009). The transition between active and inactive states is governed by both interaction with CDK-inhibitory proteins (Lim and Kaldis, 2013) and also phosphorylation or dephosphorylation events at key regulatory sites on CDKs (Cuijpers and Vertegaal, 2018; Morgan, 1995).

Though the activities of cyclins and CDKs have been studied predominantly in cultured somatic cells and single-celled eukaryotes, their roles in the germline have also been investigated (Martinerie et al., 2014; Wolgemuth and Roberts, 2010). Despite its broad expression in many cell types, *Cdk2* is not essential for mouse viability, yet its disruption causes male and female infertility (Berthet et al., 2003; Ortega et al., 2003). Specifically, *Cdk2*$^{-/-}$ meiocytes arrest during the pachytene stage of meiotic prophase I. This arrest is triggered by defective attachment of telomeres to the nuclear envelope, resulting in failed or incomplete synapsis of homologous chromosomes. In turn, these defects prevent homologous recombination repair of meiotic double strand breaks (Viera et al., 2009; Viera et al., 2015). CDK2 is expressed in spermatogonia (Johnston et al., 2008; Ravnik and Wolgemuth, 1999), but SSCs apparently remain functional because mutant males produce spermatocytes (albeit destined for meiotic arrest) into adulthood. These results suggest that, as in most somatic cells, CDK2 function is not essential in spermatogonia, but it may provide redundant function in those cells and non-canonical function(s) in meiocytes related to recombination in the latter (Berthet et al., 2003; Krasinska et al., 2008). Although a spermatogonia-specific deletion of *Cdk1* has yet to be described, this kinase is required for metaphase I entry at the end of the first meiotic prophase (Clement et al., 2015). CDK1 likely acts in concert with the meiosis-specific cyclin A1, which is also required at the same stage (Liu et al., 1998). In contrast, conditional ablation of cyclinB1 (*Ccnb1*), a CDK1 binding partner, blocks proliferation of gonocytes and spermatogonia, but doesn't impact meiosis (Tang et al., 2017).

To identify infertility alleles in human populations, we modeled a missense variant (SNP rs3087335) altering the TYR15 phosphorylation site of CDK2 in mice (Singh and Schimenti, 2015). Surprisingly, homozygotes for this allele (*Cdk2*^{Y15S}) caused an SCOS-like phenotype. Additionally, *Cdk2*^{Y15S} heterozygotes exhibited age-dependent testis histopathology and reduced sperm, indicating that *Cdk2*^{Y15S} is a gain-of-function, semidominant, allele (Singh and Schimenti, 2015). *In vitro* studies have shown that Tyr15 phosphorylation, typically catalyzed by the WEE1 kinase, negatively regulates CDK activity and thus, cell cycle progression (Gu et al., 1992; Welburn et al., 2007). We speculated that the *Cdk2*^{Y15S} allele was hyperactive by virtue of being refractory to negative regulation by WEE1 (Hughes et al., 2013; Zhao et al., 2012), thus driving excessive spermatogonial proliferation and/or differentiation over SSC regeneration and maintenance.

Here, we report that the apparent SCOS phenotype in *Cdk2*$^{Y15S/Y15S}$ testes is not due to an absence of germ cells; rather, SSC-like cells are present and can divide, but their progeny fail to differentiate and subsequently are lost before entering meiosis. The germ cell defects are first detectable at postnatal day 3, where the GST appears delayed or disrupted as determined by analyses of key markers and single cell (sc) RNA-seq data. We provide evidence that CDK2^{Y15S}-expressing cells display altered kinase activity, and that this defect underlies phenotypes observed in such cells. This study highlights the importance of precise regulation of CDK kinase activity in establishing and maintaining testis homeostasis.

RESULTS

Ablation of the TYR15 inhibitory phosphorylation site in CDK2 disrupts gonocyte and spermatogonia differentiation.

As summarized above, adult *Cdk2*$^{Y15S/Y15S}$ testes lacked evidence of spermatogenesis and were essentially devoid of cells positive for DDX4 (hereafter MVH, mouse vasa homolog), which is strongly expressed in gonocytes and all juvenile germ cells (Toyooka et al., 2000). Our working hypothesis was most gonocytes differentiated in the initial spermatogenic wave, leaving the adults devoid of a renewable SSC pool. To test this hypothesis, we first quantified gonocytes in neonatal testes. The number of MVH$^+$ cells in P0 *Cdk2*$^{Y15S/Y15S}$ testes was no different than in control littermates (Fig. 1A-B), indicating that the loss of germ cells occurred not during gestation (for example, during PGC expansion), but during postnatal development. Next, to test the prediction that all SSCs would be exhausted by adulthood, we performed immunohistochemical (IHC) analysis of mutant adult (P180) seminiferous tubule sections, which lack ongoing spermatogenesis. Remarkably, *Cdk2*$^{Y15S/Y15S}$ tubules contained ample numbers of cells positive for LIN28, which is expressed in a subset of Type A_s spermatogonia, and essentially all Type A_{pr} through A_{al} spermatogonia (Chakraborty et al., 2014b) (Fig. 1C; Fig S1B,C), demonstrating that although mutant testes had an SCOS-like appearance, there were indeed undifferentiated A-type spermatogonia present. However, they apparently were not proliferating or differentiating in a normal manner.

We next characterized maturation of germ cells using markers of gonocytes, SSCs, and progressively more differentiated spermatogonia. The transcription factor FOXO1 (Forkhead box O1), which in testis is expressed only in gonocytes and undifferentiated spermatogonia (A_s – A_{al}) but not more differentiated germ cells, transits from the cytoplasm to the nucleus as gonocytes differentiate into spermatogonia postnatally (Goertz et al., 2011). At P0, mutant gonads retained cytoplasmic localization of FOXO1 (Fig. S1A), consistent with normal prenatal development of gonocytes. However, translocation of FOXO1 to the nucleus, which normally begins at P3 and is complete by P21 (Goertz et al., 2011), was delayed in the mutant. Unlike WT, in which FOXO1 was localized in the nuclei of all positive cells at P30, >40% of *Cdk2*$^{Y15S/Y15S}$ cells exhibited cytoplasmic FOXO1 at this time; this fraction declined further to ~10% at P90 (Fig. 2A-B; Fig. S1B). These results suggest that *Cdk2*$^{Y15S/Y15S}$ germ cells are delayed in the GST and, with the exception of the unique first round of spermatogenesis, are unable to differentiate properly even after making this transition.

Studies of other markers by testis IHC confirmed abnormalities in the *Cdk2*$^{Y15S/Y15S}$ germ cell pool. At P3 and P15, there were fewer cells positive for PLZF (zinc finger and BTB domain containing 16, formally ZBTB16) and FOXO1 (Fig. S1B, C), which are expressed in all undifferentiated spermatogonia (Goertz et al., 2011). The number of cells positive for LIN28, which is expressed in both undifferentiated (A_s – A_{al}) and differentiated (A_1-A_4) spermatogonia (Chakraborty et al., 2014a; Gaytan et al., 2013), was also lower in mutants at P3 (Fig. S1C). The LIN28$^+$ cell shortfall disappeared at P15, possibly reflecting large numbers of differentiated spermatogonia resulting from the first wave of spermatogenesis, but in adulthood (P90) when the absence of ongoing spermatogenesis in mutants was manifested, LIN28$^+$ cells were again lower in the mutant, as was PLZF (Fig. S1B, C).

We next performed a series of studies on seminiferous tubule whole mounts to determine if there were disruptions to the normal patterns of Type A spermatogonia subtypes (e.g., A_s vs A_{pr} vs A_{al}). A marker for SSCs is GFRA1 (glial cell line derived neurotrophic factor family receptor alpha 1, the cell surface receptor for GDNF), which mainly labels A_s (a subset thereof) and A_{pr} spermatogonia. We found that almost all A_s and A_{pr} PLZF$^+$ SSCs from P90 *Cdk2*$^{Y15S/Y15S}$ mice also expressed GFRA1, similar to age-matched WT controls. Mutants also contained clusters of 4-8 PLZF$^+$LIN28$^+$ cells at P90, although these clusters did not have the typical appearance of A_{al} chains (possibly due to disrupted overall tubule architecture; henceforth we will refer groups of >2 cells as "clusters"). Furthermore, mutants lacked longer PLZF$^+$LIN28$^+$ chains ($A_{al16-32}$) that are typical in adult WT testes (Fig. 3A) (Buaas et al., 2004; Zheng et al., 2009). This result is consistent with the lower overall number of PLZF$^+$ and LIN28$^+$ spermatogonia in testes of young (P5) and old (P90) *Cdk2*^{Y15S} homozygotes (Fig. S1C).

To explore the basis for the defects in spermatogonial distributions, we examined GFRA1$^+$ progenitor spermatogonia in adolescent (P15, during the first round of spermatogenesis) vs adult (P90) tubules that lack spermatogonial differentiation. *Cdk2*$^{Y15S/Y15S}$ mutants had about twice as many GFRA1$^+$ cells at both ages compared to WT (Fig. 3A-C). Strikingly, mutant tubules contained large clusters of GFRA1$^+$ spermatogonia (referred to as 'GFRA1$^+$ $A_{cluster>6}$') in P15 tubules, which were virtually absent in WT (Fig. 3A,D). Hypothesizing that these GFRA1$^+$ $A_{cluster>6}$ cells might be in an abnormal, delayed state of differentiation, we examined them for co-expression of LIN28. Interestingly, many of the P15 GFRA1$^+$ cells in clusters were also LIN28$^+$ (Fig. S2). Additionally, whereas mutants had ~2-fold more A_s GFRa1$^+$ cells than WT, most were also LIN28$^+$ as in WT (Fig. 3F). The clusters of GFRA1$^+$ cells disappeared by P90, yet mutant tubules had more GFRA1$^+$ A_s and A_{pr} cells compared to WT at this age (Fig. 3B-D). The combined data indicate that the *Cdk2*^{Y15S} allele not only causes a delay in the GST, but also impacts the differentiation of progenitor SSCs beyond the A_{pr} developmental state when GFRa1 should be downregulated in chains of LIN28$^+$ A_{al} spermatogonia.

Regulation of CDK2 activity is critical for balancing spermatogonia progenitor self-renewal vs differentiation

Whereas SSC differentiation or maintenance was not obviously impacted in *Cdk2* null (*Cdk2*$^{-/-}$) mice, which demonstrate continued rounds of meiotic entry (Berthet et al., 2003; Ortega et al., 2003) (see also last section of Results and Fig. 7D), *Cdk2*^{Y15S} heterozygotes exhibited an age-related decrease in spermatogenesis. This difference in phenotypes suggests that *Cdk2*^{Y15S} is a hypermorphic or a gain-of-function allele. Given the increase in GFRa1$^+$ spermatogonia in the *Cdk2*$^{Y15S/Y15S}$ testes (Fig. 3), we hypothesized that *Cdk2*^{Y15S} either drives abnormal proliferation, and/or it skews these cells towards self-renewal instead of differentiation.

To test this, we assayed proliferation of spermatogonia by pulse labelling with the DNA analog EdU. P2 males were injected with EdU, sacrificed 4 hours later, then the testes were immunolabeled for PLZF. In both mutants and WT, >99% of PLZF$^+$ cells were negative for EdU (not shown), consistent with this being the period before gonocytes exit mitotic arrest to begin establishing the spermatogonial pool (Yang and Oatley, 2014). However, in

older mutant animals, we noticed severe proliferation abnormalities in GFRa1$^+$ and FOXO1$^+$ cells. At P15, there were ~1.6 fold more proliferating (EdU+) GFRa1$^+$ cells in mutant homozygotes than WT, and this disparity persisted through P90 (Fig. 4A,C). Furthermore, the fractions of replicating A_s and A_{pr} GFRa1$^+$ spermatogonia were higher in mutant than WT, and most dramatically, the $A_{cluster>6}$ GFRa1$^+$ category was unique to the mutant (Fig. 4A, D). Similarly, P90 *Cdk2*$^{Y15S/Y15S}$ testes, which contained cells with both nuclear (predominantly) and cytoplasmic FOXO1 (nFOXO1$^+$ and cFOXO1$^+$, respectively; Fig. 4B), had more proliferating FOXO1$^+$ cells in both categories (Fig. 4E), in aggregate representing ~50% more in the mutant than WT (22.5% vs 15%, p=0.01). In contrast, P90 mutant testes had ~40% fewer proliferating LIN28$^+$ germ cells, which include A_s spermatogonia (Fig. 4F). Collectively, our results support the notion that normal homeostasis of the stem cell niche is disrupted in *Cdk2*$^{Y15S/Y15S}$ mice, causing elevated proliferation of gonocytes and SSCs without normal differentiation.

Despite the evidence for spermatogonial cycling in the absence of differentiation, the seminiferous tubules never (up to 20 months of age) became replete with undifferentiated cells, a condition that might lead to, or resemble, tumorigenesis. Therefore, we hypothesized that such hyperproliferating progenitor spermatogonia were eliminated by apoptosis. Indeed, there were nearly 6-fold more TUNEL$^+$ seminiferous tubules containing 2-4 fold more apoptotic germ cells in mutants vs WT at P30 (Fig. S3A-C). Furthermore, double immunolabeling for FOXO1 and cleaved PARP revealed the presence of spermatogonia undergoing apoptosis in mutants (Fig. S3D). These combined results suggest that *Cdk2*$^{Y15S/Y15S}$ undifferentiated progenitor GFRa1$^+$ spermatogonia enter S phase normally but were incapable of differentiating; instead, they appear to proliferate, accumulate, and eventually undergo apoptosis.

Single cell transcriptome analysis reveals defects in differentiation of Cdk2$^{Y15S/Y15S}$ gonocytes and SSCs

The germ cell population at birth has substantial functional and molecular heterogeneity (Culty, 2013). Gonocytes in WT P3 testes, which constitute about half of all germ cells (Ohmura et al., 2004), can undergo one of three immediate fates: 1) re-enter the cell cycle, 2) migrate towards the basement membrane of the seminiferous cords to become SSCs, or 3) differentiate into spermatogonia that seed the first wave of spermatogenesis. Gonocytes decrease from 98% to 30% of all germ cells during the first week of life (Ohmura et al., 2004), thus lying within the GST interval [although the timing may differ in Swiss outbred mice (Pui and Saga, 2017)]. Given that *Cdk2*^{Y15S} mutants are born with a normal number of germ cells (Fig 1A), and that the first apparent abnormality was a deficit of PLZF$^+$/ LIN28$^+$/ FOXO1$^+$ spermatogonia at P3 (Fig. S1C), we hypothesized that the neonatal germ cell pool was defective in differentiating into SSCs.

To test this at a molecular level, we performed scRNA-seq on unsorted cells from WT and mutant P3 testes. Data were obtained from WT (5,061 cells), *Cdk2*$^{Y15S/+}$ (4,958 cells), and *Cdk2*$^{Y15S/Y15S}$ (4,403 cells) testes. There was a median of 2,239 genes and 5,751 mRNA molecules detected per cell. Somatic and germ cells were evident in the clustering analysis of 14,422 cells from testes across all genotypes. Following unbiased k-means clustering of cells based on gene-expression differences, we identified 5 major cell clusters (Fig. 5A). The cells

types within each cluster were identified by markers diagnostic of particular gonadal lineages (Fig. S4). The percentages and numbers of cells in each of the 5 clusters were as follows: MVH$^+$ germ cells (2.06%, n=299); WT1$^+$ Sertoli cells (46.88%, n=6,781); CYP11a1$^+$ Leydig cells (1.5%, n=217); MYH11$^+$ myoid cells (22.15%, n=3,204); and VCAM1$^+$ peritubular/epithelial cells (27.9%, n=3,963).

To characterize potential GST defects, we focused on the MVH$^+$ germ cell population. Interestingly, this group of cells consisted apparently of two populations that differed dramatically (> 2-fold) in the number of unique transcripts (UMIs) per cell (Fig. S5A). This broad bimodal distribution was not observed in somatic cell populations. We considered two potential explanations for this observation. One is that the population with higher UMIs represents cell doublets (Ziegenhain et al., 2017). However, two algorithms used to predict doublets from single-cell expression data, DoubletFinder (Fig. S5B) (McGinnis et al., 2018) and DoubletDetection (data not shown) (https://github.com/JonathanShor/DoubletDetection), indicated that the cell barcodes with higher UMIs (>10,000) were no more likely to be doublets than those with lower UMIs. A second explanation is that neonatal germ cells are transcriptionally more active than somatic cells, as are embryonic germ and stem cells (Percharde et al., 2017). To distinguish between these hypotheses, we quantified total RNA from equal numbers of FACS-isolated *Oct4*-GFP+ cells from neonatal testes of transgenic mice (Szabó et al., 2002). OCT4 is a pluripotency marker and its expression is restricted to undifferentiated, prepubertal gonocytes/prospermatogonia (Ohbo et al., 2003; Szabó et al., 2002). This revealed 7-fold, 8.3-fold, and 2.4-fold higher amounts of RNA in GFP$^+$ germ cells compared to lung, brain, and GFP$^-$ testis cells, respectively (Fig. S5C, D), supporting the hypothesis that neonatal germ cells have more transcripts. We next performed knee-plot analysis to identify a UMI cutoff for reliably profiled cells, revealing that the barcodes with low UMI counts fell in a low-quality region (under "knee"), indicating that those cells had poor detection (Fig. S5E) (Macosko et al., 2015). Therefore, we removed barcodes with low (≤ 10,000) UMI counts. After quality control and data filtering using Seurat (Butler et al., 2018), we identified 10,451 confidently quantified transcripts from WT and *Cdk2*$^{Y15S/Y15S}$ germ cells. Since *Cdk2*$^{Y15S/+}$ clusters closely overlapped with those from WT, we only used the latter in subsequent comparisons to homozygous mutants.

Since only a fraction of all transcripts are detected in scRNA-seq, the resulting expression matrices are sparse. We therefore employed SAVER (single-cell analysis via expression recovery) (Huang et al., 2018), which uses information across all genes and cells in a dataset, to impute gene expression values. We identified five subgroups of MVH$^+$ germ cells, designated A-E (Fig. 5B), using hierarchical clustering based on the SAVER-imputed expression of the most divergently expressed genes (n=734; see methods) (Fig. S6). As depicted in Fig. 5B, cluster A was exclusive to the WT sample, whereas D and E were exclusive to the *Cdk2*$^{Y15S/Y15S}$ sample. Clusters B and C were preferentially enriched in WT and the *Cdk2*$^{Y15S/Y15S}$ samples, respectively.

