Integrating genomic data to build more predictive models of biology

A thesis submitted by

Alexander D. Fine

in partial fulfillment of the requirements for the degree of

Ph.D.

in

Genetics

Tufts University

Sackler School of Graduate Biomedical Sciences

August 2019

Advisor: Gregory Carter, Ph.D.

Abstract

By measuring multiple characteristics of a biological system at once, one can get a more complete picture of how a perturbation affects that system. Single-omic measurements sometimes fail to correlate well with other genome-wide measurements, demonstrating the value of measuring a biological system at multiple levels. However, knowing which aspects are informative and how to best integrate them into a comprehensive model is a challenge that persists within the field of systems genetics. Here, we utilize induced and naturally-occurring perturbations across multiple biological systems to build more predictive models of biology across three distinct projects. For each project, we integrated a combination of genetic, epigenetic, transcriptional, and cellular measurements to better understand their role in the complex biological system, gaining a more comprehensive understanding than could be accomplished through examining any single measurement alone. First, we disrupted normal germ cell differentiation by knocking out the key meiotic protein PRDM9, which deposits epigenetic modifications throughout the genome to select sites for meiotic recombination. We integrated cellular and molecular phenotypes to determine how cytological and transcriptional programs remain coupled in case of developmental arrest. We found that these two developmental programs became uncoupled in the perturbed system, demonstrating the need to pair transcriptional analyses with classical molecular biology in order to put gene expression data within their proper context. Next, we employed genetic variation within and between the C57BL/6J and CAST/EiJ genomes to test the effect on binding affinity of variants within the binding site of the PRDM9 long zinc-finger array. We showed that, in addition to their individual effects, combinations of

certain base-pairs across the binding site of the zinc-finger had epistatic effects on binding affinity. Finally, we integrated genetic, epigenetic, and transcriptomic data from hepatocytes of nine diverse strains of mice to build a multi-omic model of gene expression. Our work quantified the inter-dependency among genetic, epigenetic, and gene expression variation across these strains. These projects used existing and novel tools for data integration and analysis to go beyond the scope of any single-omic analysis, demonstrating the power and necessity of multi-omic analyses in the field of modern genetics.

Dedication

To my loving family, who have encouraged, supported, and inspired me through this entire journey, from math games growing up, to starting on this path at Mass Academy, to my wild choice to come to JAX and Maine for this degree.

Acknowledgements

First and foremost, I would like to thank my mentor, Gregory Carter. Greg helped me when I was stuck, but pushed me when needed to grow. Greg maintains the philosophy that training is for the trainee, and this has made his mentorship unwaveringly supportive. Those who have left the Carter lab say that they never have, nor will, find as good of a boss as Greg. I learned how to be a better scientist from Greg, but I have also learned how to be a better boss. I am sad to move on from his formal mentorship, but I look forward to using what he has taught me as I enter my next adventure.

Second, I would like to thank my thesis advisory committee. To Jen Trowbridge, thank you for being an incredible chair to my advisory committee. Beyond your formal role, your willingness and encouragement to meet outside of our formal meetings kept me sane through the ups and downs of this process. To Mary Ann Handel, thank you for your co-mentorship. While not formally being a part of it, I have always felt welcome in the Handel lab, and your guidance has been a key part of my achievement of balance between computational power and biological interpretation. Steve Munger, you've been my harshest critic and my fiercest ally. You have never let me get away with being lazy, but you never let anyone else overlook my hard work. Amy Yee, you have kept my thinking global and have taught me how to take a step back and look at how my work fits into the broader picture. This has been one the most difficult and important lessons to learn in my training process, and I am grateful for your dedication to teaching it to me.

Next, I want to thank everyone that I have had the pleasure of working with at JAX. First, the people who first welcomed me into the Carter lab – Robyn Ball, Anna Tyler, Xulong Wang, and Bo Ji. In particular, Robyn, you have been a friend and a role

model that I will forever look up to. Anna, your constant support, collaboration, and friendship has meant the world to me. To all of the Carter lab members who have joined as I have worked here, thank you. Particularly, I thank Christoph Preuss and Cai John. You two have been the best friends I could ask for and were the ones to make this place feel like home. The Howell and Trowbridge labs have been incredibly welcoming spaces to me. The Handel, Baker, and Paigen labs have been immensively supportive collaborators for my work. In particular, Travis Kent, Yasuhiro Fujiwara, Chris Baker, and Catrina Spruce have all been essential for my success.

Next, I would like to thank the members of my graduate program. First, I thank Liz Adkins, Leah Graham, Sneha Borikar, and Qiming Wang, who welcomed me into our brand-new program and created what it is today. I would like to thank the rest of the genetics students, in particular Kate Foley, Candice Byers, Ashlee Junier, and Daniel Heller, who have been amazing friends to me. Last, but not least, I would like to thank my little cohort, Alex Neil and Jaymes Farrell. We got through the worst bits of this journey together and I've been lucky to have you around to celebrate the best bits of it all.

I would like to thank the rest of my support system, at JAX and beyond. Carrie Cowan, you are the most underappreciated staff member here at JAX. You work tirelessly to support the trainees here, from organizing our programs to righting the mistakes we make. Paige Martin, Eraj Khokhar, Sarah Neuner, Catherine Kaczorowski, Galen Squires, Becca Bell, Dylan Garceau, and Kevin Hayes have been a part of making the JAX community so wonderful for me here. I want to thank the support from long-time friends and mentors. Hannah Michlmayr, Catherine LaPlant, Joshua Mauro, Emily Durante, Alyssa Antonopoulos, Lauren Doyle, Billy Gawron, Lucy Sinacola, and Liz

Johnson, you have all been incredible friends to me and I have appreciated your everlasting support. To my previous mentors, Jesse Mager and Chelsea Marcho, you inspired me to pursue a career in research and changed my life for the better.

I would next like to thank my family. In particular, my parents, Anne Orlando Fine and Doug Fine, have been more supportive than I could have wished for. Their love and encouragement throughout graduate school has pushed me to continue on when I felt like there was no end in sight. Mom, you have taught me how to relate to people and how to communicate with them. This has been an important part of my life, including as a scientist, and it contributes to my success in many ways. Dad, you are the reason that I am a scientist. You raised me to be scientifically inquisitive, and you have inspired me to think globally about how I can change the world. To my grandmother, Nancy, from a young age, we played math games and solved puzzles, and I use those skills to this day. My brothers, Eliot and Charlie, have visited me in Maine and welcomed me back home whenever I am there. It has been amazing to have been able to share my journey with you, and I am excited to see where we all go next.

Thank you to the others not named to whom I am grateful for support. I have received so much encouragement and help through this process. Thank you to everyone who has guided me along the way and has contributed to my success.

Table of Contents

List of Figures

List of Copyrighted Materials

Fine, A. D., Ball, R. L., Fujiwara, Y., Handel, M. A., & Carter, G. W. (2019). Uncoupling of transcriptomic and cytological differentiation in mouse spermatocytes with impaired meiosis. Mol Biol Cell, 30(5), 717-728. doi:10.1091/mbc.E18-10-0681

List of Abbreviations

129 – 129S1/SvlmJ
AJ – A/J
B6 – C57BL/6J
BXD – B6 and D2 recombinant inbred
CAST – CAST/EiJ
CC – collaborative cross
ChIP – chromatin immunoprecipitation
ChIP-seq – ChIP sequencing
D – Diplotene
DBA – DBA/2J
DEG – differentially expressed gene
DHS – DNase I hypersensitivity site
DNase I – Deoxyribonuclease I
DO – diversity outbred
dpi – days post injection
dpp – days post-partum
DSB – double-strand break
EL – Early Leptotene
ENU – N-Ethyl-N-Nitrosourea
EP – Early Pachytene
eQTL – expression QTL
FDR – false discovery rate
GO – Gene Ontology
GWAS – genome-wide association study
H3K27ac – Histone 3, lysine 27 acetylation
H3K27me3 – Histone 3, lysine 27 trimethylation
H3K4me1 – Histone 3, lysine 4 monomethylation
H3K4me3 – Histone 3, lysine 4 trimethylation
iRF – iterative Random Forest
KRAB – Krüppel associated box
LL – Late Leptotene
LP – Late Pachytene
MGW – minor groove width
NOD – NOD/ShiLtJ
NZO – NZO/H1LtJ
P-Like – Pachytene-like
PCA – Principal Component Analysis
piRNA – PIWI-interacting RNA
PL – Preleptotene
PMCA – Permutation-based Maximum Covariance Analysis
PWK – PWK/PhJ
QTL – quantitative trait loci
RNA-Seq – RNA sequencing
scRNA-Seq – single-cell RNA sequencing