Clusters A and B express several well-characterized markers of undifferentiated germ cells, such as *Gfra1* and *Id4* (Fig. 5C). Based on the following distinctions between the two groups, we tentatively classified cluster A as SSCs (A_{SSC}) and cluster B as gonocytes (B_{Gono}). Cluster A was relatively depleted in cell cycle factors such as *Ccnb1* and *Cenpf* (Fig. 5C),

consistent with the idea that 'true' SSCs would have a lower proliferative index. Furthermore, this cluster was enriched for *Txnip* (inhibitor of glucose transport), indicating decreased metabolism consistent with lower proliferation**.** Finally, to validate the cluster identities, we combined previously published scRNA-seq datasets from flow sorted $OCT4^{+}$ (Liao et al., 2017) and $ID4^{+}$ cells (Song et al., 2016), and identified 50 and 100 genes uniquely expressed in gonocytes and SSCs, respectively, which we then used as diagnostic markers for cell identities (Fig. S7A,B; Table S1). GSEA analyses revealed that highly expressed genes in the SSC genesets were up-regulated in cluster A compared to the reference cluster B, whereas highly expressed genes in gonocytes showed up-regulation in cluster B (Fig. S7B). The combined data led us to conclude that cluster A consists of SSCs, and cluster B consists of gonocytes.

To better define the cellular defects in mutant germ cells present at P3, we compared key expression patterns of clusters A and D/E, which are the populations most specific to WT and *Cdk2*$^{Y15S/Y15S}$, respectively. The following features were characteristic of D/E: 1) cell cycle signature genes and E2F targets were enriched (Figs. 5C,F; 6F); 2) SSC genes were expressed at relatively low levels (Fig. 5C); and 3) the PGC/gonocyte marker NANOS3 and differentiating spermatogonia signatures (*Stra8, Lmo1, Uchl1, Dmrt1, Sohlh1, Dnmt3b*) were coexpressed and enriched (Fig. 5C,F). In sum, mutant germ cells express an unusual combination of cell cycle, differentiation and gonocyte signatures. Importantly, when we repeated these analyses using gene expression data without SAVER-based imputation, our results were consistent (Fig. S8).

We next performed pseudo-time analysis (Trapnell et al., 2014) on the germ cell cluster transcriptomes to explore the implications for developmental states and trajectories. In WT, this analysis supported a trajectory path bifurcating from cluster B (WT_{gono}) to clusters A (WT_{SSCs}) and C (differentiation-primed gonocytes, or $WT_{diff\text{-}Gono}$) (Fig. 5D), with the latter being defined by virtue of retaining both gonocyte and differentiation signatures (Fig. 5C,D). This trajectory path in WT is consistent with the known developmental progression in the germline. Interestingly, only ~8% of mutant germ cells fell into Cluster B, and these eventually differentiate into two directions: 1) a small subset towards cluster C ($YS_{diff\text{-}gono;}$ ~12%) and 2) ~80% towards mutant-specific clusters D+E (Fig. 5D). It is worth noting that trajectory analyses using non-imputed data gave a different topology. Cluster C, instead of branching out as a separate cluster, appeared as a transient stage on a developmental trajectory differentiating from cluster B to E (Fig. S7C,D). Nevertheless, the analyses indicate a profound defect in differentiation of mutant germ cells, and are consistent with the idea that mutant gonocytes do not undergo a normal GST.

To gain additional perspective on developmental defects in the mutant, we performed RNA velocity analysis of the scRNA-seq data (Fig. 5E). This method examines the expression dynamics of unspliced (nascent) vs. spliced (mature) versions of transcripts to predict the future developmental states of cells (La Manno et al., 2018). This analysis supports the conclusion that whereas most WT gonocytes will give rise to SSCs, the *Cdk2*$^{Y15S/Y15S}$ mutation causes nearly all mutant gonocytes (including those in clusters B and C; Fig. 5E) to be transitioning to an abnormal gonocyte-like state exemplified by clusters D and E (summarized in Fig. 5F).

CDK2^{Y15S} has altered kinase activity that impacts gonocyte fate

While we hypothesized that CDK2^{Y15S} is hypermorphic by virtue of lacking the target (TYR15) of inhibitory phosphorylation (Singh and Schimenti, 2015), we considered the possibility that SER15 could be phosphorylated by an unknown kinase to alter CDK2 activity. To test this, we expressed MYC-tagged WT (CDK2-TYR15), mutant (SER15) and also PHE15 cDNAs in HEK293T cells, performed mass spectrometry (LC-MS/MS) analysis on immunoprecipitates (Fig. S8A), then analyzed the mass:charge (m/z) spectra for evidence of phosphorylated of these residues. Phosphorylation was detected only at the WT CDK2^{Y15} residue, indicating that serine at this position is not a phosphorylatable substrate, at least in cultured cells (Fig. S9B).

Next, we assayed the ability of CDK2 isolated from WT and mutant P10 spleens to phosphorylate a histone H1 substrate (testis was not used as a source due to cellularity differences between mutant and WT). Consistent with ablation of the TYR15 inhibitory phosphorylation site and previous reports examining *Cdk2^{T14AY15F}* activity in mouse tissues and MEFs (Zhao et al., 2012), CDK2 immunoprecipitated from *Cdk2$^{Y15S/+}$* spleens displayed 1.5-fold more kinase activity than WT (Fig. S10B-D). Counterintuitively, material immunoprecipitated from *Cdk2$^{Y15S/Y15S}$* spleens had >5-fold reduced kinase activity compared to WT. This may reflect the consequence of excessive CDK kinase activity, which can be toxic to cell cycle progression in a mechanism involving p21-mediated inhibition of CDK2/cyclin (see discussion) (Hughes et al., 2013; Szmyd et al., 2019; Zhao et al., 2012).

As an orthogonal assessment of CDK2 activity in germ cells, we compared expression levels of 97 key CDK2 activity signature genes in the cell clusters defined earlier (Table S2) (McCurdy et al., 2017). WT$_{SSCs}$ (cluster A) had much lower expression of CDK2 activity signature genes compared to all other clusters including WT cells in cluster B and all *Cdk2^{Y15S}* clusters (Fig. 6A). Furthermore, CDK2 kinase activity, as inferred by the median of normalized expression of CDK2 activity signature genes per cell across clusters A, B, D and E, was lowest in WT$_{SSCs}$ and highest in mutant-specific clusters D and E (Fig. 6B,C). GSEA analysis indicated up-regulation of CDK2 activity signature genes in clusters B and E compared to cluster A, suggesting that first, cluster E is more like cluster B with respect to CDK2 kinase activity, and second, that cells in cluster E have a higher propensity to cycle than those in cluster A (Fig. 6C). Interestingly, in support of this observation, genes up-regulated in cluster E compared to cluster A ($q \leq 0.05$; $FC \geq 0.2$) also exhibited the enrichment of cell cycle related gene ontology (GO) terms (Fig. 6D). Overall, these data indicate that removing a layer of negative regulation to CDK2 activity disrupts the normal differentiation of gonocytes and SSCs into downstream cell types.

The molecular basis for this developmental disruption may stem from alteration of two known functions of active CDK2: 1) phosphorylating the RB1 transcriptional repressor to inactivate it, enabling timely induction of the E2F transcription factors (TFs) to drive transition to S phase (Morris et al., 2000); and 2) inhibiting cytoplasmic-to-nuclear localization of the FOXO1 TF (nFOXO1 is essential for SSC maintenance) (Huang et al., 2006)(Goertz et al., 2011). If CDK2 is indeed controlling the gene regulatory network via E2F and FOXO1, then levels or activity of their downstream might be affected in *Cdk2$^{Y15S/Y15S}$* cells. Indeed, GSEA analysis revealed that FOXO1 targets are up-regulated and highest in cluster A vs. cluster E, where expression is lowest (84 genes; nES: 3.27; p: 0)

(Fig. 6 E,G,H). In contrast, E2F target genes are up-regulated in cluster E (Fig. 6F; 66 genes; nES: −2.7; p: 0). These results indicate that FOXO1 activity is significantly greater in cluster A than E (Student's t test, $p = 3.7x10^{-19}$) (Fig. 6H).

Phosphorylation states at Tyr15 and Thr160 residues control CDK2 activity in male germ cells.

If disrupting the ability to negatively regulate CDK2 in adult GFRa1$^+$ SSCs favors cell cycle progression over differentiation, then a compensatory alteration that dampens or eliminates CDK2 activity might counteract the aberrant phenotype of *Cdk2*^{Y15S} mutant spermatogonia. To test this, we mutated threonine 160 to alanine (T160A) in the *Cdk2*^{Y15S} allele. Phosphorylation of Thr160 is required for activation of CDK2 (Gu et al., 1992; Kaldis, 1999). *Cdk2*^{T160A} is a "kinase dead" allele that causes infertility in both sexes, albeit with less severe meiotic phenotypes than in *Cdk2*$^{-/-}$ mice (Chauhan et al., 2016). Mice homozygous for the doubly mutated *Cdk2* allele (*Cdk2*^{Y15S,T160A}, abbreviated as *Cdk2*$^{YS\text{-}TA}$) exhibited three striking phenotypic differences compared to *Cdk2*^{Y15S}. First, whereas *Cdk2*^{Y15S} adult (P120) heterozygotes had small testes and a markedly reduced sperm count as previously reported (Singh and Schimenti, 2015), *Cdk2*$^{YS\text{-}TA}$ heterozygotes were indistinguishable from WT in both respects (Fig. 7A-C). Second, as with null but not *Cdk2*$^{Y15S/Y15S}$ mice, *Cdk2*$^{YS\text{-}TA}$ homozygous females were sterile (n=3 in matings to WT). Third, although *Cdk2*$^{YS\text{-}TA/YS\text{-}TA}$ adult males were severely hypogonadal (Fig. 7A-B) and azoospermic, similar to *Cdk2*$^{Y15S/Y15S}$, they exhibited differentiating spermatogonia and meiocytes, which were completely missing from age-matched *Cdk2*^{Y15S} homozygotes, indicating that the T160A alteration rescued the spermatogonial differentiation defect (Fig. 7D). At these levels of analysis, the *Cdk2*$^{YS\text{-}TA}$ allele resembles the null phenotype (Viera et al., 2009)(Chauhan et al., 2016). Collectively, our results imply that the failed spermatogonial differentiation phenotype caused by the *Cdk2*^{Y15S} allele is a result of altered kinase activity from abolition of a WEE1 phosphorylation site. As a consequence of this defective negative regulation, this allele acts semidominantly.

DISCUSSION

The longevity of spermatogenesis in both mice and humans depends upon proper establishment and homeostasis of the SSC pool. A key event in male germline establishment is the seeding of prepubertal testes with a quiescent (G1-arrested) population of gonocytes/prospermatogonia. About 2-3 days after birth, these cells re-enter the cell cycle to expand and differentiate into SSCs that establish a permanent, renewable pool of cells that can initiate spermatogenesis throughout life. Once established, there must be a fine balance of SSC self-renewal vs differentiation to maintain homeostasis and fertility, but this is not well understood and is the subject of intense research.

There is some debate as to the character of SSCs and their behavior with respect to cycling activity (Huckins, 1971b; Sharma et al., 2019). In some tissues, the ability to continuously produce differentiated cells depends upon proper maintenance of a relatively infrequently-cycling population of stem cells, as in the case of the hematopoietic system. Overproliferation of hematopoietic stem cells (HSCs) caused by deregulated cell cycle control can lead to their exhaustion or transformation (Pietras et al., 2011). It has been hypothesized that there is a slow cycling population of SSCs that maintains the germline (Huckins, 1971a; Sharma et al., 2019). However, there is also evidence for rapid turnover of SSCs (Klein et al., 2010) and for the ability of chains of differentiating spermatogonia (A_{pr} and A_{al}) to fragment and de-differentiate to become A_s SSCs (Hara et al., 2014; Nakagawa et al., 2010). Regardless of various models proposed for the identity and behavior of "true" SSCs [reviewed in (de Rooij, 2017)], proper regulation of the cell cycle is essential. This is underscored by our studies, which implicate CDK activity regulation as being crucial for the GST and SSC renewal *vs* differentiation. Indeed, it is well-recognized that in general, cell cycle progression and differentiation occur in a mutually exclusive manner (Dalton, 2015). For example a decrease in CDK activity stimulates differentiation of neural stem cells (Lim and Kaldis, 2012), whereas elevated CDK activity decreases differentiation (Lange et al., 2009) and stimulates expansion of neural stem cells in the adult mouse brain (Artegiani et al., 2011).

Phenotypes of certain mouse mutants have provided some insight into how regulation of cell cycle impacts spermatogonial maintenance and proliferation. Conditional germline knockout of the *Rb* (*Rb1*) tumor suppressor, a negative cell cycle regulator, abolished the ability of SSCs to self-renew, causing the entire germ cell pool to undergo a single round of spermatogenesis (Hu et al., 2013). Cell cycle progression in normal cells requires inactivation of Rb by CDK/cyclin-mediated phosphorylation (including by CDK2/cyclinE), thus allowing expression of genes regulated by E2F transcription factors (Rubin, 2013). Moreover, ablation of *Plzf*, which negatively regulates the cell cycle by both inhibiting key regulators (McConnell et al., 2003; Yeyati et al., 1999) and also the self-renewal signal of GDNF (Hobbs et al., 2010), causes a less severe phenotype than Rb deficiency. *Plzf*$^{/-}$ males are infertile due to a defect in SSC maintenance that leads to progressive germ cell loss and SCOS (Buaas et al., 2004; Costoya et al., 2004). Thus, a cell cycle-centric interpretation of these phenotypes is that unrestrained cycling (as in *Rb*$^{/-}$) causes efficient differentiation of all SSCs, but a moderate loss of cell cycle control (as in *Plzf*$^{/-}$) increases the propensity of SSCs to differentiate rather than self-renew. In the context of this model, CDK2^{Y15S} might have a lower impact on cell cycle control than *Rb* and *Plzf* mutants, possibly mediated by abnormal

but partial inactivation of Rb. The result is abnormal SSC proliferation but only partial differentiation, ultimately leading to loss of the aberrant cells (Clusters D and E) before meiotic entry. The defective differentiation may explain the observation that P90 *Cdk2*$^{Y15S/Y15S}$ mutants had increased proliferating GFRa1^{+} cells but fewer proliferating cells expressing LIN28, which normally marks a subset of A_s "true" SSCs).

MEFs (mouse embryonic fibroblasts) derived from mice in which both the Thr14 and Tyr15 negative phosphoregulatory sites were mutated exhibited accelerated entry into S phase (Zhao et al., 2012). Interestingly, though no data was presented, the only significant defect in these mice was male infertility due to an apparent absence of germ cells in homozygotes and also in heterozygotes in one strain background. In that regard, this model appears to resemble our *Cdk2*^{Y15S} mouse. This might indicate that the Thr14 residue is redundant or not important for negative regulation of SSCs.

Counterintuitively, while we observed elevated CDK2 kinase activity in spleens from heterozygous mice, the activity was greatly reduced in homozygous spleens. However, CDK2 kinase over-activation has been previously shown to be deleterious to cell cycle progression (Hughes et al., 2013). We postulate that, at least in spleen, cell autonomous regulation maintains CDK2 activity within an appropriate range. Mutation of both negative phosphoregulatory sites in CDK2 (Thr14 and Tyr15) in human Hct116 cell lines caused premature S phase progression, DNA damage accumulation, and genomic instability leading to S phase arrest (Hughes et al., 2013). In these cells, degradation of cyclin E was found to be increased and this could be relieved by p21 depletion, suggesting the presence of cellular feedback loops to attempt to reduce the levels of CDK2 activity upon premature activation (Hughes et al., 2013). In consideration of the aforementioned data of others, and our data showing elevated replicative activity and apoptosis of GFRa1^{+} cells in *Cdk2*$^{Y15S/Y15S}$ testes, we propose a model (Fig. 8) in which CDK2-associated kinase activity must be tightly regulated during specific stages of the cell cycle in SSCs, otherwise those cells attempting to differentiate will eventually die from cell cycle dysregulation.

A caveat to the model is that it is based on cell cycle studies in homogeneous cell cultures, as opposed to germ cells developing in a complex milieu that provide a stem cell niche containing several somatic cell types (Hofmann, 2008; Oatley and Brinster, 2012). In particular, the Sertoli cells, which associate intimately with the germ cells, provide key instructive signals including RA to induce differentiation (Endo et al., 2015). Furthermore, along with endothelial cells, Sertoli cells produce germline derived neurotrophic factor (GDNF, which binds GFRa1) that is essential for spermatogonial maintenance (Bhang et al., 2018; Meng et al., 2000). It is possible that the phenotypes of *Cdk2*^{Y15S} mice may not be a manifestation of germ cell-intrinsic defects exclusively, but rather to defects in one or more somatic cell types in addition to those of germ cells. Disruption to individual components of the testis niche could impact the transcriptome and behavior of germ cells, and *vice versa*. While there were differences between mutant and WT Sertoli cells transcriptomes, EdU labeling of P30 mice revealed no evidence for active DNA replication in mutant or WT SOX9^{+} Sertoli cells (data not shown), suggesting that CDK2^{Y15S} affects the cell cycle of germ cells (Fig. 4) rather than Sertoli cells. Ultimately, these issues of potential niche defects can be addressed via germ cell transplantation or cell-type specific ablation experiments.

Whereas there are mutations that reduce the number of perinatal gonocytes (typically stemming from PGC defects) (Hamer and de Rooij, 2018), the *Cdk2*^{Y15S} phenotype is unique in that the GST is defective. Nevertheless, studies of other mutants give insight into how perturbations to cell cycle regulation impact gonocytes. *Cdk2*$^{Y15S/Y15S}$ germ cells fail to relocalize FOXO1 from the cytoplasm to the nucleus, a hallmark of the GST which normally occurs between P3 and P21 (Goertz et al., 2011; Kang et al., 2016; Pui and Saga, 2017). This cytoplasmic localization occurs as a result of CDK2-dependent phosphorylation (inactivation) of FOXO1 (Huang et al., 2006), which is essential for SSC maintenance and differentiation (Goertz et al., 2011). Since FOXO1 nuclear localization is disrupted in *Cdk2*^{Y15S} mutants, we conclude that proper cell cycle regulation and/or CDK2 kinase activity is a pre-requisite for the GST, and thus lies upstream in the developmental evolution/progression. Mice lacking the transcription factor *Glis3* partially resemble *Cdk2*^{Y15S} mutants; they lack FOXO1 nuclear localization in neonatal gonads and fail to undergo a normal GST, as indicated by decreased expression of genes associated with the permanent pool of undifferentiated spermatogonia (Kang et al., 2016). However, it is unknown whether GLIS3 regulates these genes directly or indirectly, so its relationship to CDK2 in the developmental hierarchy is unclear.