SNP – single nucleotide polymorphism
Sp – Spermatogonia
TAD – topologically associated domain
TF – transcription factor
TPM – transcripts per million
TSS – transcription start site
WSB – WSB/EiJ
WT – wild-type
Z – Zygotene
ZF – zinc-finger

CHAPTER 1: Introduction

1.1. Systems genetics

1.1.1. Rationale

In this research, we applied systems genetics approaches to several different biological questions. We built computational models of genetic systems based on multiple genome-wide ('omic) measurements made on perturbed systems. We found that a comprehensive and integrated analysis is necessary to fully understand the cellular and molecular effects in a complex system, as there are many interacting and interdependent biological levels that should be considered (i.e. epigenetics and gene expression). Moreover, models developed based on multi-omic measurements of a system have improved predictive power and reproducibility.

1.1.2. Origin and advancement of the field gives rise to integrative genomic analyses

Systems biology comprises the perturbation of a biological system, the integration of multi-omic measurements, and the utilization of these data to mathematically model the biological system (Chuang et al., 2010; Ideker et al., 2001; Kitano, 2002a, b). However, there is no single common definition for systems biology, despite it not being a new concept (Baliga et al., 2017; Breitling, 2010; Kirschner, 2005; Likic et al., 2010). Systems biology has been getting mainstream attention for over 15 years (Ideker et al., 2001; Kitano, 2002a, b). Since then, systems biology has grown in popularity across subdisciplines for its ability to facilitate more biologically accurate model building (Likic et al., 2010), yet many questions remain open in the field (Baliga et al., 2017).

The field of genetics has frequently used a systems approach to uncovering the effects of genetic variation. Moving beyond the historically important single-gene, reductionist models, the field of genetics has recently opened up to be heavily based in system-wide measurements. Interactions among genetic elements have long been established, and thus studying the effect of a perturbation system-wide is somewhat intuitive. This has been accomplished most frequently by performing genomic measurements – genetic, transcriptomic, proteomic, etc – that survey the entire genome, rather than measuring only genes or phenotypes of interest. These efforts were propelled in part by completion of the Human Genome Project just as systems biology grew in popularity (Lander et al., 2001; Venter et al., 2001). At that time, the ability to measure transcript abundance at a genome-wide scale had already been established with microarrays (Schena et al., 1995), and RNA-sequencing (RNA-Seq) emerged in the years following (Mortazavi et al., 2008; Nagalakshmi et al., 2008). Further, the development of high-throughput techniques to measure chromatin state and protein abundance (Barski et al., 2007; Buenrostro et al., 2013; Johnson et al., 2007a; Kim et al., 2014; Mikkelsen et al., 2007) added to the ability to fully annotate a genetic system. Systems approaches have become a common practice and popular solution to the complicated nature of genetic regulation of phenotype.

Thus multi-omic methodologies are widely used by geneticists to study systems biology. Perturbations can impact a system at genetic, epigenetic, expression, and phenotype levels simultaneously and in non-redundant ways. A genetic mutation may alter translation rate without affecting transcript abundance, which would be hard to infer without 'omic measurements of both data types. Further, incorporating multi-omic

measurements increases the interpretability of a single-omic measurement. For example, the effect of distal epigenetic changes can be hard to interpret without the corresponding gene expression data. These complexities make it increasingly valuable to thoughtfully integrate, analyze, and visualize multiple data types. By strengthening our capacity to process data and increasing our power to interpret it, we posit that a systems approach of integrating multiple 'omic data types would significantly improve our ability to understand and model biology.

1.1.3. Gene regulatory complexity and lack of concordance between 'omic measurements

Models of gene regulation have frequently been based on a linear and unidirectional relationship from genetics to gene expression; however, in the past fifty years, there have been numerous examples of gene expression being both a complex and non-linear process (Berthelot et al., 2018; Chen et al., 2018; Holter et al., 2000; Jansen et al., 1995; Strambio-De-Castillia et al., 2010). Many intricacies of the regulation of transcription, translation, and product stability have been ignored due to a limited ability to survey them. At the most basic level, transcription is not a one-to-one process with DNA, as variance in the copies of a gene can, but does not always, correspond to variance in transcript abundance (Coate and Doyle, 2010; Henrichsen et al., 2009; Zhang et al., 2010). Moreover, DNA content, chromatin structure, DNA methylation, and histone modifications all contribute to the regulation of transcription (Bannister and Kouzarides, 2011; de Laat and Duboule, 2013; Dixon et al., 2012; Nora et al., 2012). Such complexity is not limited to the gene body, as many proximal and distal regulatory elements, like enhancers, have been found to be functionally relevant and regulate gene

expression, further complicated our understanding of gene regulation (Bulger and Groudine, 2011; Dao et al., 2017; Maston et al., 2006; Shlyueva et al., 2014; Stadhouders et al., 2012; Wamstad et al., 2014).

Translational regulation is also not as simple as a one-to-one relationship between RNA and protein. Recently, multiple studies have shown that transcript abundance is not predictive of protein abundance (Battle et al., 2015; Chick et al., 2016; Goncalves et al., 2017; Wu et al., 2014). Additionally. there are entire classes of genes that produce noncoding RNA, which are not translated to protein, further complicating the RNA-to-protein relationship (Eddy, 2001). Proteins, as well as noncoding RNA, can alter gene expression. Noncoding RNAs play roles in transcriptional and translational regulation (Cech and Steitz, 2014; Lee, 2012; Ponting et al., 2009). Proteins can act as transcriptional regulators to alter gene expression, in part by regulating epigenetic modifications and genome organization (Chen and Dent, 2014; Lambert et al., 2018; Lee and Young, 2013; Spitz and Furlong, 2012).

Proteins can do more than affect gene expression, they can alter genetics as well. Proteins can directly, and indirectly through epigenetic modifications, create DNA damage, for example, during meiotic recombination (Baker et al., 2014; Brick et al., 2012; Hayashi et al., 2005; Mahadevaiah et al., 2001; Parvanov et al., 2010). The repair of DNA damage is an error-prone process, potentially leading to impactful genetic changes (Brick et al., 2012; Rodgers and McVey, 2016; Tubbs and Nussenzweig, 2017). The regulation of gene expression is incredibly complicated and therefore, when building computation models in our study, we have focused on the complex relationships of gene regulation and gene function.

1.1.4. Relevance to Experimental Reproducibility

There is an ongoing discussion of a reproducibility crisis – the observation that many scientific findings cannot be reproduced in other institutions, other laboratories or even by other researchers within a laboratory. It is quite possible that the complexity of the biological systems we study, but corresponding lack of complexity in the models that we use to describe them, has contributed to the reproducibility crisis in science. One survey of scientists revealed that 90% of the survey participants perceived there to be a reproducibility crisis in science (Baker, 2016). While there is evidence for and against the validity of this growing concern (Fanelli, 2018; Lithgow et al., 2017), it is worth taking time to examine how we can better design, annotate, and report our science to increase reproducibility within the field.

One hypothesis for the cause of this failure to reproduce findings is that the hyper-controlled research conditions used by a single researcher uses that ends up being nearly impossible to reproduce by others elsewhere. Therefore, while one result may be apparent under one researcher's hands, on one strain of mice, using one measurement technique, any variation to that methodology can produce different results. But by broadening the scope of our data, by diversity or multi-omic analyses, we may find greater consistency from study to study. First, a finding that is consistent across multi-omic measurements and diverse systems may be more robust and thus, more likely to be reproduced, as it would not necessarily rely on a specific methodology to be apparent. Second, when incongruent results are observed across 'omic measurements, those incongruencies may help to form a novel hypothesis about the system. For example, gene expression that is upregulated at the transcript level but not the protein level may reflect a stochastic

difference in transcript abundance and therefore be less likely to be biologically relevant. Such a result would not be interpreted as producing a differential effect on cell function, thus guarding the researcher from making claims that would not be apparent by another form of analysis. Therefore, I propose findings observed across multiple levels of measurements – using a systems approach – will be more reproducable in other research condition. While systems biology alone will not solve a reproducibility crisis in science, it could set guidelines that would help to limit the apparent pervasiveness of repruducability issues.

1.1.5. Systems Approach to Model Development

While not linearly related, the genetic, epigenetic, and gene expression levels of molucalar function are interrelated by a complex network of interactions at each biological level that cumulates in an outward phenotype (Figure 1.1). In my work, I have studied each of these biological levels to understand the connections among them. By collecting multi-omic data, I have examined the relationships among genetics, epigenetics, gene expression, and phenotype. My results suggest that a genetic model that does not integrate multi-omic datasets may be incomplete and may fail to be predictive, providing further evidence that systems biology is a necessary methodology for modern genetics.

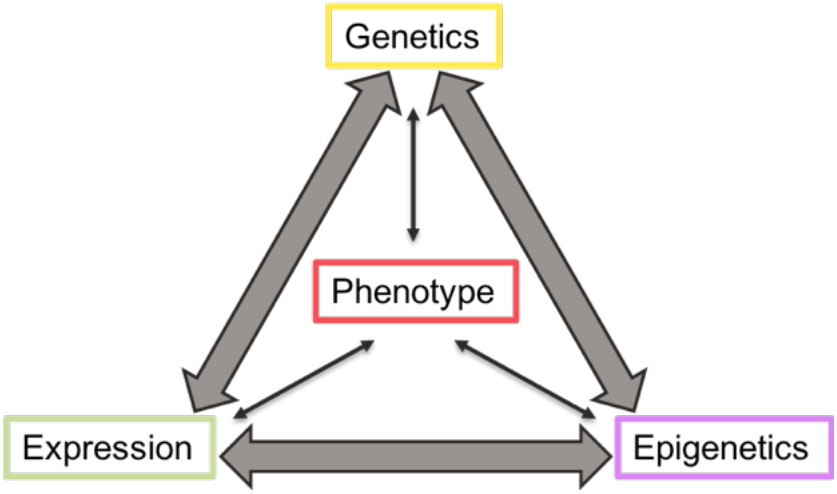

Figure 1.1. Genetics, epigenetics, and expression are all interrelated.
Genetics, epigenetics, and gene expression levels of function all affect each other, demanding that they should not be studied in isolation.

1.2. Modern Genomics for Systems Genetics

Genomics, the study the entire genome, rather than individual genes, is the toolset we have used to study systems genetics (Hawkins et al., 2010). Since the term was first coined in 1986 by geneticist Tom Rodrick of the Jackson Laboratory, it has grown to involve the study of many levels of biological systems, from genetics, to epigenetics, to gene expression (Figure 1.2) (Hawkins et al., 2010). Not only can such 'omic measurements be conducted genome-wide, but recent work has also demonstrated that these measurements can be made at the single-cell level (Buenrostro et al., 2015; Chen et al., 2018; Hwang et al., 2018; Macosko et al., 2015; Zheng et al., 2019). The amount of information that can now be collected from a single tissue – or even a single cell – has led to a new challenge for the field of genetics: how do we thoughtfully process 'omic data across multiple data types and the genome?

1.2.1. Increasing Capabilities of Genetic Analyses

There are multiple ways to measure an individual's genetics that are frequently

used in both humans and model organisms. For a broad overview of allelic differences across the genome, genotyping chip arrays exist that target specific known sites and assess their genetic sequence (LaFramboise, 2009). These markers are strategically spread throughout the genome, and are accurate in their measurement of local alleles, but imprecise as the arrays fail to identify variants not directly represented. Therefore, genetic sequencing, while more costly, is often the most favorable method for complete

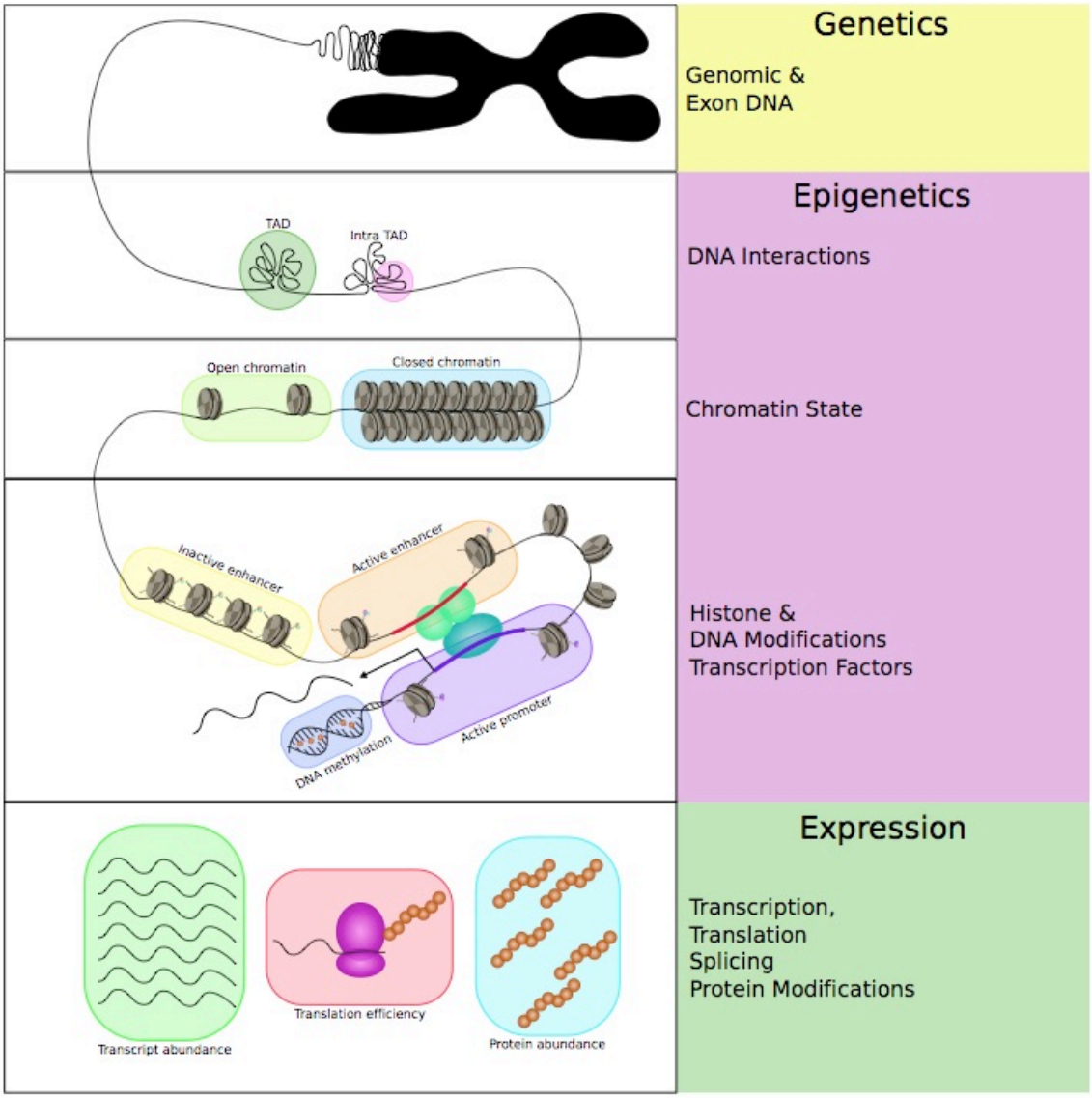

Figure 1.2. Various 'omic measurements across genetics, epigenetics, and gene expression.
Summary diagram for various levels of 'omic measurements. Genetic, epigenetic, and expression measurements are in yellow, purple, and green respectively.