Interestingly, the first round of spermatogenesis in *Cdk2*$^{Y15S/Y15S}$ mice occurs normally, although meiosis is still interrupted as in null mice. This first wave of spermatogenesis is claimed to arise directly from a subset of gonocytes that do not express *Ngn3* (Yoshida et al., 2006). We can conclude that differentiation of this subtype of gonocytes neither requires CDK2 nor is affected by its abnormal activity in *Cdk2*^{Y15S} mutants. This underscores the caution needed when extrapolating data from cultured cells or from somatic cells (such as spleen) to germ cells, and even from one germ cell type to another.

These investigations into the *Cdk2*^{Y15S} mouse stemmed from a project to investigate the genetic basis of human infertility, and involves modeling human SNPs in mice to determine possible pathogenicity (Singh and Schimenti, 2015). A lesson learned from this and other alleles is that very few predicted deleterious variants are nulls, and that extensive phenotyping is required to understand their impact. Generally speaking, unless this mutation is entirely unusual in its effects, it raises the possibility that mice and people that present with non-obstructive azoospermia, and which have histopathology superficially resembling SCOS, may in fact have an occult pool of SSCs that have lost the ability to differentiate, but might be stimulated to do so after appropriate intervention. Additionally, as an example of an autosomal semi-dominant male infertility allele, it emphasizes the importance of considering monoallelic alterations when attempting to identify genetic causes of an individual patient's infertility.

MATERIALS and METHODS

Mouse strains and breeding. Mice used were on a mixed genetic background (FVB/NJ and B6(Cg)-Tyr^{c-2J}/J). Experiments with the animals were performed under a protocol (2004-0038) approved by Cornell's Animal Care and Use Committee. For female fertility tests, 8-10 wk-old WT or mutant homozygote females were mated with age-matched B6(Cg)-Tyr^{c-2J}/J males. The *Cdk2*^{tm1Kald}, *Cdk2*^{T160A}, and *Cdk2*^{Y15S} alleles have been described previously (Berthet et al., 2003; Chauhan et al., 2016; Singh and Schimenti, 2015).

Production of CRISPR/Cas9-Edited Mice. The *Cdk2*$^{Y15S\text{-}T160A}$ (formally, *Cdk2*^{em2Jcs}) and *Cdk2*$^{-/-}$ (formally *Cdk2*^{em3Jcs}) alleles were generated using CRISPR/Cas9 genome editing, essentially as described previously (Singh et al., 2015; Varshney et al., 2015). sgRNAs and ssODNs are listed in (Table S1). Briefly, *In vitro* transcription of sgRNA were performed using MEGAshortscript™ T7 Transcription Kit (Ambion; AM1354), and ssODN were availed from IDT. sgRNA, ssODN, Cas9 protein and mRNA (44758; Addgene) were co-microinjected into zygotes [F1 hybrids between strains FVB/NJ and B6(Cg)-Tyrc-2J/J] using the reagent concentrations listed in Table S1. Edited founders were identified either by subcloning followed by Sanger sequencing using primers and annealing temperatures listed in Table S1. For generating the *Cdk2*$^{Y15S\text{-}T160A}$ allele, *Cdk2*$^{Y15S/Y15S}$ females were used as embryo donors, thus, the *Cdk2*^{Y15S} alteration was not re-introduced by CRISPR. The *Cdk2*$^{-/-}$ allele contained a 2nt deletion in exon1, leading to a premature termination codon, leading to a predicted truncated CDK2 protein of 31 aa.

Testes histology and Immunohistochemistry. For histological analyses, testes were fixed for 24 h at room temperature (RT) in Bouin's solution, paraffin-embedded, sectioned at 7 μm, and then stained with H&E. For immunohistology, testes were fixed in 4% paraformaldehyde for ~24 h, paraffin-embedded, sectioned at 7 μm, and deparaffinized. Antigen retrieval for different antibodies was performed as indicated in Table S4. Sections were blocked in PBS containing 5% goat serum for 1 h at RT, followed by incubation with primary antibodies for 12 h at 4°C and detection by the secondary antibody (see antibody details in Table S4).

Whole-Mount Seminiferous Tubule Staining. Seminiferous tubule whole mounts were prepared as described previously (Savitt et al., 2012), with minor modifications. Briefly, seminiferous tubules were manually isolated in phosphate-buffered saline (PBS) and interstitial tissue was washed thoroughly followed by fixation in 2% paraformaldehyde for 2 h. After extensive washing, tubules were processed for detection of progenitor spermatogonial markers and EdU labeling. To detect ZBTB16, FOXO1, cPARP1 and LIN28, tubules were blocked in PBSS buffer (5% goat serum, 0.5 % Triton X-100 in 1x PBS) for 2 h at RT following primary antibody incubation for 12 h-18 h at 4^{O}C in the same buffer. Tubules were washed at room temperature twice for 15 minutes, five times for 1 hour in PBSS, and then incubated overnight at 4^{O}C with secondary antibodies. To detect GFRA1, the blocking and all incubations and washing steps are done exclusively in PBS-AT (1% BSA, 0.1% Triton X-100 in 1x PBS) buffer. Briefly, isolated tubules were blocked for 2 h and incubated in goat anti-GFRA1 antibody diluted in PBS-AT buffer. This was followed by five subsequent washing (30 min each) and detection by secondary antibodies. While performing co-immunolabeling of GFRA1 with other spermatogonial markers, ZBTB16, FOXO1, cPARP1 and LIN28 were always detected prior to GFRA1.

For EdU incorporation, mice were injected intraperitoneally with 100 mg/kg EdU (PY7562; Berry & Associates) before harvesting and immediately processive testes 4 hours later for whole-mount staining, or fixed in 4% PFA for IHC. Following antibody staining, freshly prepared Click reaction cocktail (10mM (+)-sodium-L-ascorbate, 0.1mM 6-

Caboxyfluorescein-TEG azide (FF 6110; Berry & Associates) or Alexa Fluor® 594 Azide (A10270; Invitrogen™), and 2mM copper (II) sulfate in water) was added into whole mount tubules or testes cross sections and removed quickly after 45 sec. Excess EdU was removed by extensive washings (2x 1h and 2x 12h at 4 ^{o}C) using PBS-AT buffer. Following washing, tubules were mounted in Vectashield (Vector Laboratories, Burlingame, CA). Click reaction was always performed following antibody detection. In each animal, we counted all A_s, A_{pr} and A_{al} GFRa1$^+$ cells positive or negative for EdU seminiferous tubules at least >8 mm in length.

Imaging. Slide preparations were scanned and tiled images of testes cross sections or seminiferous tubule whole mounts were acquired using a Laser Scanning Confocal microscope (u880, Carl Zeiss, Germany) using Plan Apo 40× water immersion objective (1.1 NA) and Zen black software. Initial laser power adjustment was performed to avoid saturation of signal: argon laser-488 nm, blue-diode-405 nm, DPSS laser—561 nm and HeNe-633. Additional images of testes cross sections were obtained through an Olympus BX51 microscope with objectives 100×/1.35 NA infinity/0.17 or objective or 10×/0.3 NA, respectively using Olympus Cell Sens software (Olympus). Following identical background adjustments for all images, cropping, color, and contrast adjustments were made with Adobe Photoshop CC 2017.

In vitro kinase assay. These were performed as described elsewhere (Lim et al., 2017). Briefly, 1μg of anti CDK2 antibody was added to 0.5μg of protein lysate and rotated overnight (approx. 16hrs) at 4^{o}C. Twelve μl of protein A beads was added to each sample and left to mix for a further 2 hours. Prior to their use, these beads were washed twice previously using 1ml of EBN buffer (80 mM beta-glycerophosphate; 15 mM $MgCl_2$; 20 mM EGTA- adjusted to pH7.3 with KOH; followed by adding 150mM NaCl, 0.5% NP40 to 0.and 1X protease inhibitor cocktail). After binding, the beads were spun down and washed with EBN buffer for 3 times. Pellets were resuspended in 14μl EBN buffer (containing protease inhibitors: cOmplete, Mini, 11836153001; Roche) followed by adding 6μl of Histone H1 (260ng/ul) and 9ul of kinase assay buffer (15mM EGTA, 25mM NaF, 250mM sodium beta glycerophosphate, 5mM DTT, 20mM $MgCl_2$, 21μM ATP). Following 20 min incubation at room temperature, 1μl of ATP [γ-32P] containing a specific radioactivity of 5μCi (Perkin Elmer: #BLU502A) was added to each sample and left for 30 minutes at 30^{o}C, 400rpm. 6X SDS sample buffer (2M beta-mercaptoethanol 0.375 M Tris pH 6.8.12% SDS,60% glycerol, 0.6M DTT,0.06% bromophenol blue) was added to a final concentration of 1X and each sample was boiled at 95^{o}C for 5 minutes. 12μl of each sample were analyzed for Coomassie stained Histone H1 protein bands and Phosphosignal on a phosphor screen cassette for 6-24hrs. Phosphosignal was quantified using FLA7000 phosphimager (Fujifilm).

Western blotting. Protein was extracted using EBN buffer (see above) with protease inhibitors. Samples were boiled at 95°C for 5 minutes and were separated by SDS/PAGE (12% acrylamide). Separated proteins were electrotransfer to 0.2μM nitrocellulose (Biorad #1620112) or PVDF membrane (Immobilon-P membrane, IPVH00010; EMD Millipore), and then blocked in 5% nonfat milk for 45 min at RT. Membrane was washed 3X 10

minutes in TBST (0.14M NaCl, 15mM KCl, 25mM Tris Base, 1% Tween20) at room temperature followed by ~16 h incubation with primary antibody, washing, and ~1 h incubation with HRP conjugated secondary antibody (as stated in Table S4) for 45 minutes at room temperature. Signal was detected using Luminata Classico Western HRP substrate (WBLUC0100; EMD Millipore). Densiometric analysis of western blot bands was performed using Fujifilm Multi Gauge software Ver. 3.1.

Site directed mutagenesis. Site-directed mutagenesis for obtaining CDK2-Y15F and CDK2-Y15S cDNA was accomplished using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent, La Jolla, CA; Cat# 210518) using primers sets mentioned in Table S3. All mutations were confirmed by Sanger sequencing.

CDK2 overexpression and immunoprecipitation. The (Myc-DDK-tagged)-CDK2 cDNA was obtained from Origene (Cat#RC200494). Y15F and Y15F mutant cDNA were generated as described above and transfected into HEK-293T cells (ATCC; CRL-3216) using TransIT®-LT1 Transfection Reagent (Mirus; MIR2305) following user instructions. Lysates were prepared from harvested cells in lysis buffer (50mM Tris pH7.5, 150mM NaCl, 0.5% Triton X-100, 5 mM EDTA) containing phosphatase (Thermo Scientific; 78428) and protease inhibitors (Roche; 04693159001). 400 μg of each lysate was immunoprecipitated using c-Myc-Tag IP/Co-IP Kit (Pierce, 23620). Proteins were separated on 12% poly acrylamide gel by SDS-PAGE and CDK2 protein was visualized by western blot analysis.

Protein Identification by nano LC/MS/MS Analysis. In-gel trypsin digestion of immunoprecipitated CDK2 protein was performed as described earlier (Yang et al., 2007). The tryptic digests were subjected to nanoLC-ESI-MS/MS analysis on Orbitrap Fusion™ Tribrid™ (Thermo-Fisher Scientific, San Jose, CA) mass spectrometer equipped with a nanospray Flex Ion Source, and a Dionex UltiMate3000RSLCnano system (Thermo, Sunnyvale, CA) following a protocol described earlier (Thomas et al., 2017; Yang et al., 2018). The column was re-equilibrated with 0.1% FA for 23 min prior to the next run. For data-dependent acquisition (DDA) analysis, the instrument was operated using FT mass analyzer in MS scan to select precursor ions followed by 3 second "Top Speed" data-dependent CID ion trap MS/MS scans at 1.6 m/z quadrupole isolation for precursor peptides with multiple charged ions above a threshold ion count of 10,000 and normalized collision energy of 30%. MS survey scans at a resolving power of 120,000 (fwhm at m/z 200), for the mass range of m/z 375-1575. Dynamic exclusion parameters were set at 30 s of exclusion duration with ±10 ppm exclusion mass width. All data were acquired under Xcalibur 3.0 operation software (Thermo-Fisher Scientific).

Mass spec data analysis. The DDA raw files for CID MS/MS were subjected to database searches using Proteome Discoverer (PD) 2.2 software (Thermo Fisher Scientific, Bremen, Germany) with the Sequest HT algorithm. All 3 raw MS files for three samples were used for database search. Processing workflow for precursor-based quantification. The PD 2.2 processing workflow containing an additional node of Minora Feature Detector for precursor ion-based quantification was used for protein and ptms identification. The

database search was conducted against a *mouse* database containing ~20153 entries downloaded from NCBI on Jan. 12, 2018 plus some common contaminants (246 entries) database. Two-missed trypsin cleavage sites were allowed. The peptide precursor tolerance was set to 10 ppm and fragment ion tolerance was set to 0.6 Da. Variable modification of methionine oxidation, deamidation of asparagines/glutamine, phosphorylation of serine, threonine and tyrosine and fixed modification of cysteine carbamidomethylation, were set for the database search. Only high confidence peptides defined by Sequest HT with a 1% FDR by Percolator were considered for the peptide identification. The final protein IDs contained protein groups that were filtered with at least 2 peptides per protein.

Sperm counting. Sperm counting was done as described earler (Singh and Schimenti, 2015).

scRNA seq sample preparation. 3.5 DPP testes were decapsulated in HBSS (Mediatech Inc.) and digested with 0.642 ml of Trypsin/EDTA (0.25%, Invitrogen Inc.) and 0.071 ml DNase I (1 mg/ml in HBSS, Sigma Inc.) at 37°C for 3-5 min. Digestion was stopped by adding 0.5 ml of media (HBSS + 10% FBS), the cells were filtered through a 70-µm cell strainer and resuspended in HBSS with 0.04% FBS. Single-cell 3' RNA-seq sequencing libraries for Illumina were constructed using a 10X Genomics Chromium instrument, using the Chromium Single Cell 3' Reagent kits (v2) following the manufacturer's protocols. The final libraries were quantified by digital PCR and sequenced on Illumina sequencers, using an Illumina NextSeq500/550 75 bp kit (26 bp + 8 bp index read + 58 bp). The target number of cells captured from the input, single-cell suspension was 8700. The data were demultiplexed and aligned to the reference genome using the 10X Genomics Cellranger software (v2.2) and visualized using the 10X Genomics Loupe Cell Browser software (v2.0) packages.

scRNA-seq data analysis. The raw scRNA-seq data was analyzed using cell ranger from 10X platform to generate a matrix of raw read counts, which was further analyzed in R using Seurat (Butler et al., 2018; Satija et al., 2015) and Monocle (Qiu et al., 2017). A cluster of germ cells from WT and *Cdk2*$^{Y15S/Y15S}$ were separated from rest of the testis cells based on *Ddx4* gene expression. We then isolated cell barcodes that had more than 10,000 UMIs per cell (see Results). Raw read counts for these selected germ cells were further filtered to keep only those genes that are detected in at least 10% of the cells. This resulted in an expression matrix of 10,451 genes across 141 cells (WT=69; *Cdk2*$^{Y15S/Y15S}$ =72). Due to high rate of drop-outs in mRNA-capture in scRNA-seq approaches, the expression values for mid- and lowly-expressed genes are often unreliable due to missing information. Hence, we recovered expression for each gene using SAVER (Huang et al., 2018), an approach that estimates gene expression from UMI-based scRNAseq data using information across all genes and cells. This estimated gene expression was used for further analysis. The most variable genes across all cells were identified using principle component analysis (PCA). Specifically, PCA was performed using log-transformed counts-per-million (CPM) expression values. The top 300 genes with highest absolute loading for PC1, top 300 genes for PC2 and top 300 genes for PC3 were chosen, which resulted into 744 unique genes. Furthermore, this list of variable genes was filtered to remove mitochondrial genes (n=10). The 141 germ cells were clustered

into five groups using "Ward.D2" method based on expression estimates for the 734 highly variable genes. The pseudotime and trajectory analysis were performed based on the expression of 734 most variable genes using Monocle R code. RNA velocity analysis (La Manno et al., 2018) was performed using Velocyto, available at velocyto.org. Gene set enrichment analysis (GSEA) (Subramanian et al., 2005) was performed on the web server (software.broadinstitute.org/gsea/index.jsp).

Statistics. P values were calculated from unpaired Student's t-test for all IHC and whole mount experiments.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Institutes of Health (R01 HD082568 to JCS and P50HD076210 to AG and DAP), and contract CO29155 from the NY State Stem Cell Program (NYSTEM). The authors would like to thank R. Munroe and C. Abratte of Cornell's transgenic facility for generating the *Cdk2*^{T160A} allele alone and in *cis* to the *Cdk2*^{Y15S} change. Authors also thank the Proteomics Facility of Cornell University for providing the mass spectrometry data and NIH SIG 1S10 OD017992-01 grant support for the Orbitrap Fusion mass spectrometer. Confocal Imaging data was acquired in the Cornell BRC-Imaging Facility using the NIH-funded (S10OD018516) Zeiss LSM880 confocal/multiphoton microscope (u880).

COMPETING INTERESTS

The authors declare no competing or financial interests.

AUTHOR CONTRIBUTIONS

P.S. conducted all experiments related to mouse breeding, cytology, histology, and preparing samples for sc sequencing. She also contributed to bioinformatics analyses and writing the paper. R.K.P. did most of the scRNA-seq analyses, with contributions from J.K.G. and supervision from A.G. N.P performed the kinase activity assays under the supervision of P.K., and they wrote the relevant sections and contributed to overall interpretations. D.A.P. made the initial observations that *Cdk2*$^{Y15S/Y15S}$ mutant adult testes were not devoid of spermatogonia. J.C.S. oversaw the entire study and shared the bulk of manuscript writing with P.S.

FUNDING

This research was funded by National Institutes of Health grants (R01 HD082568 to J.C.S. and P50 HD076210 to A.G.), the Biomedical Research Council, Agency for Science, Technology and Research (A*STAR) to P.K., by SINGA (Singapore International Graduate Award) to N.P., by the Biomedical Research Council – Joint Council Office Grant (1231AFG031 to P.K.); by the National Medical Research Council Singapore, NMRC (NMRC/CBRG/0091/2015) to P.K., and by National Research Foundation Singapore grant

NRF2016-CRP001-103 to PK.

DATA AVAILABILITY

The 10x single-cell RNAseq data is available at GEO with accession GSE130554.

FIGURES

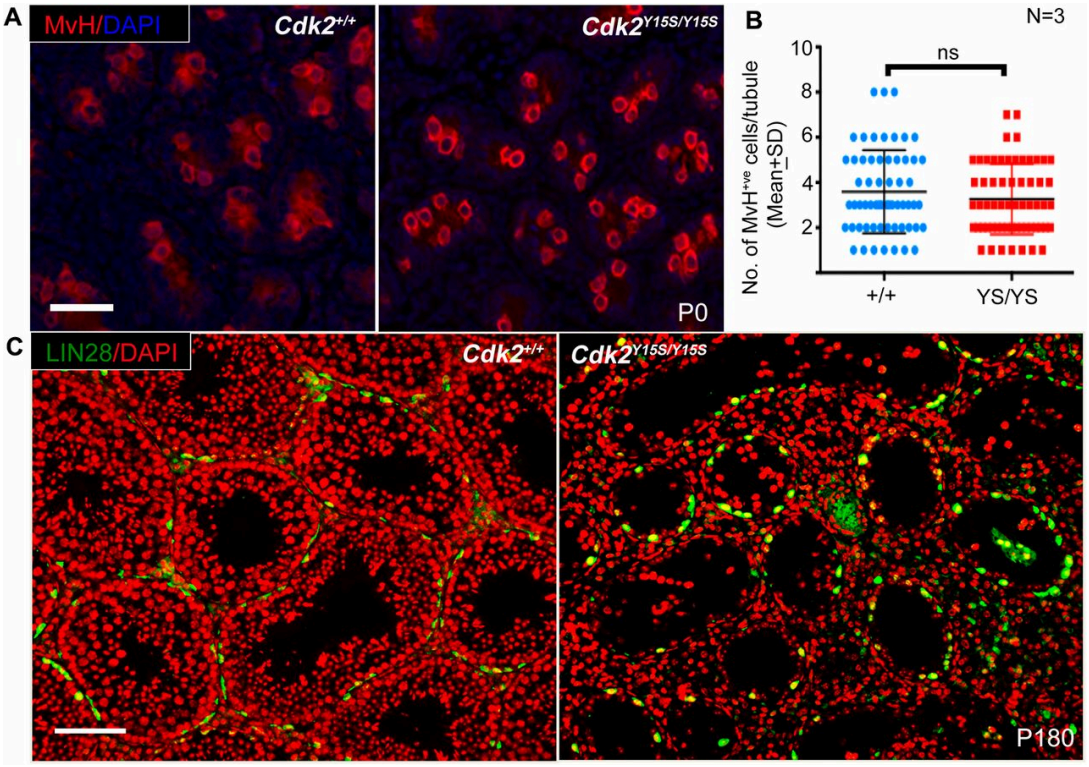

Figure 1. ***Cdk2*$^{Y15S/Y15S}$ mice are born with a normal germ cell complement, but exhibit defective spermatogonial differentiation. (A)** IHC of newborn testis sections germ cells with germ cell marker MVH. Scale bars = 50µm. **(B)** Quantification of data in "A". YS = *Cdk2*^{Y15S}. **(C)** IHC on adult testis sections. Scale bars, 100µm.

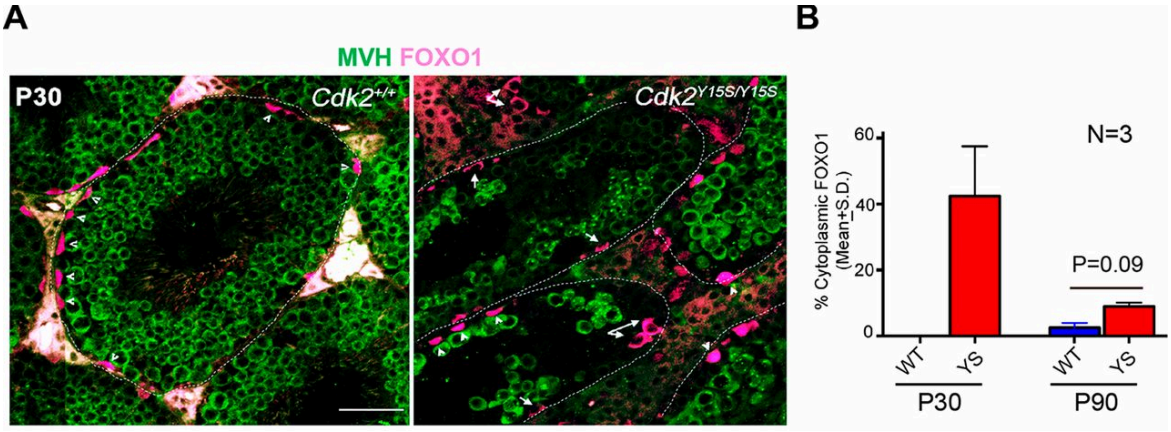

Figure 2. ***Cdk2*$^{Y15S/Y15S}$ mice have a defective gonocyte-to-spermatogonia transition.** **(A)** Immunolabeling of testis histological sections. Arrows and arrowheads indicate germ cells with cytoplasmic or nuclear FOXO1, respectively. The locations of some seminiferous tubule boundaries are indicated by dashed lines. **(B)** Percentage of cytoplasmic FOXO1 in P30 and P90 tubule sections. YS = *Cdk2*$^{Y15S/Y15S}$. Scale bar = 50μm.

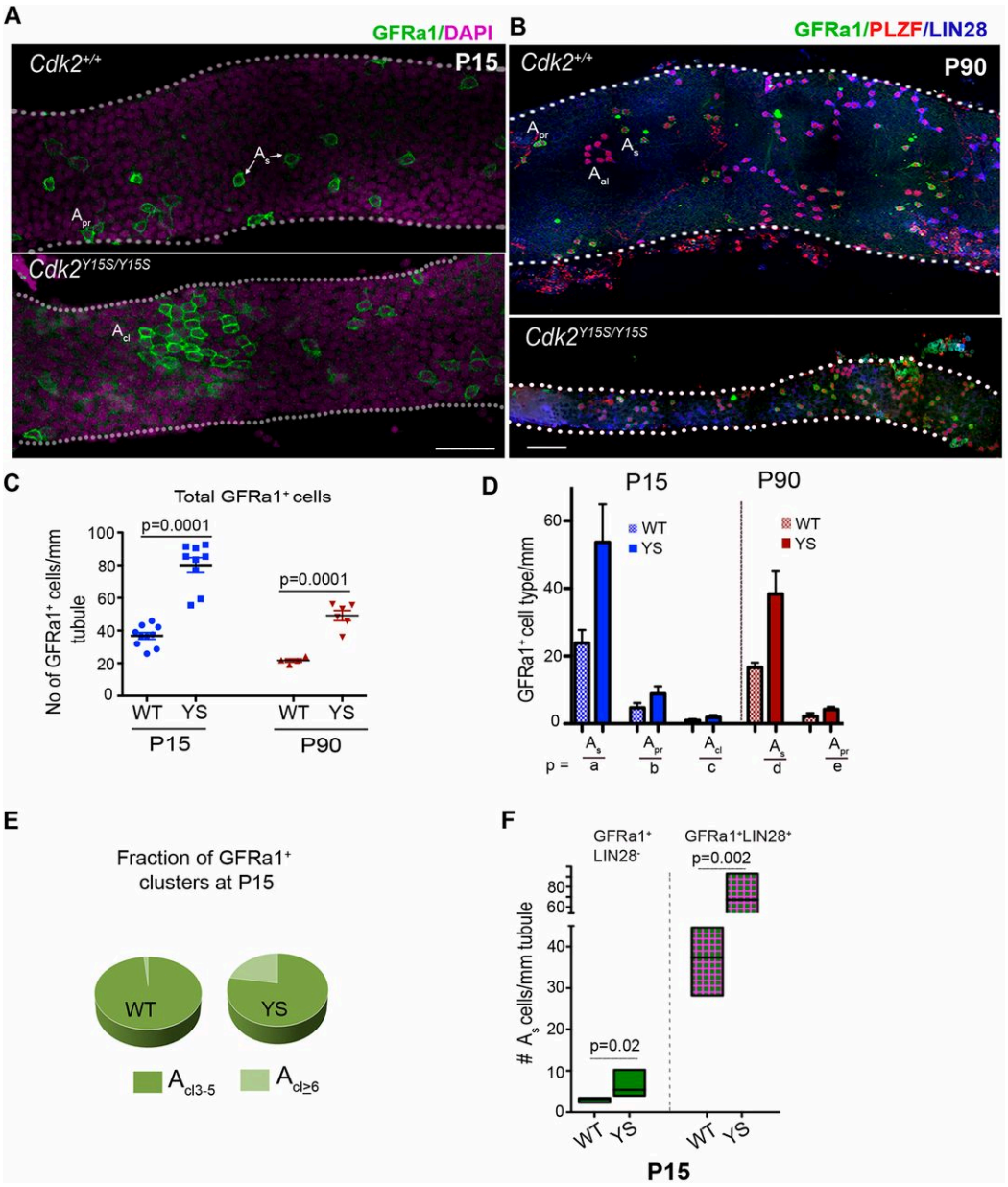

Figure 3. ***Cdk2*$^{Y15S/Y15S}$ seminiferous tubules contain increased undifferentiated spermatogonia. (A)** Confocal images of P15 seminiferous tubules showing density of GFRa1$^+$ spermatogonia. Scale bar, 50µm. **(B)** Confocal images of P90 tubules immunolabelled for markers of undifferentiated spermatogonia. Outermost 2-3 optical sections are projected in A and B. Scale bar, 200µm. Density of **(C)** total and **(D)** A_s, A_{pr} and $A_{cluster}$ GFRa1$^+$ spermatogonia in >8mm long tubule segments. n≥3 for each genotype (except n=2 for P90 WT). p values: a<0.0001, b=0.0001, c=0.0006, d=0.0002, e=0.0039. **(E)** Distribution of GFRa1$^+$ subpopulations. Cl = cluster. **(F)** Quantification of A_s spermatogonia subtypes. YS = *Cdk2*$^{Y15S/Y15S}$. Two or more tubules (6-8 mm length) from 3 different animals were quantified.

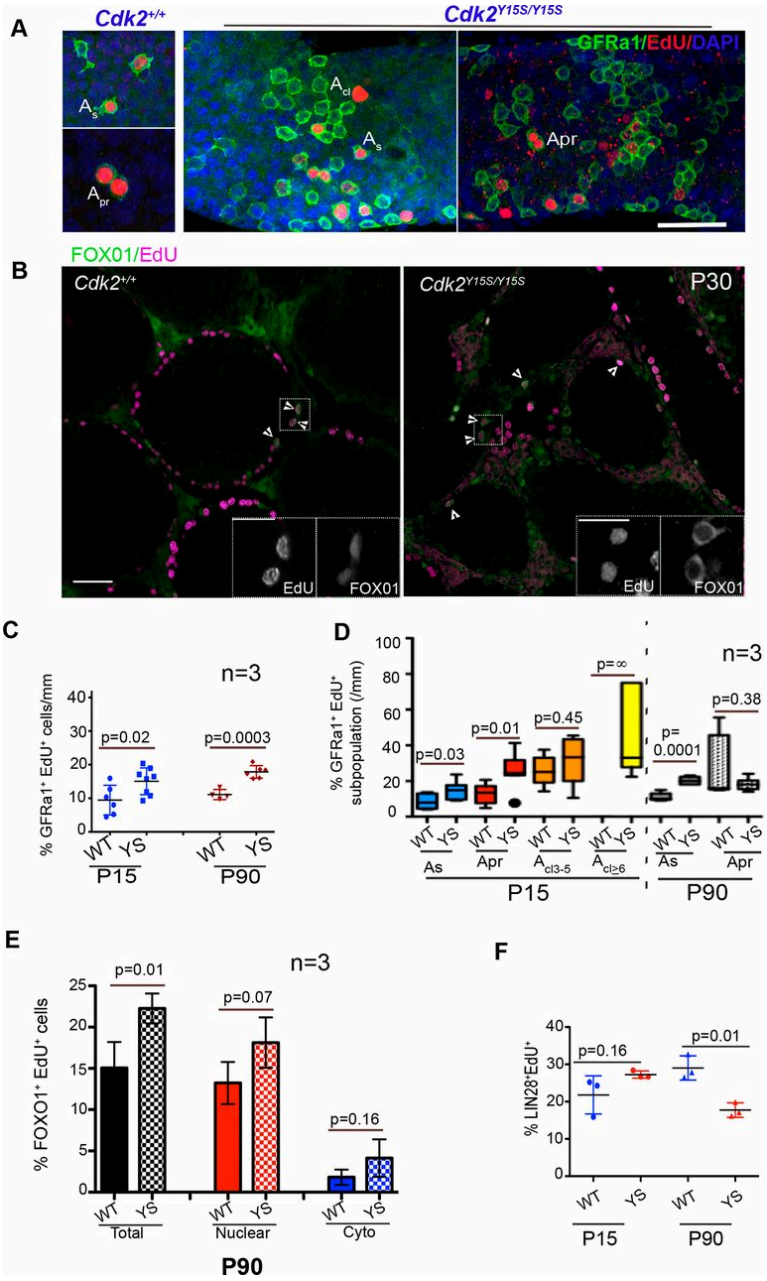

Figure 4. Mutation of the CDK2^{Tyr15} phosphorylation site impedes differentiation of GFRa1^{+} A$_{s}$ spermatogonia. (A) Seminiferous tubules (P15) whole mounts showing abnormally long clusters of GFRa1^{+} and EdU^{+} spermatogonia. A_{cl} = $A_{cluster}$. The data are quantified in **(C). (B)** DNA replication in FOXO1^{+} germ cells. Arrowheads indicate examples of double positive cells. FOXO1 is mostly nuclear in WT, and cytoplasmic in mutants (insets). **(D)** Quantification of replicating spermatogonial subtypes (mean ± SD, n>3 for each group). **(E)** Replicating FOXO1^{+} cells at P90. Over 100 and 55 tubules were analyzed from three different animals of each. Error bars are Std Dev. **(F)** Quantification of replicating LIN28+ spermatogonia. Scale Bar, 50µm in A and B; inset in B, 25µm. YS = *Cdk2$^{Y15S/Y15S}$*.

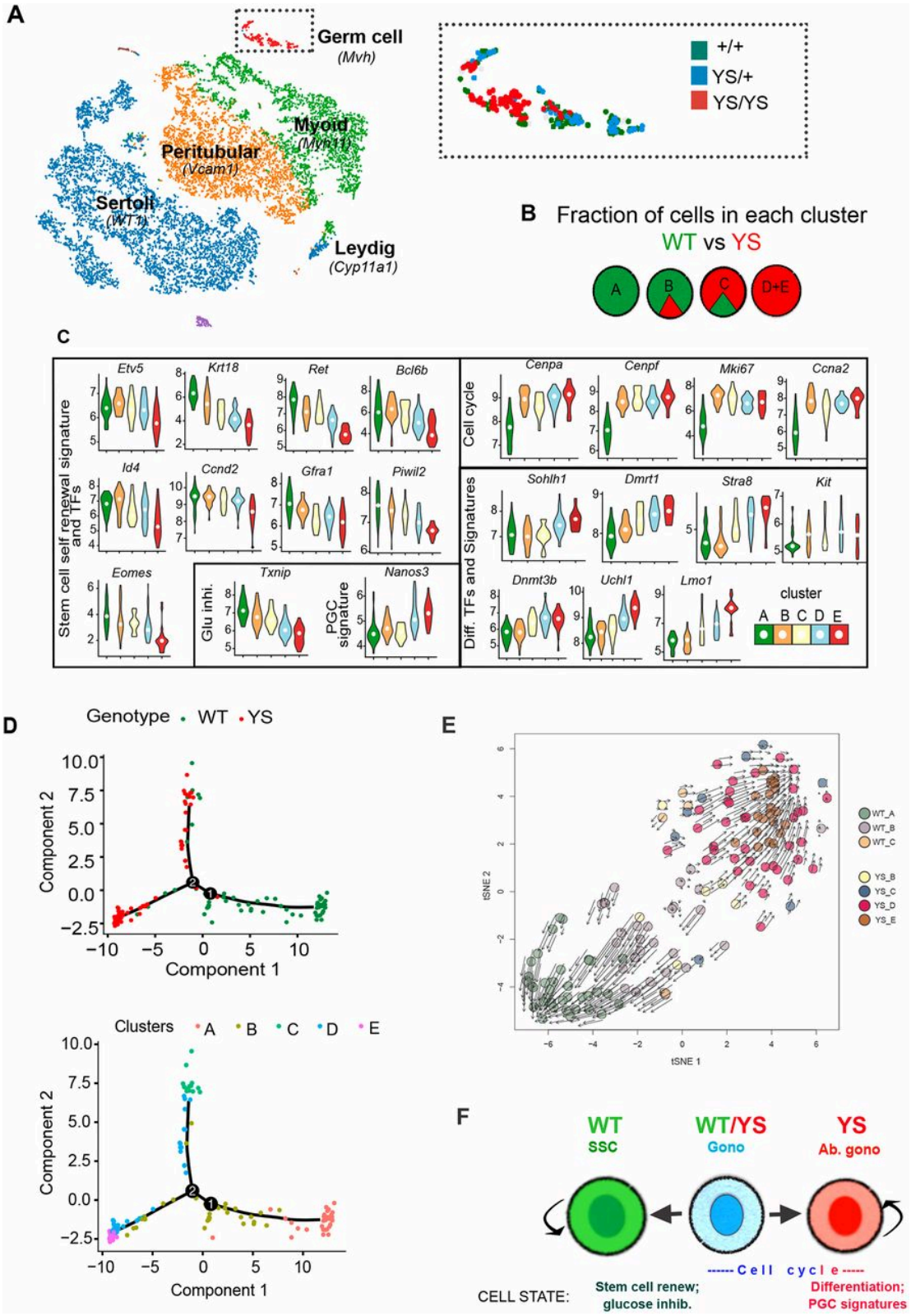

Figure 5. Single cell RNA sequencing of P3.5 testes reveals abnormal differentiation of *Cdk2*$^{Y15S/Y15S}$ gonocytes. (A) A t-SNE (t-distributed stochastic neighbor analysis) plot of 14,422 testicular cells with K-means clustering . Cell types were classified by expression of

key marker genes (e.g., *Myh11* for myoid cells). Inset shows MVH+ 263 total germ cells (WT = 106, *Cdk2*$^{Y15S/Y15S}$ = 99, *Cdk2*$^{Y15S/+}$ = 58), color coded by genotype. YS = *Cdk2*^{Y15S}. **(B)** Identification of germ cell clusters and distribution by genotype. For this panel and also C-D, the WT and *Cdk2*$^{Y15S/Y15S}$ germ cells from "A" were filtered for those that had >10,000 UMIs/cell, leaving in an expression matrix of 10,451 genes across 141 cells (WT=69; *Cdk2*$^{Y15S/Y15S}$ =72). 734 highly variably expressed genes were used to identify 5 distinct clusters, A-E. **(C)** Violin plots showing log2-transformed reads per million (RPM) of key genes in cells from each of the 5 clusters (see color coded legend) defined in "B". **(D)** Pseudotime trajectory analysis plots, showing transitions from one cell state to another, as predicted based on expression of the aforementioned 734 most variable genes. The cells are colored by genotype or clusters A-E. **(E)** RNA velocity plotted in tSNE space. Shaded circles represent cells. Arrows indicate their estimated differentiation trends. The lower left is enriched for cells moving towards a WT cluster A identity, and the upper right towards mutant (cluster C-E). **(F)** Schematic summary of cell states and transitions. The germ cell clusters identified in "B" can be classified into three states: normal SSCs (cluster A), intermediate (clusters B & C), and mutant (clusters D&E). However, based on pseudotime and RNA velocity analyses, the WT cells in clusters B & C (blue) are predisposed to differentiate towards normal SSCs (cluster A), and the mutant cells towards abnormal gonocytes (clusters D & E). SSC= spermatogonial stem cells; Gono.= Gonocytes; Ab. gono. = Abnormal gonocytes; YS = *Cdk2*$^{Y15S/Y15S}$.