genetic profiling (Heather and Chain, 2016; Mardis, 2008; Shendure et al., 2017; Shendure and Ji, 2008). Sequencing can be targeted specifically to exons, if coding variation is the primary interest, or to the whole genome (Hodges et al., 2007). When interpreting genetic variants, it can be important to capture the entire genome, as a vast majority of the genome is noncoding (Venter et al., 2001) and these noncoding elements influence cell, tissue, and organism function (Maurano et al., 2012; Venter et al., 2001). That said, with the current technical limitations of read size, complex genetic variations, such as repetitive elements, can be missed with the current genome sequencing methods (Treangen and Salzberg, 2011). Nonetheless, with all of these technologies, we are able to measure underlying genetic variation, at either the allele or variant level (Figure 1.2) to survey changes at a genome-wide level.

How genetic data is interpreted largely relies on the other accompanying measurements. Genetic variation in a system is often used as the causal disruption to other levels of biology, be it epigenetics, gene expression, or outward phenotype, although it can be interpreted in isolation as well. One common way to interpret genomic data is to associate it with phenotypes of interest, such as with a genome-wide association study (GWAS) or quantitative trait loci (QTL) mapping experiment (Abiola et al., 2003; Manolio, 2010; Visscher et al., 2017). Both of these compare variations in a secondary measurement to underlying genetic differences, including assigning genetic variants to a disease state. This type of association can also be made with molecular measurements, such as protein or transcript abundance, or chromatin state (Brem et al., 2005; Chick et al., 2016; Kumasaka et al., 2016; Nica and Dermitzakis, 2013). However, the mechanism by which these identified genetic variants affect any of these other measurements is often

hard to interpret, especially when looking beyond the coding sequence of the genome. For example, a vast majority of significant GWAS variants fall in noncoding parts of the genome, often limiting their interpretation solely to correlation, without the integration of measurements of intermediate biological levels to provide insight into potential mechanisms (Altshuler et al., 2008; Maurano et al., 2012). Genetic data can be interpreted in isolation as well; however, there are notable limitations to the current methods that predict variant effects. There are tools to infer the functional consequences of coding variations that alter the amino acid sequence of a protein, but other genetic variation is considerably more challenging to interpret (Altshuler et al., 2008; Brodie et al., 2016; Witte, 2010). For example, deciphering how variation in regulatory elements will influence gene expression remains a largely unresolved problem, partially due to incomplete models of gene assignment to regulatory elements (Brodie et al., 2016; Petersen et al., 2013). Many current analyses follow a 'nearest gene model' that asserts that variants will alter the expression of the gene nearest to them, but many long-range regulatory interactions have disproved blanket utility of this method (Matthews and Waxman, 2018; Nishizaki and Boyle, 2017; Petersen et al., 2013). Genetic variation is biologically meaningful and capable of being measured, but with the genomic measurements being made today, we often need other corresponding measurements to be made to thoughtfully interpret the biological implications.

1.2.2. Increasing Capabilities of Epigenetic Analyses

Epigenetics – the changes made to DNA that affect gene expression without altering genetic sequence – is a multi-faceted molecular aspect of a cell. The 3-dimentional architecture of DNA – often simplified to defining topologically associated

domains (TADs) of DNA or enhancer-promoter interactions – can be measured through chromatin confirmation capture assays (Dekker, 2006; Dekker et al., 2002; Dixon et al., 2016; van Steensel and Dekker, 2010). These identify physically proximal sections of DNA in 3-dimensions, rather than just proximity along a strand of DNA, under the assumption that interacting DNA regions will be physically near each other. At the histone level, we can also measure histone modifications – covalent post-translational modifications to histones – with Chromatin Immunoprecipitation (ChIP) sequencing (Blat and Kleckner, 1999; Park, 2009; Ren et al., 2000; Solomon et al., 1988). The combination of marks in a region are used to define its function, e.g. an enhancer, promoter, etc (Consortium, 2012; Shlyueva et al., 2014). These marks can also inform how accessible DNA is – how tight (closed) or loose (open) the DNA stand is wrapped around histones – as some marks are more highly associated with open or closed chromatin (Shlyueva et al., 2014). Generally, if a gene or regulatory element falls in a region of open chromatin, it is more likely to be active than if it were to be in a region of closed chromatin. Finally, direct modifications to the DNA – such as DNA methylation – can be measured by techniques such as bisulfite sequencing. DNA methylation, particularly in gene bodies, is involved in gene regulation by altering the accessibility of regions of DNA (Frommer et al., 1992; Lister et al., 2009). Together, measuring epigenetics can be an informative way to study gene regulation, but is intrinsically multi-leveled, complex, and limited by our imperfect understanding of the functional significance of many marks. The multiple aspects of epigenetics at a given locus are often interrelated, defining that region's chromatin state, but the tools to measure them are largely divergent.

Similar to analyzing genomic data, interpreting gene targets of regulatory elements from most epigenetic data in isolation relies on previous annotations of the epigenome. A major exception to this rule is studying the 3-dimensional architecture of the genome, as chromatin contact can imply function – a section of DNA looped over to a transcription start site is likely an enhancer-promoter loop (Shlyueva et al., 2014). For other aspects of chromatin state, predicting how variation in epigenetics in a noncoding region of DNA will alter gene expression is usually done with nearest-gene models (Matthews and Waxman, 2018). For this type of analysis, active or inactive chromatin is assumed to enhance or repress nearby gene expression, respectively. While this appears to commonly be the case, chromatin changes have frequently been shown to be more related to farther away genes as well (Matthews and Waxman, 2018; Stadhouders et al., 2012). But successfully linking epigenetic changes to downstream effects can be a powerful way to make sense of intergenic variants. While there are ways to interpret changes to the chromatin landscape alone, the power to utilize those data to learn about a biological system increases with multi-omic measurements.

1.2.3. Increasing Capabilities of Expression Analyses

Gene expression is a dynamic process that requires equivalently dynamic assessments. Broadly, these methods can be classified into two groups: those assessing regulation and those measuring abundance. Transcriptional and translational regulation can be measured by utilizing the tools that facilitate those processes. Identifying the location and abundance of transcriptional machinery, or coopting the machinery to tag newly transcribed RNA molecules, can reveal how many transcripts of specific genes are being actively created (Cleary et al., 2005; Rabani et al., 2011; Sandoval et al., 2004;

Schwanhausser et al., 2011). Similar methods can be applied to translational machinery in the cell to measure active translation to protein (Ingolia et al., 2009; Johnson et al., 2010; Sanz et al., 2009). Molecule abundance can also be measured at the RNA or protein level. RNA abundance is usually quantified by methods such as microarray or RNA sequencing (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Schena et al., 1995). Arrays are cheaper and target pre-defined genes of interest, while sequencing measures the abundance of all transcripts in a relatively unbiased manner. Recent developments in single-cell RNA sequencing (scRNA-seq) have enabled researchers to survey transcript abundance at a single-cell level, revealing a striking degree of heterogeneity, but these methods are prone to complications such as low coverage and drop outs (Hwang et al., 2018; Macosko et al., 2015). Protein abundance can also be measured in a high-throughput manner, largely with tools like mass spectrometry (Steen and Pandey, 2002). The progression of gene expression assays towards unbiased, genome-wide analyses has increased the research potential of the systems genetics field.

Gene expression assays are powerful tools to gain insights into molecular mechanisms of cell function. Transcript and protein abundance have been demonstrated to be highly correlated with cellular function in healthy cells, even if their abundance is not linked to protein abundance and therefore does not directly impace specific functions (Ball et al., 2016; Harrison et al., 2012; Sul et al., 2009; Vu et al., 2015). Curated databases, such as Gene Ontology and Ingenuity Pathway Analysis, summarize known gene functions and thus allow for the interpretation of gene expression (Harris et al., 2004; 2014). However, predicting cellular functions from gene expression levels ise challenging. As stated earlier, RNA abundance frequently fails to match protein

abundance, and even protein abundance can be buffered from cell function through numerous regulatory mechanisms, limiting the utility of applying annotations to transcript or protein abundances (Ball et al., 2016; Chick et al., 2016; Veitia and Birchler, 2015). Further, transcriptional regulation can be difficult to infer from transcript measurements alone. For instance, cases of differential gene expression can be problematic attribute to genetic variation, chromatin alterations, variable mRNA stability, or a higher-level change, such as cell identity, without additional measurements (Gaffney, 2013; Mazo et al., 2007). While powerful, and deservingly frequently used, current gene expression assays can more effectively inform researchers of cell function and gene regulation when measured in tandem with other 'omics.

1.3. Multi-Omic Data Integration to Address System Complexity

Despite the power of the aforementioned genomic methods, their utility can frequently be improved through the data integration inherent to systems-level approaches. By inducing or exploiting variation, and assessing other features of the system with mutli-omic methods, researchers can obtain a broader understanding of mechanisms and resulting implications of how variation affect the system. When, and how, to integrate these data is an outstanding question that should be handled both thoughtfully and in a case-by-case manner to create the fullest models of biology allowed by the data (Lu et al., 2005).

1.3.1. Data Integration at Different Stages

There are three stages at which data can be integrated: early, intermediate, and late (Hamid et al., 2009), each refering to the degree of processing of each independent data type before integration. Depending on the data types and the goal of one’s analysis,

any of these integration stages may be appropriate. Early data integration comprises integration before any transformation of the data has been performed, but can occur before or after data processing. For example, combining RNA-seq transcript abundance from multiple datasets would fall into the category of early data integration. This is in contrast to intermediate data integration, where some data transformation has occurred. This would include the comparison of summary vectors of data, such as principal components from multiple data types. Finally, late data integration compares data types following independent analyses on the different measurements, such as the comparison of correlation coefficients and p-values from different experiments. Features of the datasets and analytical needs of a project, determine which of these three implementations of data integration would be the most appropriate.

1.3.2. Data Integration Methods Within or Across Data Types

When studying multiple measurements of a single data type, it can be useful to integrate those data as a part of their analysis. This approach often falls into the category of early or intermediate data integration, as data could likely be combined better in raw form than as statistical summaries. Particularly congruent measurements, like two RNA-seq datasets, can often be simply merged together and treated as one complete dataset, modeling early integration. There are even tools to estimate known or unknown systematic variance and account for it across a dataset (Leek et al., 2012). However, integrating less congruent measurements within a data type is more challenging. For example, while RNA-seq data measures gene expression by molecule abundance, their differences in scale and precision require more strategic integration (Hamid et al., 2009). Epigenetic data are similarly complicated to integrate. Epigenetic measurements do not

have the same standardized set of annotated bins that gene expression measurements do – gene expression can typically be simplified to counts for genes, while epigenetic marks can differ in their localization, width, and magnitude. A common way to combine these types of data is transform all of the measurements to a binary presence or absence across the genome and overlap the transformed measurements through a combinatorial Hidden Markov model (HMM), falling into intermediate integration (Ernst and Kellis, 2012, 2017). While this can accommodate for localization and width differences between the measurements, it suffers the loss of magnitude information for all measurements. An alternative method can be performed by calling peak ranges in each sample, defining a common peakset across all samples, and then re-calling read counts for the common peakset in all samples (Ross-Innes et al., 2012; Stark and Brown, 2011). This also accounts for localization and width differences between measurements, while maintaining scale information; however, it adds bias to the system through peak-calling cut-offs that are challenging to normalize across diverse epigenetic measurements. Therefore, defining peaksets is perhaps more useful for integrating samples from a single epigenetic measurement, while binary transformations may be more informative for the integration of multiple epigenetic measurements. Genetic data can also be integrated, but our limited ability to infer valuable information genetic data can restrict this capability. The current standard for genetic integration is to put variant information onto a single scale, like a reference alignment, and annotate deviations from that sequence. There can be times when late-stage integration is most useful for integration within a data type. For example, genetic association data are often combined with meta-analyses based effect sizes or p-values, rather than integrating the raw genetic information and re-performing

statistical quantifications (Hamid et al., 2009). Within data type, integration techniques can be used to increase the sample size and improve the power of an analysis.

There are times when it is necessary to integrate multiple data types within the context of a single experiment. This is considered to be the most challenging type of data integration (Hamid et al., 2009). Methods currently exist for the integration of cellular, molecular, or organismal phenotypes with genetic information, such as quantitative trait loci (QTL) mapping or a genome-wide association study (GWAS) (Abiola et al., 2003; Manolio, 2010; Mills and Rahal, 2019). These associate genetic variants with a secondary measurement to infer mechanistic relationships. Even once statistically significant relationships have been defined, it can still be challenging to identify causative relationships. For example, the vast majority of disease-associated variants from GWAS publications fall in noncoding regions of the genome without a definitive link to gene function (Maurano et al., 2012). The same problem occurs when studying epigenetic-to-expression relationships, as the abundance and breadth of epigenetic data across the genome means variation doesn't often fall in a gene body (Capell and Berger, 2013; Shlyueva et al., 2014; Thurman et al., 2012). A flawed, but common, solution to this problem is to adopt a nearest-gene model, which simplifies gene regulation by assuming that genetic or epigenetic variation will only affect the expression of the nearest gene. Despite the weaknesses and limitations of this model, there is currently no simple alternative to interpreting distal variation (Matthews and Waxman, 2018). Integrating 'omic data with cellular data is also a challenge, as unlike multi-omic integration, their data formats and structures are drastically different (Hamid et al., 2009). One solution is the implementation of covariance analyses, as these do not rely on data type similarities,

and these have proven successful for relating cellular and molecular phenotypes (Ball et al., 2016). Integration across data types can expand the scope of a project, but doing so is challenging and will require further innovation to improve in accuracy and interpretation.

1.4. The Mouse as a Model of Human Disease

Mice have long been used to model human biology, to varied success and public opinion (Perlman, 2016; Vanhooren and Libert, 2013). Phylogenetically they are relatively similar to humans, while being easier to breed, maintain, and perturb than other, larger mammals (Berry and Bronson, 1992; Mouse Genome Sequencing et al., 2002; Perlman, 2016). There is a substantial body of work demonstrating the value of mouse models, not just for basic biology but also for modeling disease (Birling et al., 2017; Liu et al., 2017; Perlman, 2016; Vanhooren and Libert, 2013). One of the major benefits of mouse models is their capability to be used for systems genetics approaches. There are many strains that have been inbred to homozygosity, which can be perturbed in a controlled manner, and recently developed outbred populations contain genetic diversity that is equivalent to estimates of human diversity (Casellas, 2011). Utilization of mouse models in this manner can reveal novel and translatable insights for the betterment of human health.

1.4.1. Mouse Inbreeding for Systems Biology

While not originally done for research purposes, the inbreeding of mice has been one of the main contributors to their establishment as a key model organism for biomedicine (Casellas, 2011; Perlman, 2016). The origin of the inbred mice used for research today is similar to the creation of breeds of dogs – they were bred specifically for desirable traits and sold as so-called fancy mice (Nishioka, 1995; Takada et al., 2013;

Yoshiki and Moriwaki, 2006). Coopted by biomedical researchers, originally at The Jackson Laboratory, the genetic consistency between mice in a strain enabled a controlled genetic backdrop for years of productive research (Perlman, 2016; Yoshiki and Moriwaki, 2006). External perturbations, such as drug treatment, radiation exposure, or behavioral conditioning, could be performed on mice without confounding effects of underlying genetic variants (Perlman, 2016). These studies were able to resolve thes perturbation effects with far greater certainty than would otherwise be possible. While creating divergence from the concept of genetically diverse populations, such as humans, mouse inbreeding has enabled significant progress in the field of biomedicine.