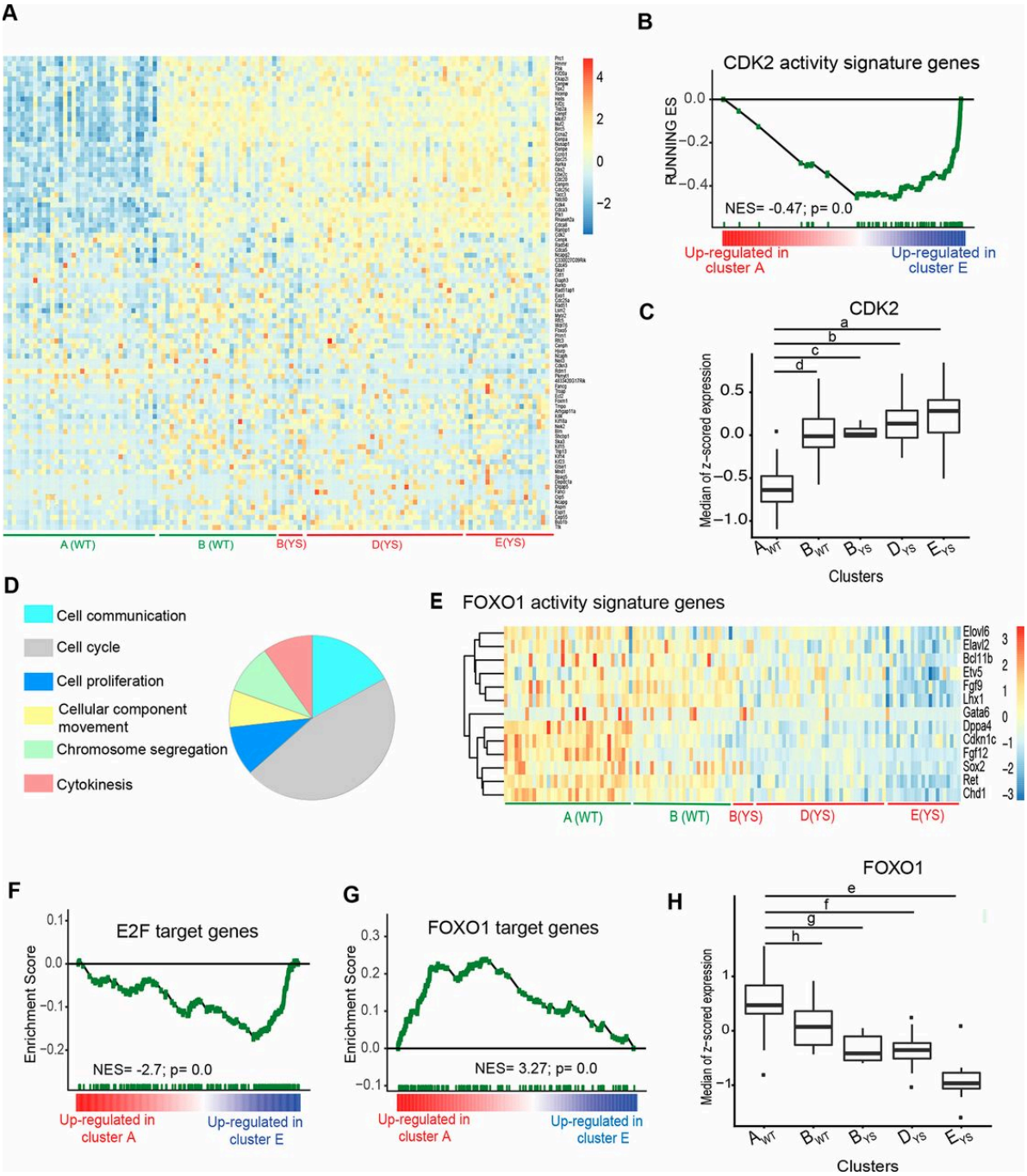

Figure 6: Downstream transcriptional effects of CDK2^{Y15S}. (A) Heatmap showing expression of CDK2 activity signature genes (McCurdy et al., 2017) in indicated clusters of WT *a*nd *Cdk2$^{Y15S/Y15S}$* germ cells. Key: log_2-transformed RPM expression. **(B)** GSEA enrichment plots for CDK2 activity signature genes. **(C)** CDK2 activity scores defined as median of normalized expression of CDK2 target genes per cell, across indicated clusters. (Student's t-test p-values: a= $5.0x10^{-11}$; b= $4.5x10^{-12}$; c= $5.7x10^{-19}$; d= $1.2x10^{-11}$). YS = *Cdk2$^{Y15S/Y15S}$*. **(D)** Enrichment analysis for GO-slim "biological process" terms in genes up-regulated in clusters A vs E. **(E, F)** GSEA enrichment score plots for E2F and FOXO1

target-genes (MSigDb) in clusters A vs E. **(G)** Heatmap showing expression of FOXO1 targets in indicated germ cell clusters and genotypes. Key same as in "A". **(H)** FOXO1 activity defined as median of z-score normalized expression of FOXO1 target genes per cell. (Student's t-test p-values: e= $8.9x10^{-5}$; f= $6.2x10^{-5}$; g= $1.5x10^{-13}$; h= $3.7x10^{-19}$).

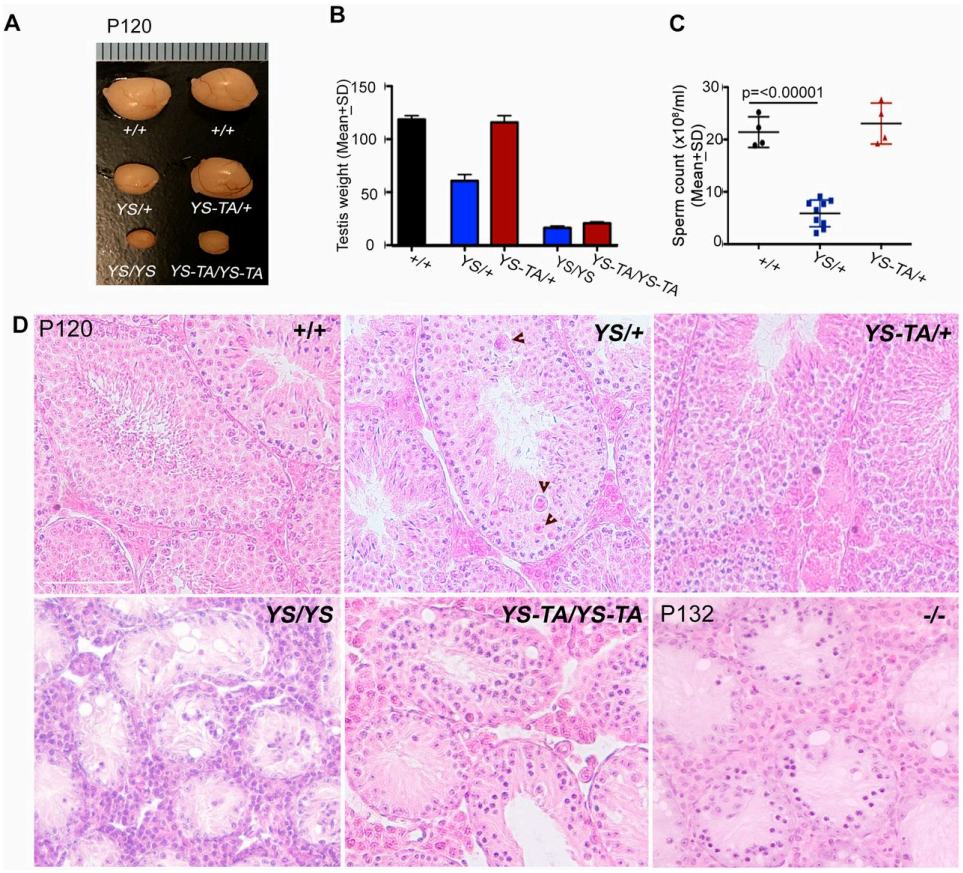

Figure 7. Phenotypic effects of CDK2 Tyr15 and Thr160 phosphorylation during spermatogenesis. (A) Images and **(B)** weights of P120 testes. **(C)** Sperm counts of age-matched genotypes at P120. **(D)** H&E-stained testis cross-sections from adult animals. Abnormal degenerating cells, possibly multinucleate, are indicated by the arrowheads. Error bars in B and C represents +SD. Scale bar in D, 100μM. YS = *Cdk2*^{Y15S}; TA = *Cdk2*^{T160A}; -/- corresponds to a putative null allele of *Cdk2* generated by CRISPR mutagenesis as described in Methods. The formal name is *Cdk2*^{em3Jcs}.

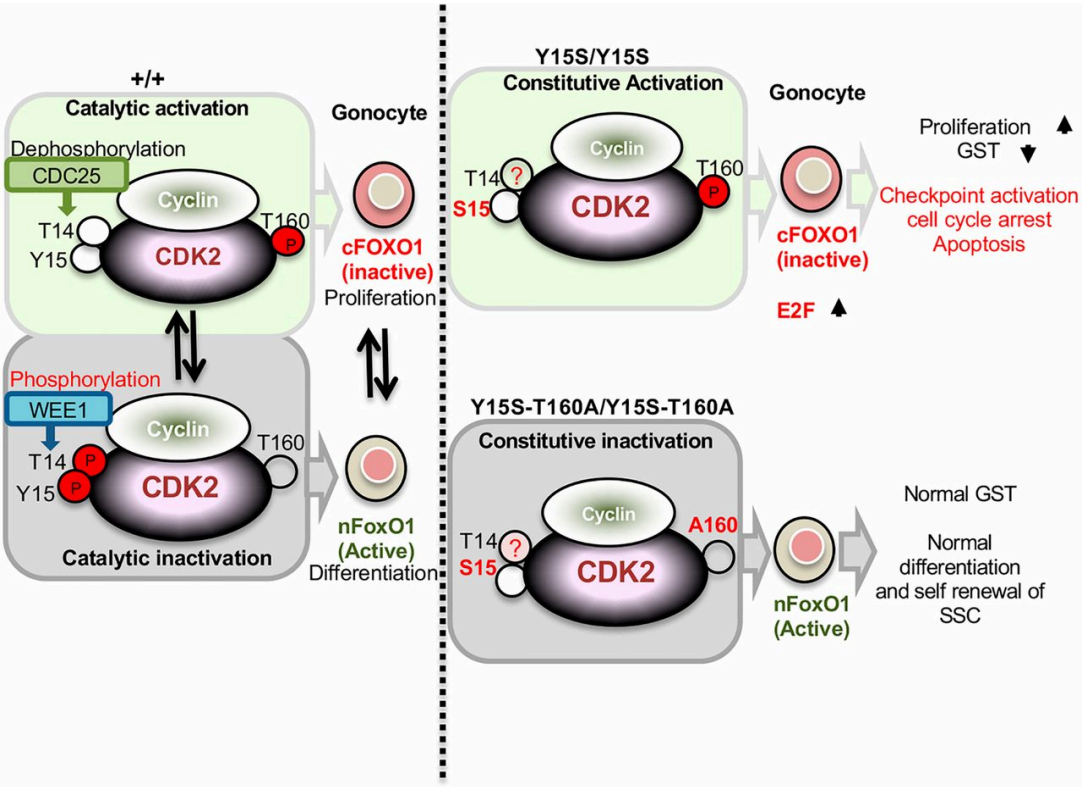

Figure 8. Model for regulation of SSC differentiation by CDK2 phosphorylation.

SUPPLEMENTAL FIGURES

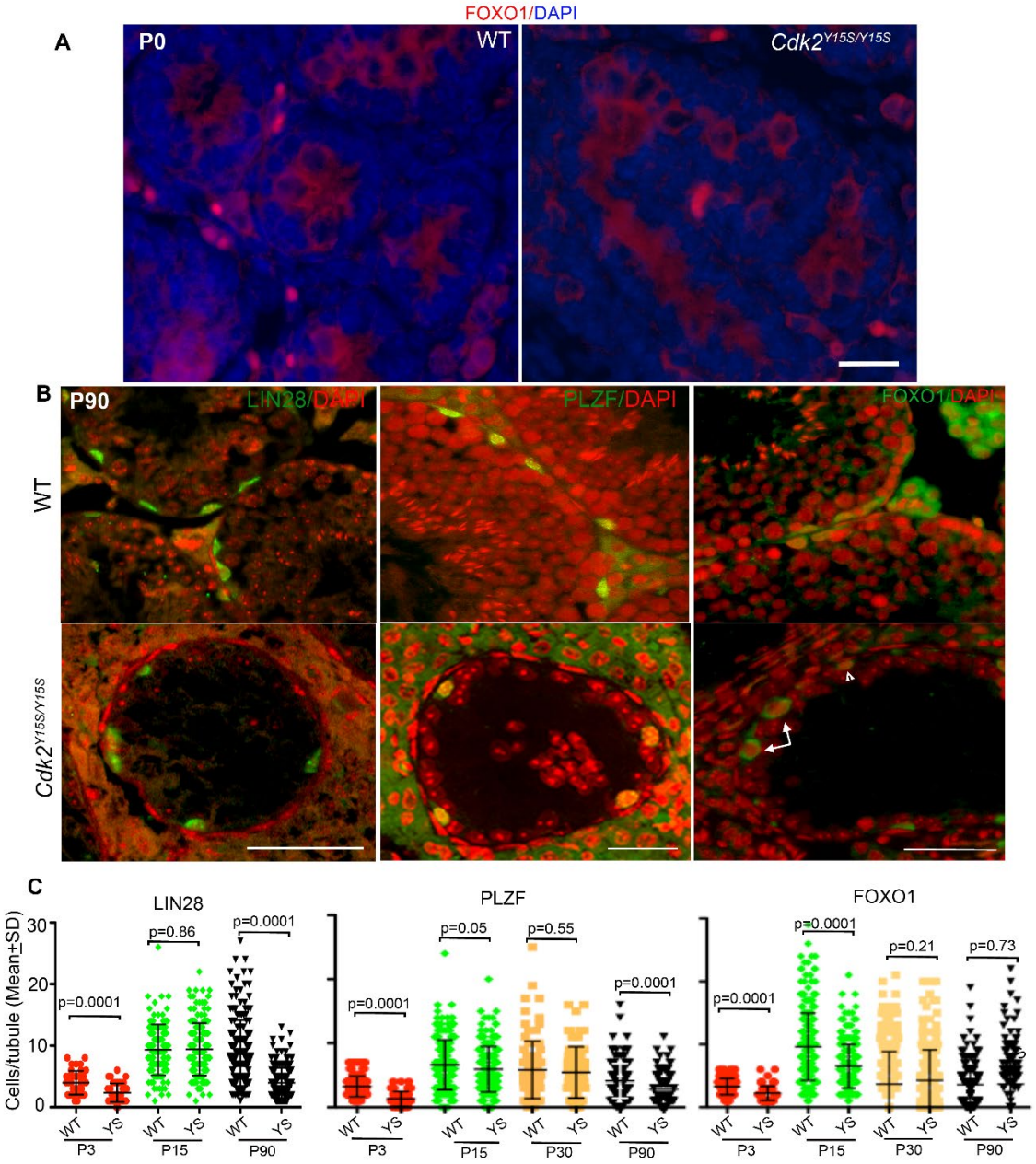

Figure S1: Cdk2Y15S/Y15S testes contain undifferentiated spermatogonia that are unable to differentiate. (A) P0 testis cross sections stained with FOXO1 (red). (B) Sections of WT and mutant P90 gonads immunolabeled with indicated antibodies (all green). (C) Quantification of immunostaining data in B at indicated ages and genotypes. FOXO1 quantification combines cytoplasmic (c) and nuclear (n) FOXO1+ spermatogonia. Arrow, cFOXO1; Arrowhead, nFOXO1. YS = Cdk2Y15S/Y15S. Scale bars, 20 μm in A and 50 μm in B.

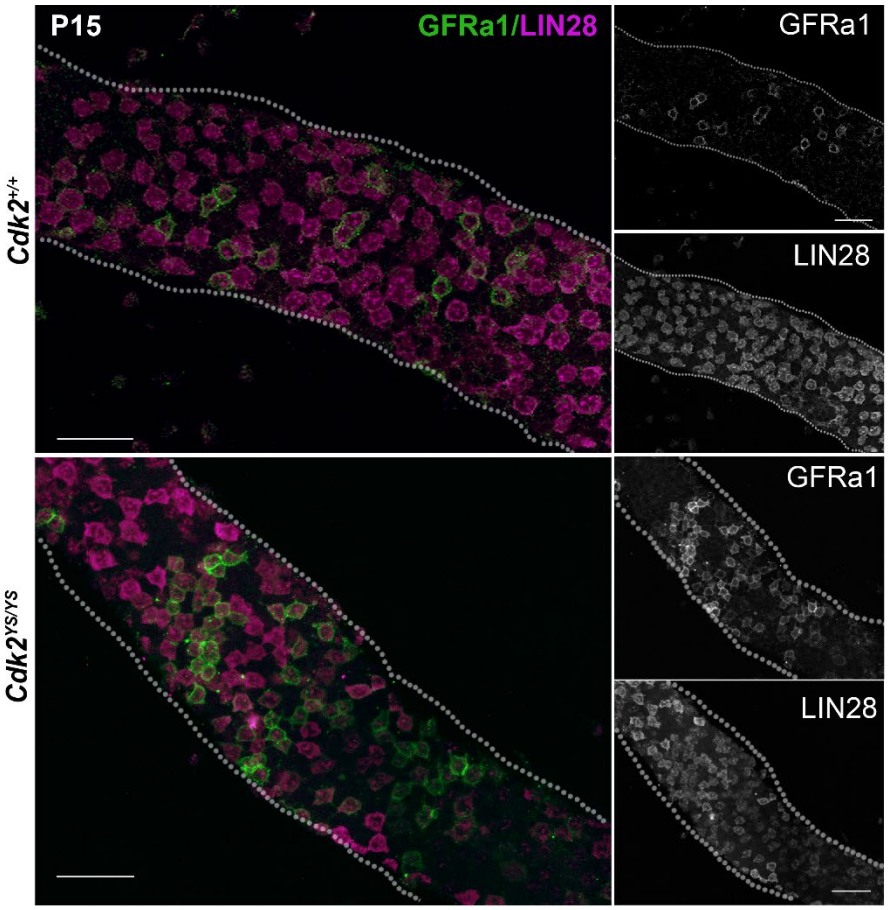

Figure S2. Cells in GFRa1+ clusters in mutant testes co-express LIN28. Scale bar, 50μm.

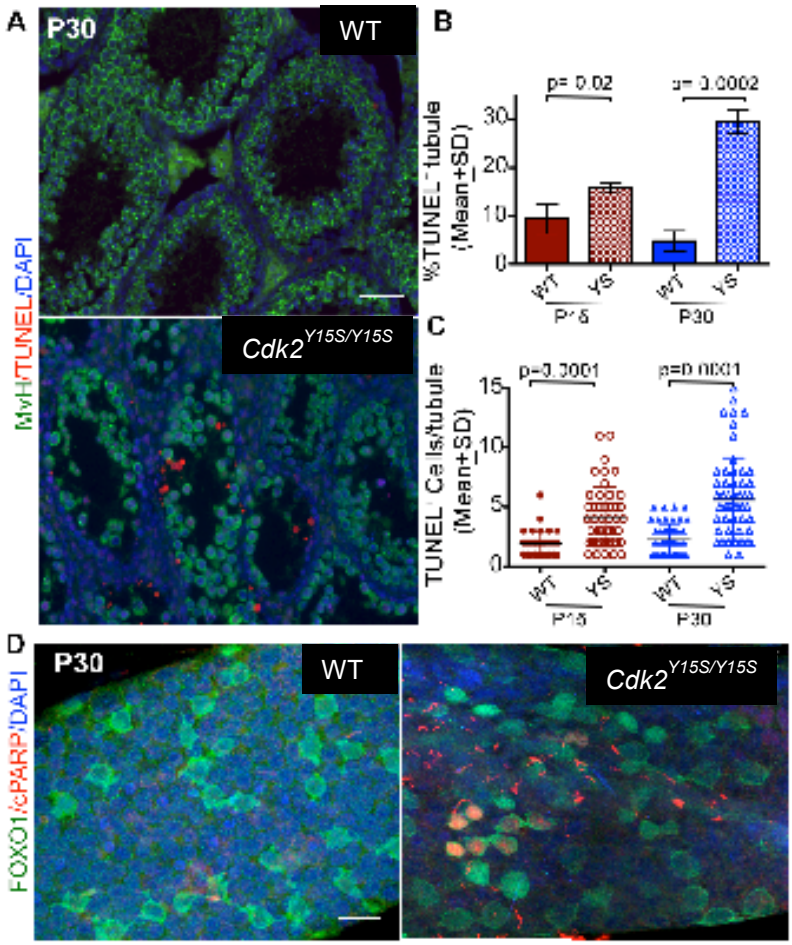

Figure S3: Progressive loss of germ cells in *Cdk2*$^{Y15S/Y15S}$ mutant gonads. (A) Testis cross-section showing presence of TUNEL+ cells in MVH labeled (green) mutant seminiferous tubule. (B,C) Bar graphs of TUNEL+ tubules and TUNEL+ cells/tubule, respectively. (D) Double immunolabeling of WT and mutant seminiferous tubule with cleaved PARP (cPARP, red) and FOXO1 (green). Scale Bar, 20 mm in D and 50 mm in A. YS = *Cdk2*$^{Y15S/Y15S}$.

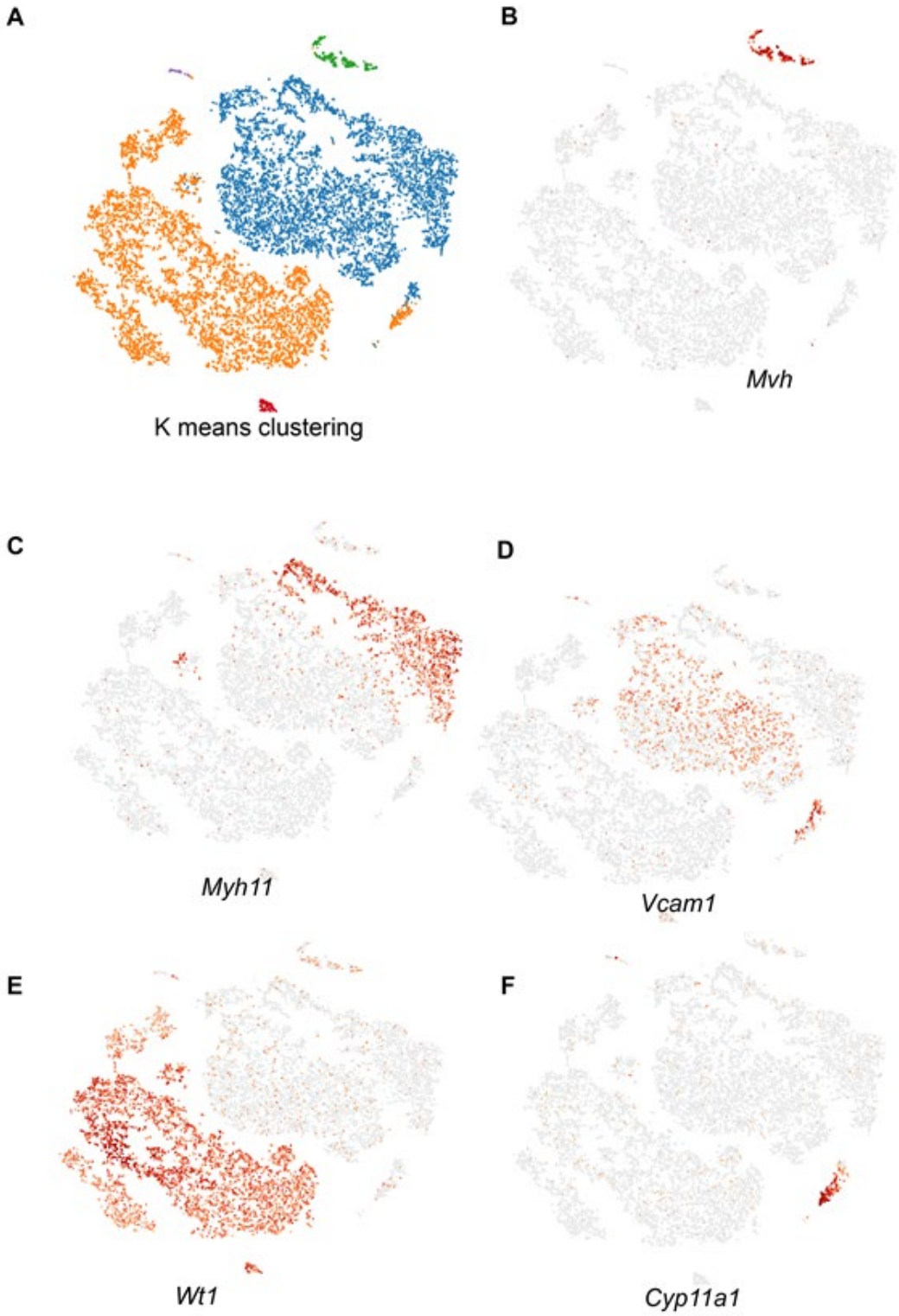

Figure S4. The t-distributed stochastic neighbor embedding (t-SNE) plot identifies 5 clusters of spermatogenic cell types based on K means (A). Cells at different developmental stages are shown in different colors. (B-F) Diagnostic markers used for identification of cell types.

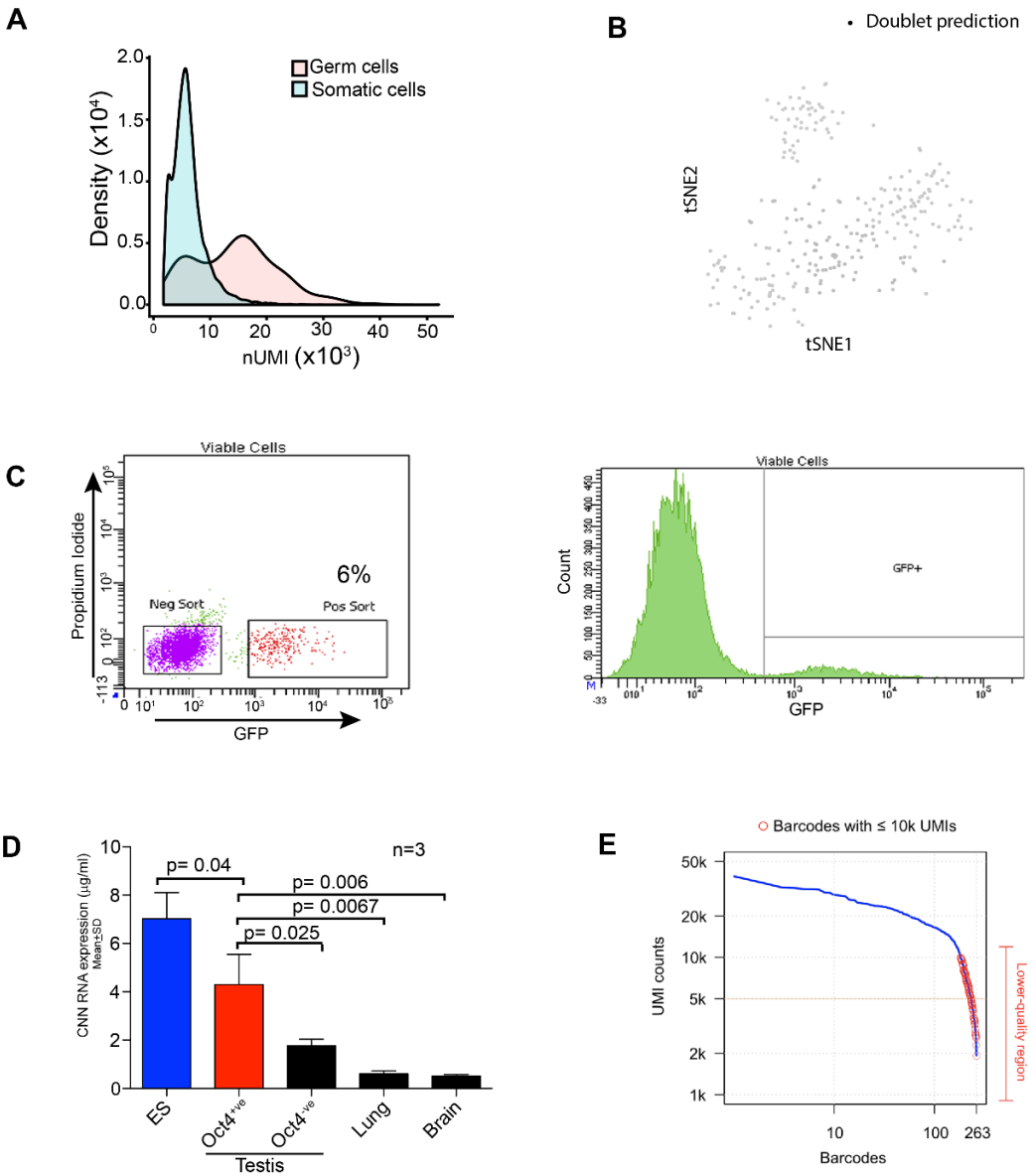

Figure S5. Neonatal germ cells produce more transcripts than somatic cells. (A) Distribution of UMI numbers (nUMI) in germ and somatic cells. The right half of the bimodal germ cell distribution represents 65%, 69% and 73% of the MVH+ germ cell populations in WT, Cdk2Y15S/+ and Cdk2Y15S/Y15S samples, respectively. (B) tSNE plot showing outcomes from the DoubletDetection algorithm analysis of germ cells of all genotypes. Each dot represents one germ cell (263 total). There were no "cells" called as doublets, which would have had been represented as a dark black dot. (C) FACS analysis of GFP+ testis cells from a neonatal Oct4-GFP transgenic mouse. (D) Quantification of total RNA per cell in P3-P5 testicular germ, embryonic stem, and somatic cells. CNN, cell number normalized. (E) 'Knee'-plot analysis identifying cells with lower nUMIs. These are represented as red dots under the "knee".

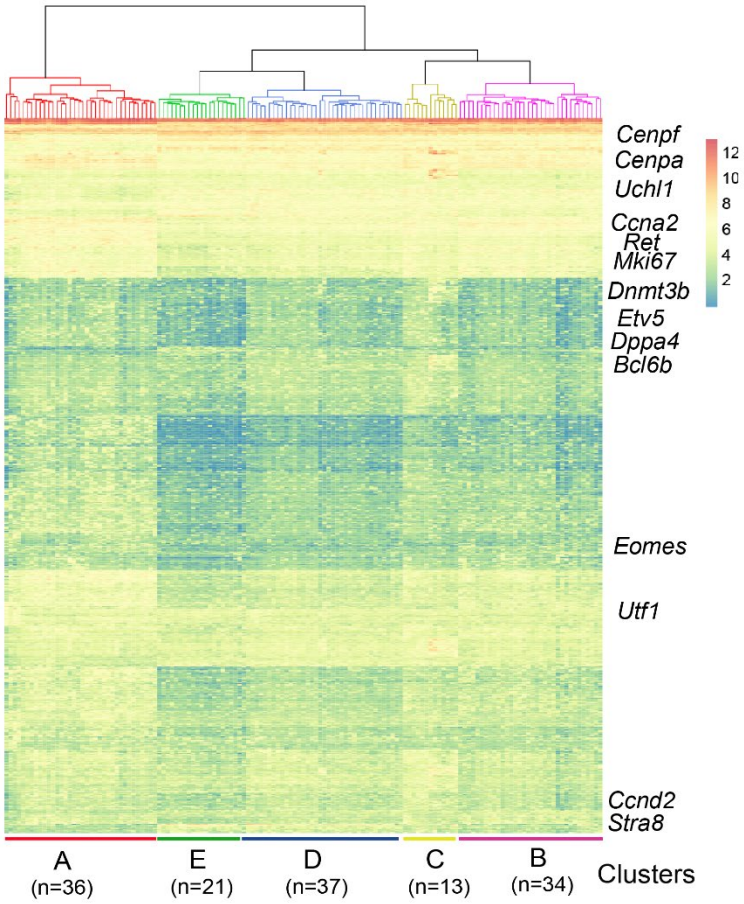

Figure S6. Hierarchical clustering of 141 germ cells from WT (69 cells) and Cdk2Y15S/Y15S (72 cells) testes. Clustering into 5 categories (A-E) is based on the most divergently expressed genes. Notable marker genes are highlighted to the right of the heatmap. Color key represents the relative gene expression levels (normalized and Z-scored).