1.4.2. Controlled Genetic Perturbation

In addition to environmental perturbations, inbred mice can be genetically mutated to study the functional effect of specific genetic variants (Caspary, 2010; Daxinger et al., 2013; Hrabe de Angelis and Balling, 1998; Rinchik, 1987). Particularly since the development of CRISPR/Cas9 genome editing, it has become possible to make precise mutations to the DNA to alter expression or function of genes (Cong et al., 2013; Mali et al., 2013). Large-scale mutagenesis projects have been run using mutation-generating systems, like N-Ethyl-N-Nitrosourea (ENU) chemical induction, to create random mutations that can then be phenotyped and mapped (Acevedo-Arozena et al., 2008; Salinger and Justice, 2008). Conversely, mutations have been purposefully generated in many genes in numerous, one-off studies, as well as for the long-term by the International Mouse Phenotyping Consortium (IMPC) (Meehan et al., 2017). A major limitation and criticism of these types of projects is that they are largely performed on a single strain of mouse, most frequently the C57BL/6J strain. Even across a panel of

multiple inbred strains of mice, the lack of allelic shuffling will limit the broad applicability of any findings (Montagutelli, 2000). So, while these genetic perturbations have granted insights into the molecular mechanism of many genes, their translatability to a more diverse population, like humans, can be limited in nature.

1.4.3. Introduction of Diversity to the Mouse Model

In response to concerns about the limitations of a purely inbred system, multiple mouse models of diversity have been utilized. These can be subdivided into two broad classes: panels and populations. Diversity panels, such as the Collaborative Cross (CC) and BXD (C57BL/6J and DBA/2J) panels, are composed of recombinant inbred mice derived from known founder strains (Churchill et al., 2004; Peirce et al., 2004; Taylor et al., 1999; Threadgill and Churchill, 2012). While mice from these panels are genetically homozygous genome-wide, and therefore fundamentally different from a human population, the advantage is that each mouse is reproducible (Srivastava et al., 2017). This enables varying perturbations to be made on a single strain within the panel. In contrast, populations, like the Diversity Outbred (DO) population, are heterozygous at most loci (Churchill et al., 2012; Svenson et al., 2012). This better represents individual genomes in the human population, but has limitations in mouse-to-mouse reproducibility. For all of these examples, allele shuffling generates a greater spectrum of many phenotypes, without increasing the number of alleles in the population. This demonstrates the power of diverse mouse models, which have been used numerous times to improve our functional understanding of cellular and molecular biology (Andreux et al., 2012; Bogue et al., 2015; Threadgill and Churchill, 2012; Toth et al., 2014). In a systems genetics context, natural genetic variation within a panel or population generates an

abundance of measurable perturbations to a system, making genetic diversity an intriguing resource for biomedical research in model systems (Brekke et al., 2018).

1.5. Project Rationale

In this project, I have hypothesized that studying genetics at a systems level, through multi-omic data integration, has the capacity to improve the predictive power of computational analyses. However, because many of the tools for these types of analysis are new, or even theoretical, thorough demonstration of their methodologies and capabilities are required to standardize them in the field (Endrullat et al., 2016). Across three biologically distinct projects, I performed multi-omic analyses on perturbed genetic systems, each described within the followiung independent manuscripts (Chapters 2-4). Using a combination of new and previously developed techniques, I demonstrate that multi-omic integration generates more predictive computational models of biological systems.

1.6. Research Projects

The multi-omic analyses I conducted were performed across three different biological contexts. Each one perturbed an aspect of the biological system (Figure 1.1) and used genomic technologies to measure multiple cellular and molecular phenotypes (Figure 1.2). First, we knocked-out a key meiotic histone methyltransferase and measure its effects on gene expression and cellular development. Next, we used natural genetic variation within and between two strains of inbred mice to assess how genetics alters the binding affinity of a long zinc-finger protein. Finally, we utilized a panel of nine genetically-distinct inbred strains of mice to build a model of how underlying genetic variation corresponds to gene expression differences, dependent or independent of

epigenetic changes. In total, these projects demonstrated the value of multi-omic analyses in systems genetics to build predictive models of biology.

1.6.1. Uncoupling of transcriptomic and cytological differentiation in mouse spermatocytes with impaired meiosis

Preventing deposition of proper histone modifications at recombination hotspots, by the removal of functional PRDM9, has drastic downstream effects in spermatocytes of mice (Hayashi and Matsui, 2006; Hayashi et al., 2005). We assessed these effects at levels of gene expression and cellular phenotype (Figure 1.3). We asked how well transcriptional arrest corresponds with the well-characterized cytological arrest of *Prdm9*$^{-/-}$ germ cells in male mice (Figure 1.4). We found that while gene expression and cellular development are normally tightly coupled in spermatogenesis, the knockout of *Prdm9* causes these processes to become uncoupled, presenting a challenge for inferring one from the other, and thus necessitating the measurement of both of these biological levels.

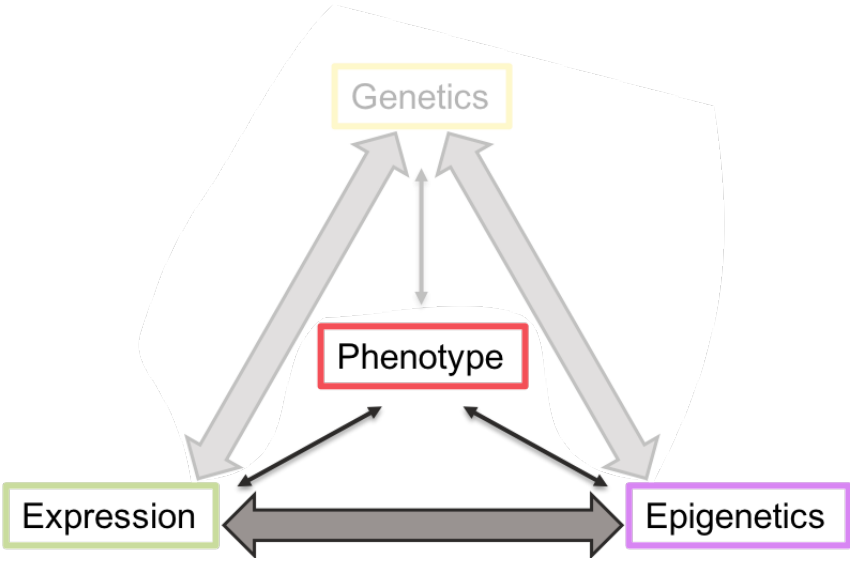

Figure 1.3. Aim 1: Pertubation of epigenetics to assess effects on gene expression and cellular phenotype.
Through the knockout of *Prdm9*, we disrupted genome-wide histone modifications and measured how that alters expression and phenotype.

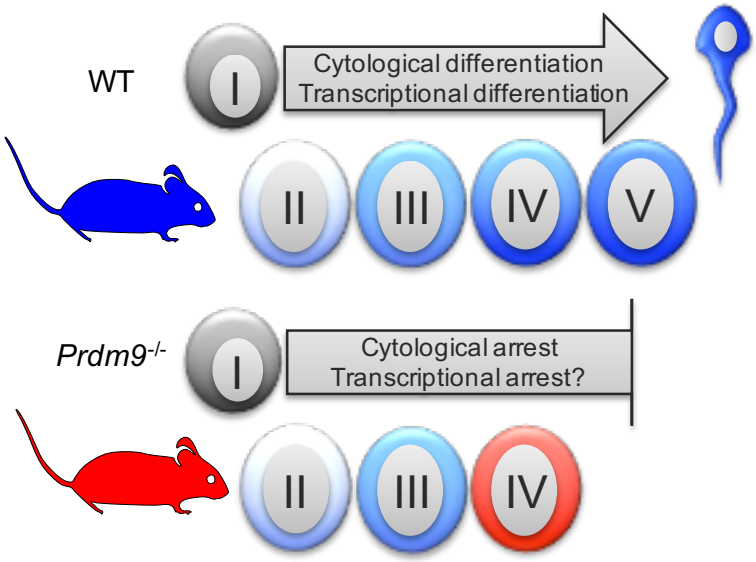

Figure 1.4. Do cytological and transcriptional programs remain coupled in Prdm9$^{-/-}$ germ cells?
In healthy mice (WT), cytological and transcriptional differentiation processes are tightly coupled through spermatogenesis, the process of spermatogonia developing into sperm cells. In systems of arrested development, like *Prdm9*$^{-/-}$ mice, it had not yet been shown how faithfully transcriptional differentiation co-arrested with cytological arrest.

1.6.2. Modeling the multiple zinc finger protein PRDM9 binding affinity using Affinity-seq

Modeling the effects of variants in the binding sites of long zinc-finger arrays is an outstanding challenge in the field of genetics. By providing the long zinc-finger, histone methyltransferase PRDM9 with multiple genomes to bind to, it is possible to assess how such a protein selects where to bind, and with what affinity (Figure 1.5) (Walker et al., 2015). We integrated two multi-omic affinity-seq datasets measuring the binding affinity of PRDM9^{Dom2} on the C57BL/6J and CAST/EiJ backgrounds to build a model that would predict differential SET activity of PRDM9 based on genetic variation in its binding site (Figure 1.6). The inclusion of binding activity with underlying variant data from within and between the mouse strains allowed us to develop a computational tool to predict the effect of those variants.

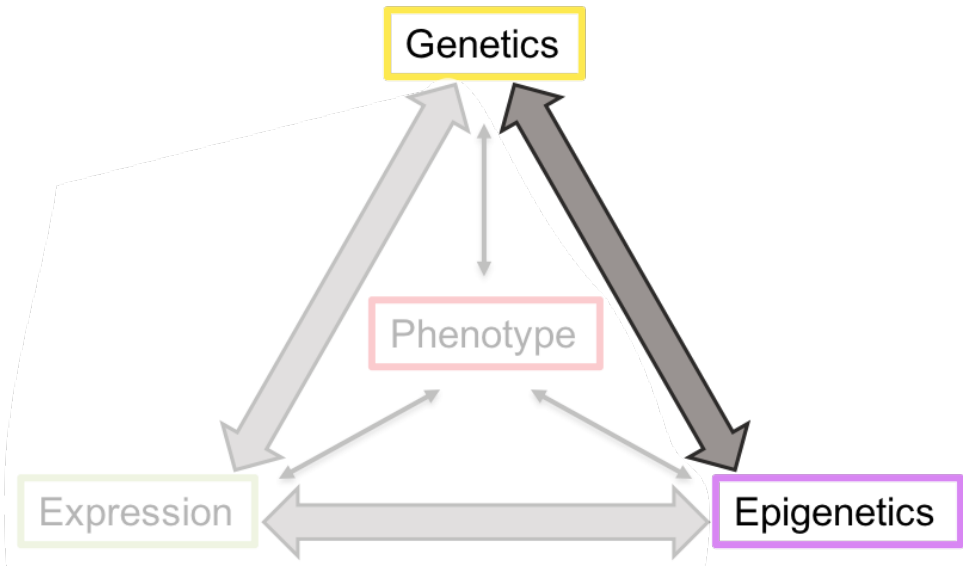

Figure 1.5. Aim 2: Natural variation in genetics utlized to predict changes in binding affinity of long zinc-finger protein and histone methyltransferase.
By integrating genome-wide genetic and protein affinity data from two strains of mice, we were able to model how genetic variants alter the binding affinity of a long zinc-finger array.

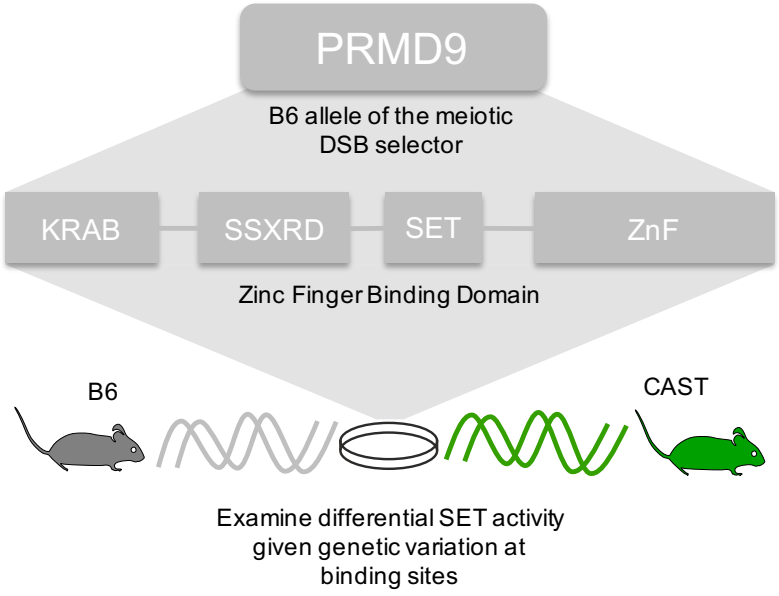

Figure 1.6. How does genetic variation in the binding site of PRDM9 alter its binding affinity?
We used the endogenous allele of PRDM9 in B6 mice, PRDM9^{Dom2}, and performed Affintiy-seq on it within the genetic backgrounds of B6 and CAST. The genetic variation between these two strains enabled us to examine the differential binding affinity of PRDM9 with variation in its binding site, aiming to build a predictive model of its SET domain activity *in vivo*.

1.6.3. Modeling the effect of genetic and epigenetic variation on gene expression in mouse hepatocytes

Genetic diversity between inbred strains of mice corresponds with both epigenetic and gene expression diversity. We built upon this knowledge by examining how well allelic differences, histone modifications, and transcript abundance are correlated at the gene level (Figure 1.7). We identified how genetic variation around a gene body alters gene expression levels, both dependent on and independent of corresponding epigenetic changes (Figure 1.8). While there is a relationship at all of these levels, no single measurement could fully explain differences in transcript abundance, signifying the value in multi-omic measurements for predicting gene expression.

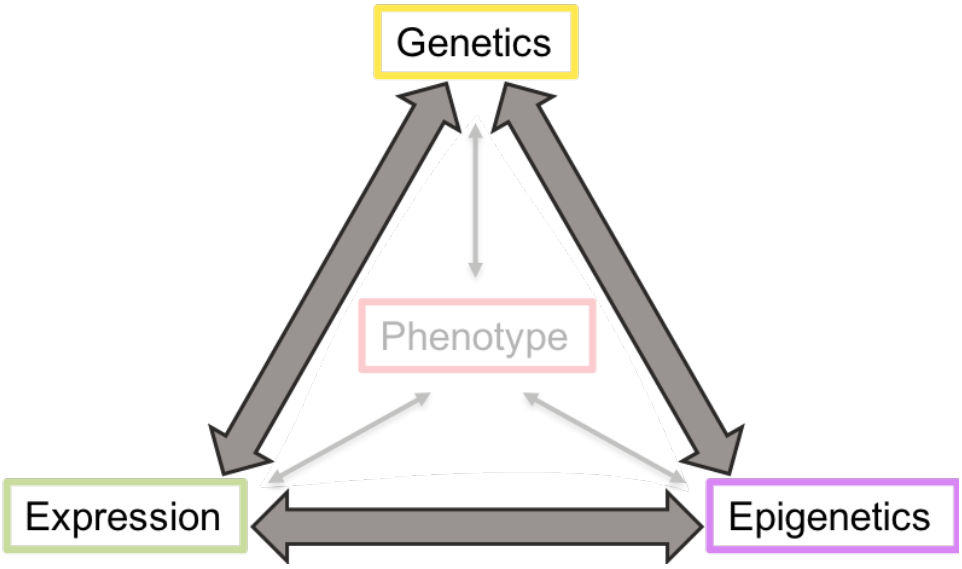

Figure 1.7. Aim 3: Inter-strain genetic variation alters gene expression both dependent on and independent of corresponding epigenetic changes.
We built a computational model of gene expression by integrating genetic, epigenetic, and transcriptomic data from mouse hepatocytes.