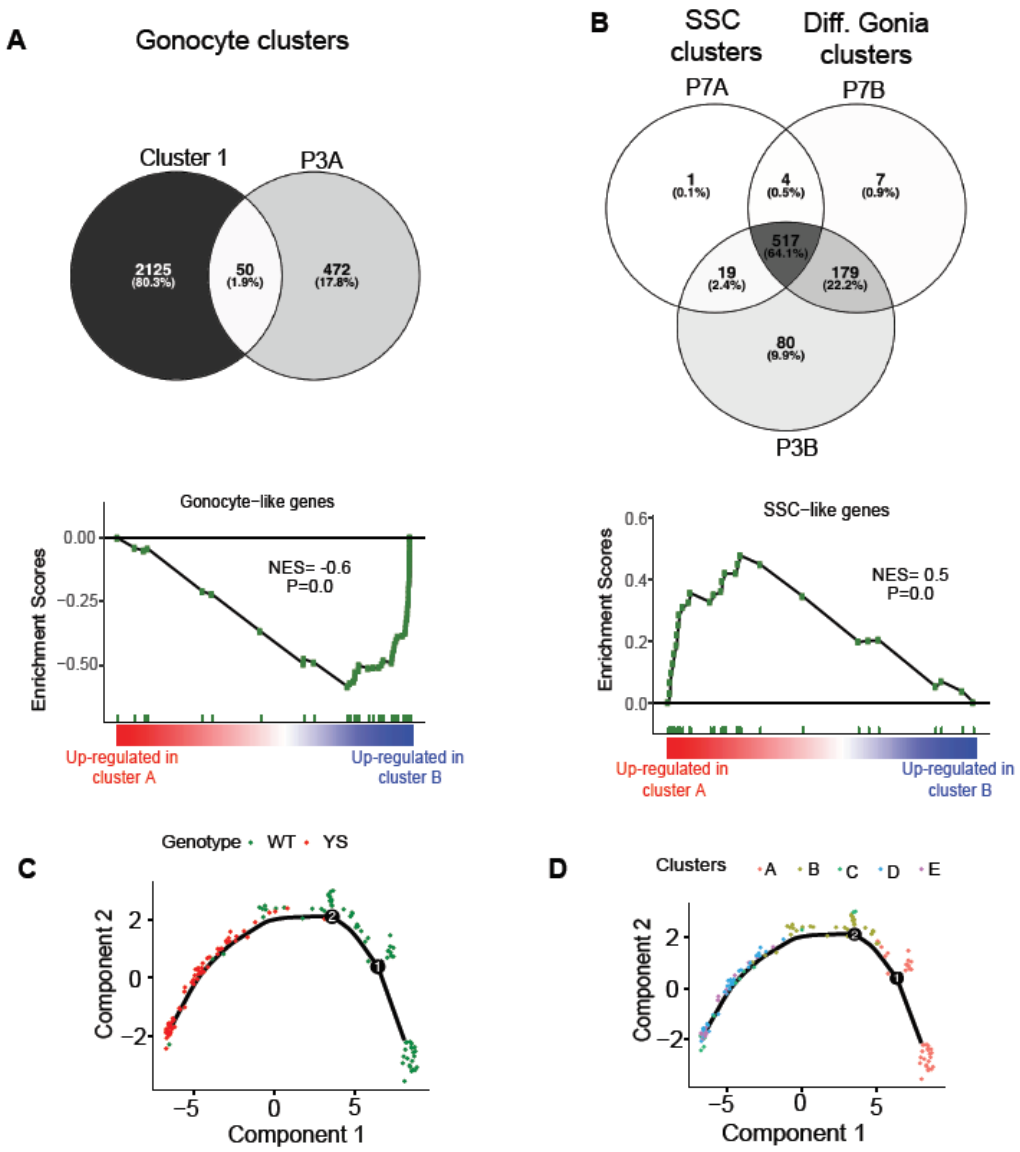

Figure S7. Identification of germ cell subtypes for characterization of mutant cells. Gene set enrichment analysis (GSEA) of gonocyte-like (A) and SSC-like (B) genes. Published scRNA-seq datasets (Liao et. al., 2018; Song et al., 2016) were compared to define reference sets for our data. We selected a gonocyte-like set of 50 genes identified as those in common between "Cluster 1" and "P3A" from the aforementioned publications that were claimed to be gonocytes. We selected an SSC-like set of 100 genes by taking genes from putative SSC-like datasets P3B and P7A, and subtracting those genes expressed in the P7B gene set classified as 'differentiating spermatogonia'. Clusters "A" and "B" in the GSEA plots refer to the clusters we defined from our analyses as described in the text. (C-D) Monocle pseudotime trajectory analysis on data without SAVER-based imputations. The results project the developmental timeline of germ cell subset clusters from WT and mutant ("YS") gonads defined in A,B and Fig 5.

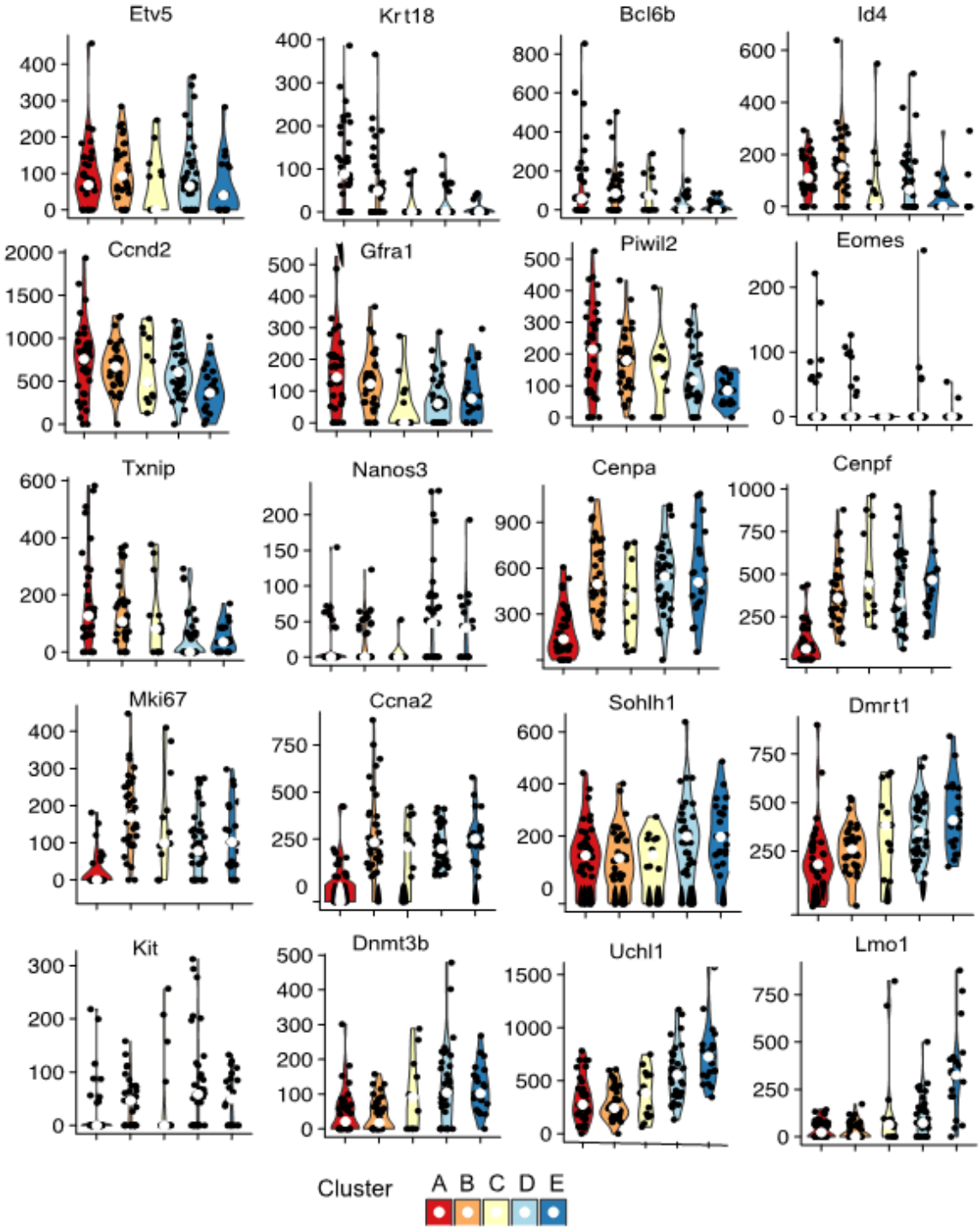

Figure S8. Violin plots of raw data (without SAVER-based imputation) showing expression of selected genes in different cell clusters. The black dots represent cells and white dots indicate median gene expression for the cluster.

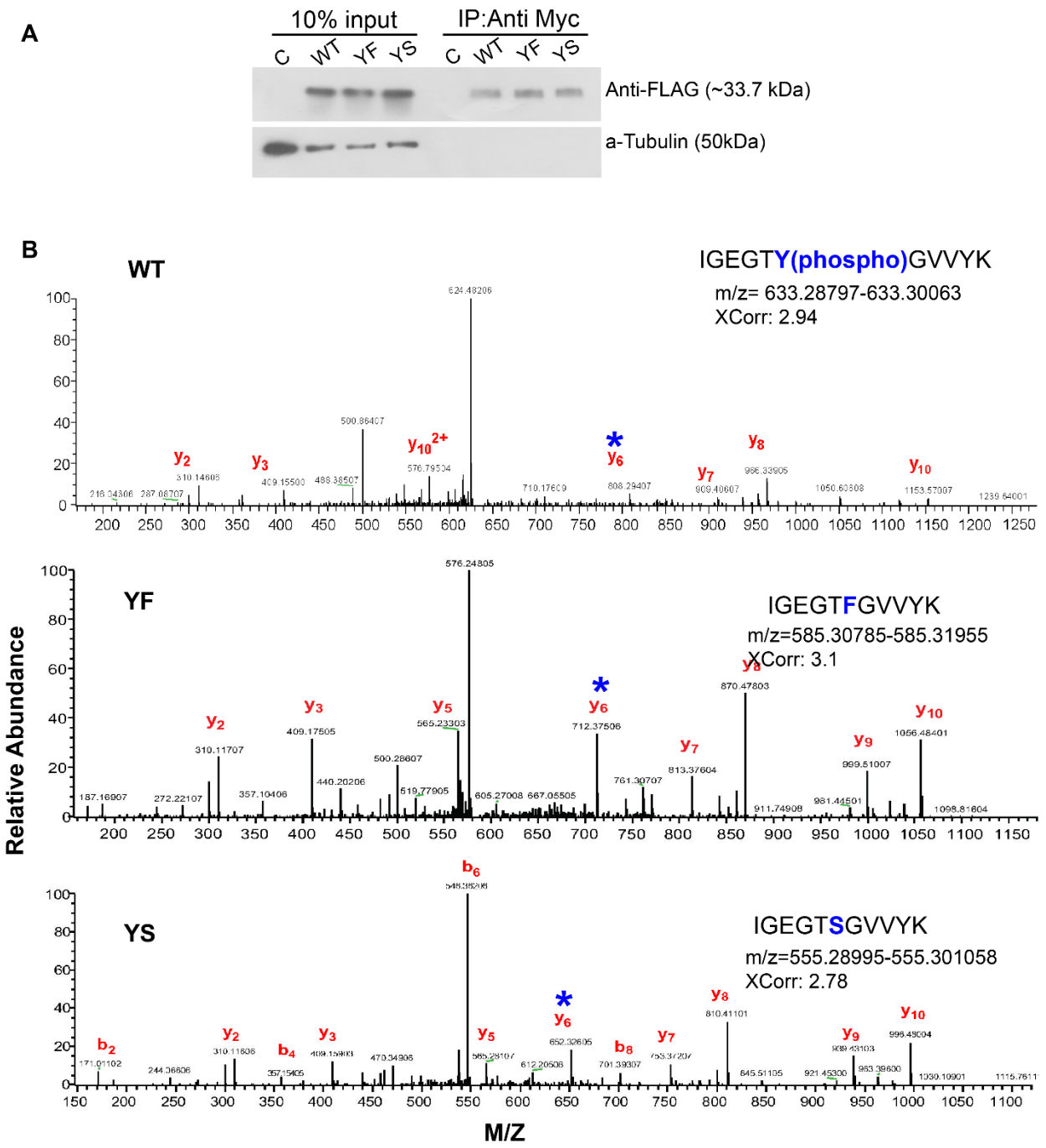

Figure S9: Phosphorylation status of Y15 and S15 residues in CDK2. (A) Western blot showing overexpression and immunoprecipitation of WT and mutant CDK2. FLAG Myc-CDK2 cDNA was transfected into HEK293T cells, then immunoprecipitated and probed with anti-Myc and anti-FLAG antibodies, respectively. (B) MS/MS spectra showing phosphorylation at the Y15 residue only (WT condition), while phosphorylation is not detected in protein immunoprecipitated from cells bearing F15 or S15 substitution transgenes.

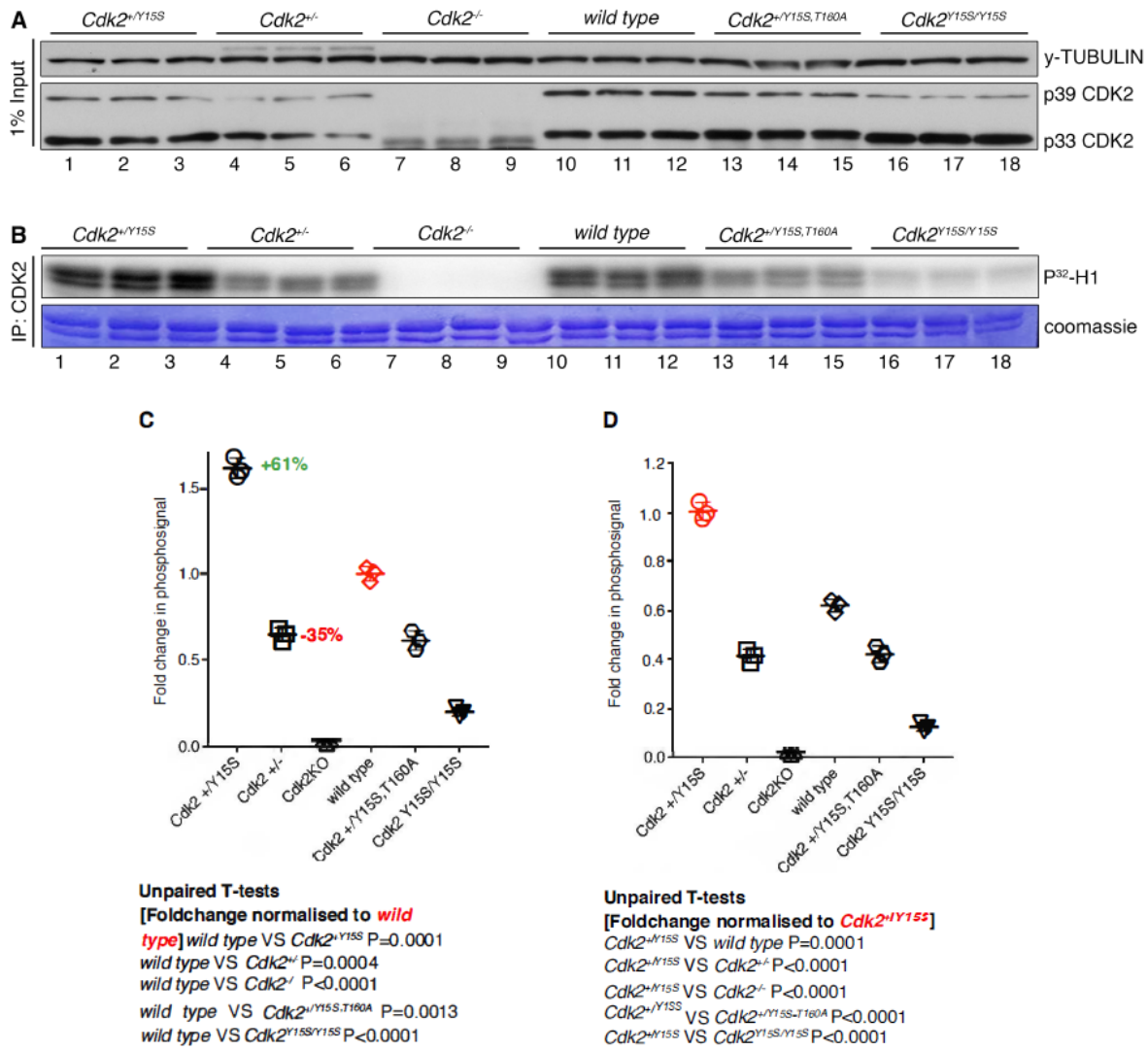

Figure S10. Quantification of CDK2 kinase activity in spleens of Cdk2 mutant mice. (A) Western blots of 1 % (1 0ug) of total protein lysate used for immunoprecipitation (input). lmmunoblotting was performed using the specified antibodies on whole spleen lysates of Cdk2+/Y15S (1-3),Cdk2+/- (4-6),Cdk2-/- (7-9), wild type (10-12), Cdk2+/Y15S-T160A (13-15) or Cdk2Y15S/Y15S (16-18) mice. All spleens were extracted from 10 days old mice. (B) Kinase assays of CDK2-immunoprecipitates from spleen lysates shown in panel A. CDK2 immunoprecipitation was performed against 1 mg of whole spleen lysate using 1 ug of anti-CDK2 antibody. The fold change in phosphosignal for each sample was calculated relative to the mean wildtype (C) or mean Cdk2+/Y15S (D) as displayed. Statistical tests were performed via unpaired-t test between each genotype to determine whether fold change was significant.

REFERENCES

Artegiani, B., Lindemann, D. and Calegari, F. (2011). Overexpression of cdk4 and cyclinD1 triggers greater expansion of neural stem cells in the adult mouse brain. *J. Exp. Med.* 208, 937–948.

Bellve, A., Cavicchia, J., Millette, C., O'Brien, D., Bhatnagar, Y. and Dym, M. (1977). Spermatogenic cells of the prepubertal mouse. Isolation and morphological characterization. *J. Cell. Biol.* 74, 68–85.

Berthet, C., Aleem, E., Coppola, V., Tessarollo, L. and Kaldis, P. (2003). Cdk2 knockout mice are viable. *Curr. Biol.* 13, 1775–1785.

Bhang, D. H., Kim, B.-J., Kim, B. G., Schadler, K., Baek, K.-H., Kim, Y. H., Hsiao, W., Ding, B.-S., Rafii, S., Weiss, M. J., et al. (2018). Testicular endothelial cells are a critical population in the germline stem cell niche. *Nat. Commun.* 9, 4379.

Buaas, F. W., Kirsh, A. L., Sharma, M., McLean, D. J., Morris, J. L., Griswold, M. D., de Rooij, D. G. and Braun, R. E. (2004). Plzf is required in adult male germ cells for stem cell self-renewal. *Nat. Genet.* 36, 647–652.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420.

Chakraborty, P., Buaas, F. W., Sharma, M., Snyder, E., de Rooij, D. G. and Braun, R. E. (2014a). LIN28A marks the spermatogonial progenitor population and regulates its cyclic expansion. *Stem Cells* 32, 860–873.

Chakraborty, P., William Buaas, F., Sharma, M., Smith, B. E., Greenlee, A. R., Eacker, S. M. and Braun, R. E. (2014b). Androgen-dependent sertoli cell tight junction remodeling is mediated by multiple tight junction components. *Mol. Endocrinol.* 28, 1055–1072.

Chauhan, S., Diril, M. K., Lee, J. H. S., Bisteau, X., Manoharan, V., Adhikari, D., Ratnacaram, C. K., Janela, B., Noffke, J., Ginhoux, F., et al. (2016). Cdk2 catalytic activity is essential for meiotic cell division in vivo. *Biochem. J.* 473, 2783–2798.

Clement, T. M., Inselman, A. L., Goulding, E. H., Willis, W. D. and Eddy, E. M. (2015). Disrupting cyclin dependent kinase 1 in spermatocytes causes late meiotic arrest and infertility in mice. *Biol. Reprod.* 93, 137.

Costoya, J. A., Hobbs, R. M., Barna, M., Cattoretti, G., Manova, K., Sukhwani, M., Orwig, K. E., Wolgemuth, D. J. and Pandolfi, P. P. (2004). Essential role of Plzf in maintenance of spermatogonial stem cells. *Nat. Genet.* 36, 653–659.

Cuijpers, S. A. G. and Vertegaal, A. C. O. (2018). Guiding Mitotic Progression by Crosstalk between Post-translational Modifications. *Trends Biochem. Sci.* 43, 251–268.

Culty, M. (2013). Gonocytes, from the fifties to the present: is there a reason to change the name? *Biol. Reprod.* 89, 46.

Dalton, S. (2015). Linking the cell cycle to cell fate decisions. *Trends Cell Biol.* 25, 592–600.

de Rooij, D. G. (2017). The nature and dynamics of spermatogonial stem cells. *Development* 144, 3022–3030.

E. F. Oakberg (1971). Spermatogonial stem cell renewal in the mouse and timing of stages of the cycle of the seminiferous epithelium. *Anat. Rec.* 169, 515–531.

Endo, T., Romer, K. A., Anderson, E. L., Baltus, A. E., de Rooij, D. G. and Page, D. C.

(2015). Periodic retinoic acid-STRA8 signaling intersects with periodic germ-cell competencies to regulate spermatogenesis. *Proc. Natl. Acad. Sci. USA* 112, E2347-56.
Gaytan, F., Sangiao-Alvarellos, S., Manfredi-Lozano, M., García-Galiano, D., Ruiz-Pino, F., Romero-Ruiz, A., León, S., Morales, C., Cordido, F., Pinilla, L., et al. (2013). Distinct expression patterns predict differential roles of the miRNA-binding proteins, Lin28 and Lin28b, in the mouse testis: studies during postnatal development and in a model of hypogonadotropic hypogonadism. *Endocrinology* 154, 1321–1336.
Ginsburg, M., Snow, M. H. and McLaren, A. (1990). Primordial germ cells in the mouse embryo during gastrulation. *Development* 110, 521–528.
Goertz, M. J., Wu, Z., Gallardo, T. D., Hamra, F. K. and Castrillon, D. H. (2011). Foxo1 is required in mouse spermatogonial stem cells for their maintenance and the initiation of spermatogenesis. *J. Clin. Invest.* 121, 3456–3466.
Gu, Y., Rosenblatt, J. and Morgan, D. O. (1992). Cell cycle regulation of CDK2 activity by phosphorylation of Thr160 and Tyr15. *EMBO J.* 11, 3995–4005.
Hamer, G. and de Rooij, D. G. (2018). Mutations causing specific arrests in the development of mouse primordial germ cells and gonocytes. *Biol. Reprod.* 99, 75–86.
Hara, K., Nakagawa, T., Enomoto, H., Suzuki, M., Yamamoto, M., Simons, B. D. and Yoshida, S. (2014). Mouse spermatogenic stem cells continually interconvert between equipotent singly isolated and syncytial states. *Cell Stem Cell* 14, 658–672.
Hobbs, R. M., Seandel, M., Falciatori, I., Rafii, S. and Pandolfi, P. P. (2010). Plzf regulates germline progenitor self-renewal by opposing mTORC1. *Cell* 142, 468–479.
Hofmann, M.-C. (2008). Gdnf signaling pathways within the mammalian spermatogonial stem cell niche. *Mol. Cell. Endocrinol.* 288, 95–103.
Hu, Y.-C., de Rooij, D. G. and Page, D. C. (2013). Tumor suppressor gene Rb is required for self-renewal of spermatogonial stem cells in mice. *Proc. Natl. Acad. Sci. USA* 110, 12685–12690.
Huang, H., Regan, K. M., Lou, Z., Chen, J. and Tindall, D. J. (2006). CDK2-dependent phosphorylation of FOXO1 as an apoptotic response to DNA damage. *Science* 314, 294–297.
Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M. and Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542.
Huckins, C. (1971a). The spermatogonial stem cell population in adult rats. I. Their morphology, proliferation and maturation. *Anat Rec* 169, 533–557.
Huckins, C. (1971b). The spermatogonial stem cell population in adult rats. 3. Evidence for a long-cycling population. *Cell Tissue Kinet.* 4, 335–349.
Hughes, B. T., Sidorova, J., Swanger, J., Monnat, R. J. and Clurman, B. E. (2013). Essential role for Cdk2 inhibitory phosphorylation during replication stress revealed by a human Cdk2 knockin mutation. *Proc. Natl. Acad. Sci. USA* 110, 8954–8959.
Johnston, D. S., Wright, W. W., Dicandeloro, P., Wilson, E., Kopf, G. S. and Jelinsky, S. A. (2008). Stage-specific gene expression is a fundamental characteristic of rat spermatogenic cells and Sertoli cells. *Proc. Natl. Acad. Sci. USA* 105, 8315–8320.
Kaldis, P. (1999). The cdk-activating kinase (CAK): from yeast to mammals. *Cell Mol. Life Sci.* 55, 284–296.

Kang, H. S., Chen, L.-Y., Lichti-Kaiser, K., Liao, G., Gerrish, K., Bortner, C. D., Yao, H. H.-C., Eddy, E. M. and Jetten, A. M. (2016). Transcription factor GLIS3: A new and critical regulator of postnatal stages of mouse spermatogenesis. *Stem Cells* 34, 2772–2783.
Klein, A. M., Nakagawa, T., Ichikawa, R., Yoshida, S. and Simons, B. D. (2010). Mouse germ line stem cells undergo rapid and stochastic turnover. *Cell Stem Cell* 7, 214–224.
Krasinska, L., Cot, E. and Fisher, D. (2008). Selective chemical inhibition as a tool to study Cdk1 and Cdk2 functions in the cell cycle. *Cell Cycle* 7, 1702–1708.
La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498.
Lange, C., Huttner, W. B. and Calegari, F. (2009). Cdk4/cyclinD1 overexpression in neural stem cells shortens G1, delays neurogenesis, and promotes the generation and expansion of basal progenitors. *Cell Stem Cell* 5, 320–331.
Law, N. C., Oatley, M. J. and Oatley, J. M. (2019). Developmental kinetics and transcriptome dynamics of stem cell specification in the spermatogenic lineage. *Nat. Commun.* 10, 2787.
Liao, J., Ng, S. H., Tu, J., Luk, A. C. S., Qian, Y., Tang, N. L. S., Feng, B., Chan, W.-Y., Fouchet, P. and Lee, T.-L. (2017). Single-cell RNA-Seq Resolves Cellular Heterogeneity and Transcriptional Dynamics during Spermatogonia Stem Cells Establishment and Differentiation. *BioRxiv*.
Lim, S. and Kaldis, P. (2012). Loss of Cdk2 and Cdk4 induces a switch from proliferation to differentiation in neural stem cells. *Stem Cells* 30, 1509–1520.
Lim, S. and Kaldis, P. (2013). Cdks, cyclins and CKIs: roles beyond cell cycle regulation. *Development* 140, 3079–3093.
Lim, S., Bhinge, A., Bragado Alonso, S., Aksoy, I., Aprea, J., Cheok, C. F., Calegari, F., Stanton, L. W. and Kaldis, P. (2017). Cyclin-Dependent Kinase-Dependent Phosphorylation of Sox2 at Serine 39 Regulates Neurogenesis. *Mol. Cell. Biol.* 37,.
Liu, D., Matzuk, M. M., Sung, W. K., Guo, Q., Wang, P. and Wolgemuth, D. J. (1998). Cyclin A1 is required for meiosis in the male mouse. *Nat. Genet.* 20, 377–380.
Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214.
Martinerie, L., Manterola, M., Chung, S. S. W., Panigrahi, S. K., Weisbach, M., Vasileva, A., Geng, Y., Sicinski, P. and Wolgemuth, D. J. (2014). Mammalian E-type cyclins control chromosome pairing, telomere stability and CDK2 localization in male meiosis. *PLoS Genet.* 10, e1004165.
McConnell, M. J., Chevallier, N., Berkofsky-Fessler, W., Giltnane, J. M., Malani, R. B., Staudt, L. M. and Licht, J. D. (2003). Growth suppression by acute promyelocytic leukemia-associated protein PLZF is mediated by repression of c-myc expression. *Mol. Cell. Biol.* 23, 9375–9388.
McCurdy, S. R., Pacal, M., Ahmad, M. and Bremner, R. (2017). A CDK2 activity signature predicts outcome in CDK2-low cancers. *Oncogene* 36, 2491–2502.
McGinnis, C. S., Murrow, L. M. and Gartner, Z. J. (2018). DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *BioRxiv*.
Meng, X., Lindahl, M., Hyvönen, M. E., Parvinen, M., de Rooij, D. G., Hess, M. W.,

Raatikainen-Ahokas, A., Sainio, K., Rauvala, H., Lakso, M., et al. (2000). Regulation of cell fate decision of undifferentiated spermatogonia by GDNF. *Science* 287, 1489–1493.
Morgan, D. O. (1995). Principles of CDK regulation. *Nature* 374, 131–134.
Morris, L., Allen, K. E. and La Thangue, N. B. (2000). Regulation of E2F transcription by cyclin E-Cdk2 kinase mediated through p300/CBP co-activators. *Nat. Cell Biol.* 2, 232–239.
Nagano, R., Tabata, S., Nakanishi, Y., Ohsako, S., Kurohmaru, M. and Hayashi, Y. (2000). Reproliferation and relocation of mouse male germ cells (gonocytes) during prespermatogenesis. *Anat Rec* 258, 210–220.
Nakagawa, T., Sharma, M., Nabeshima, Y., Braun, R. E. and Yoshida, S. (2010). Functional hierarchy and reversibility within the murine spermatogenic stem cell compartment. *Science* 328, 62–67.
Oatley, J. M. and Brinster, R. L. (2012). The germline stem cell niche unit in mammalian testes. *Physiol. Rev.* 92, 577–595.
Ohbo, K., Yoshida, S., Ohmura, M., Ohneda, O., Ogawa, T., Tsuchiya, H., Kuwana, T., Kehler, J., Abe, K., Schöler, H. R., et al. (2003). Identification and characterization of stem cells in prepubertal spermatogenesis in mice. *Dev. Biol.* 258, 209–225.
Ohmura, M., Yoshida, S., Ide, Y., Nagamatsu, G., Suda, T. and Ohbo, K. (2004). Spatial analysis of germ stem cell development in Oct-4/EGFP transgenic mice. *Arch Histol Cytol* 67, 285–296.
Ortega, S., Prieto, I., Odajima, J., Martín, A., Dubus, P., Sotillo, R., Barbero, J. L., Malumbres, M. and Barbacid, M. (2003). Cyclin-dependent kinase 2 is essential for meiosis but not for mitotic cell division in mice. *Nat. Genet.* 35, 25–31.
Percharde, M., Wong, P. and Ramalho-Santos, M. (2017). Global hypertranscription in the mouse embryonic germline. *Cell Rep.* 19, 1987–1996.
Pietras, E. M., Warr, M. R. and Passegué, E. (2011). Cell cycle regulation in hematopoietic stem cells. *J. Cell Biol.* 195, 709–720.
Pui, H. P. and Saga, Y. (2017). Gonocytes-to-spermatogonia transition initiates prior to birth in murine testes and it requires FGF signaling. *Mech. Dev.* 144, 125–139.
Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309–315.
Ravnik, S. E. and Wolgemuth, D. J. (1999). Regulation of meiosis during mammalian spermatogenesis: the A-type cyclins and their associated cyclin-dependent kinases are differentially expressed in the germ-cell lineage. *Dev. Biol.* 207, 408–418.
Rubin, S. M. (2013). Deciphering the retinoblastoma protein phosphorylation code. *Trends Biochem. Sci.* 38, 12–19.
Sasaki, H. and Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat. Rev. Genet.* 9, 129–140.
Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502.
Satyanarayana, A. and Kaldis, P. (2009). Mammalian cell-cycle regulation: several Cdks, numerous cyclins and diverse compensatory mechanisms. *Oncogene* 28, 2925–2939.
Savitt, J., Singh, D., Zhang, C., Chen, L.-C., Folmer, J., Shokat, K. M. and Wright, W. W. (2012). The in vivo response of stem and other undifferentiated spermatogonia to the

reversible inhibition of glial cell line-derived neurotrophic factor signaling in the adult. *Stem Cells* 30, 732–740.
Sharma, M., Srivastava, A., Fairfield, H. E., Bergstrom, D., Flynn, W. F. and Braun, R. E. (2019). Identification of EOMES-expressing spermatogonial stem cells and their regulation by PLZF. *Elife* 8,.
Singh, P. and Schimenti, J. C. (2015). The genetics of human infertility by functional interrogation of SNPs in mice. *Proc. Natl. Acad. Sci. USA* 112, 10431–10436.
Singh, P., Schimenti, J. C. and Bolcun-Filas, E. (2015). A mouse geneticist's practical guide to CRISPR applications. *Genetics* 199, 1–15.
Song, H.-W., Bettegowda, A., Lake, B. B., Zhao, A. H., Skarbrevik, D., Babajanian, E., Sukhwani, M., Shum, E. Y., Phan, M. H., Plank, T.-D. M., et al. (2016). The homeobox transcription factor RHOX10 drives mouse spermatogonial stem cell establishment. *Cell Rep.* 17, 149–164.
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102, 15545–15550.
Szabó, P. E., Hübner, K., Schöler, H. and Mann, J. R. (2002). Allele-specific expression of imprinted genes in mouse migratory primordial germ cells. *Mech. Dev.* 115, 157–160.
Szmyd, R., Niska-Blakie, J., Diril, M. K., Renck Nunes, P., Tzelepis, K., Lacroix, A., van Hul, N., Deng, L.-W., Matos, J., Dreesen, O., et al. (2019). Premature activation of Cdk1 leads to mitotic events in S phase and embryonic lethality. *Oncogene* 38, 998–1018.
Tam, P. P. and Snow, M. H. (1981). Proliferation and migration of primordial germ cells during compensatory growth in mouse embryos. *J Embryol Exp Morphol* 64, 133–147.
Tang, J.-X., Li, J., Cheng, J.-M., Hu, B., Sun, T.-C., Li, X.-Y., Batool, A., Wang, Z.-P., Wang, X.-X., Deng, S.-L., et al. (2017). Requirement for CCNB1 in mouse spermatogenesis. *Cell Death Dis.* 8, e3142.
Thomas, C. J., Cleland, T. P., Zhang, S., Gundberg, C. M. and Vashishth, D. (2017). Identification and characterization of glycation adducts on osteocalcin. *Anal. Biochem.* 525, 46–53.
Toyooka, Y., Tsunekawa, N., Takahashi, Y., Matsui, Y., Satoh, M. and Noce, T. (2000). Expression and intracellular localization of mouse Vasa-homologue protein during germ cell development. *Mech. Dev.* 93, 139–149.
Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386.
Varshney, G. K., Pei, W., LaFave, M. C., Idol, J., Xu, L., Gallardo, V., Carrington, B., Bishop, K., Jones, M., Li, M., et al. (2015). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res.* 25, 1030–1042.
Viera, A., Rufas, J. S., Martínez, I., Barbero, J. L., Ortega, S. and Suja, J. A. (2009). CDK2 is required for proper homologous pairing, recombination and sex-body formation during male mouse meiosis. *J. Cell Sci.* 122, 2149–2159.
Viera, A., Alsheimer, M., Gómez, R., Berenguer, I., Ortega, S., Symonds, C. E., Santamaría,

D., Benavente, R. and Suja, J. A. (2015). CDK2 regulates nuclear envelope protein dynamics and telomere attachment in mouse meiotic prophase. *J. Cell Sci.* 128, 88–99.
Welburn, J. P. I., Tucker, J. A., Johnson, T., Lindert, L., Morgan, M., Willis, A., Noble, M. E. M. and Endicott, J. A. (2007). How tyrosine 15 phosphorylation inhibits the activity of cyclin-dependent kinase 2-cyclin A. *J. Biol. Chem.* 282, 3173–3181.
Wolgemuth, D. J. and Roberts, S. S. (2010). Regulating mitosis and meiosis in the male germ line: critical functions for cyclins. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* 365, 1653–1662.
Yang, Q.-E. and Oatley, J. M. (2014). Spermatogonial stem cell functions in physiological and pathological conditions. *Curr Top Dev Biol* 107, 235–267.
Yang, Y., Thannhauser, T. W., Li, L. and Zhang, S. (2007). Development of an integrated approach for evaluation of 2-D gel image analysis: impact of multiple proteins in single spots on comparative proteomics in conventional 2-D gel/MALDI workflow. *Electrophoresis* 28, 2080–2094.
Yang, Y., Anderson, E. and Zhang, S. (2018). Evaluation of six sample preparation procedures for qualitative and quantitative proteomics analysis of milk fat globule membrane. *Electrophoresis* 39, 2332–2339.
Yeyati, P. L., Shaknovich, R., Boterashvili, S., Li, J., Ball, H. J., Waxman, S., Nason-Burchenal, K., Dmitrovsky, E., Zelent, A. and Licht, J. D. (1999). Leukemia translocation protein PLZF inhibits cell growth and expression of cyclin A. *Oncogene* 18, 925–934.
Yoshida, S., Takakura, A., Ohbo, K., Abe, K., Wakabayashi, J., Yamamoto, M., Suda, T. and Nabeshima, Y.-I. (2004). Neurogenin3 delineates the earliest stages of spermatogenesis in the mouse testis. *Dev. Biol.* 269, 447–458.
Yoshida, S., Sukeno, M., Nakagawa, T., Ohbo, K., Nagamatsu, G., Suda, T. and Nabeshima, Y. (2006). The first round of mouse spermatogenesis is a distinctive program that lacks the self-renewing spermatogonia stage. *Development* 133, 1495–1505.
Zhao, H., Chen, X., Gurian-West, M. and Roberts, J. M. (2012). Loss of cyclin-dependent kinase 2 (CDK2) inhibitory phosphorylation in a CDK2AF knock-in mouse causes misregulation of DNA replication and centrosome duplication. *Mol. Cell. Biol.* 32, 1421–1432.
Zheng, K., Wu, X., Kaestner, K. H. and Wang, P. J. (2009). The pluripotency factor LIN28 marks undifferentiated spermatogonia in mouse. *BMC Dev. Biol.* 9, 38.
Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* 65, 631–643.e4.