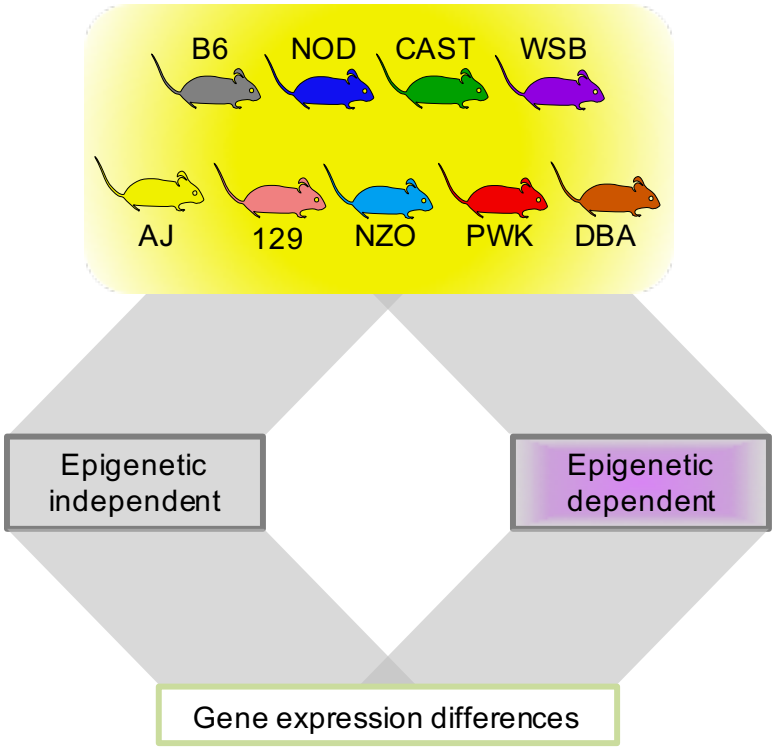

Figure 1.8. What gene expression differences can be explained by underlying genetic and epigenetic variation?
We measured transcript abundance and multiple histone modifications in nine diverse strains of mice and looking for multi-omic correlations at the gene level. We could then identify which genes appear to be most altered by underlying genetic and/or epigenetic variation.

CHAPTER 2: Uncoupling of transcriptomic and cytological differentiation in mouse spermatocytes with impaired meiosis

Fine, A. D., Ball, R. L., Fujiwara, Y., Handel, M. A., & Carter, G. W. 2019. Mol Biol Cell. 30(5), 717-728

2.1. Introduction

Cellular differentiation unfolds via a combination of genetic, epigenetic, transcriptional, translational, and cytological sub-programs that establish specific cellular identities. Transcriptional regulation is the best studied of these concurrent programs across diverse cell types (Chen and Dent, 2014), but the degree to which these sub-programs are coordinated is not well understood. Often progression through differentiation is signified by the hallmark expression of one or more well-established marker genes. Protein quantification of the expression of marker gene products is the most reliable method to stage cellular development; however, transcript abundance determined by deep sequencing allows for unbiased genome-wide analyses of gene expression. Although transcript abundance is frequently used as a surrogate for protein abundance, numerous examples have highlighted that transcript abundance does not necessarily correspond with protein abundance (Battle et al., 2015; Chick et al., 2016; Gan et al., 2013; Goncalves et al., 2017; Wu et al., 2014). Further, the reliability of transcript abundance of marker genes to reflect cytological cell state in abnormal cells is rarely assessed.

Mammalian gametogenesis is an instructive model system for cell differentiation in that it is characterized by marked transcriptomic changes paralleling morphologically dramatic stages of cytodifferentiation - particularly in the male germline. Spermatogenesis in mammals entails differentiation of committed germ cells from stem progenitor cells, followed by the germ-cell-specific process of meiosis, then cytological transformation into the highly specialized sperm cells (Eddy, 1998; Hammoud et al., 2014). The cytological and developmental stages of spermatogenesis have been well

characterized in the mouse, and, in particular, meiotic prophase substages have been defined by precise cytological and molecular criteria (Bolcun-Filas and Handel, 2018; Handel and Schimenti, 2010), which can be related to transcriptomic states (Ball et al., 2016). During meiotic prophase, spermatocyte nuclei progressively pass through well-defined and unique structural configurations as chromosomes undergo synapsis and recombination, followed by desynapsis; together these define the well-characterized prophase substages of leptonema, zygonema, pachynema, and diplonema (Baudat et al., 2013; Handel and Schimenti, 2010). Concurrent with the unfolding of these cytological events, a complex gene expression program that is highly correlated with cytologically defined meiotic prophase substages unfolds in a precise temporal pattern (Ball et al., 2016; Fallahi et al., 2010; Green et al., 2018; Ramskold et al., 2009; Schultz et al., 2003; Soumillon et al., 2013). Interestingly, an exceptionally large number and diverse array of transcripts are expressed in spermatocytes during meiotic substages, i.e., over 20,000 transcripts, including approximately 5,000 noncoding RNAs (Ball et al., 2016; Xie et al., 2014). This complex transcriptome supports spermatogenesis in at least two major ways: it provides templates for proteins that are required for meiotic progression, and it produces transcripts that are stored in an inactive state, but later support the post-meiotic differentiation known as spermiogenesis (Ball et al., 2016; da Cruz et al., 2016; Fallahi et al., 2010; Gan et al., 2013). Although expression of spermiogenic transcripts does not lead to immediate translation, their production is specific to cytologically defined meiotic prophase substages (Ball et al., 2016; da Cruz et al., 2016). Experimental analysis of the parallel programs of cytological differentiation and transcription is complicated by a practical problem: the extreme heterogeneity of both somatic and germ-cell types in the

mammalian testis makes isolation of purified cell populations and stage-specific transcript profiling difficult. Recently, this problem was addressed by a novel computational strategy, Permutation-based Maximum Covariance Analysis (PCMA), which identifies transcripts that co-vary with proportions of meiotic substages in complex cell populations (Ball et al., 2016). Although PCMA analysis associates gene expression states and cytological states, it alone cannot determine if the cytological or transcriptomic differentiation programs are sequential, or, more importantly, if one drives the other. To infer these relationships, controlled perturbation of the system is helpful.

Genetic mutations that disrupt meiotic prophase differentiation provide experimental models to analyze the coupling between concurrent differentiation and transcriptomic programs. Here we investigate the relationship between transcript abundance and cellular programs in spermatocytes from mice bearing a null mutation in the *Prdm9* gene, a key gene for meiotic prophase progression (Hayashi et al., 2005). This gene encodes PRDM9 (PR/SET Domain 9), a zinc-finger DNA-binding protein with histone methyltransferase activity that is expressed in early meiotic prophase, during leptonema and zygonema (Sun et al., 2015), and is required for activation of recombination (Parvanov et al., 2010). Recent work has demonstrated that, despite its deposition of classically gene-activating marks, PRDM9 does not directly regulate gene expression (Thibault-Sennett et al., 2018). Mice bearing inactivating mutations of *Prdm9* (herein designated as $Prdm9^{-/-}$) are infertile and meiosis is arrested at early to mid-prophase, with the failure of reciprocal recombination and absence of cytologically normal germ cells past the zygotene substage. In addition to effects on cytodifferentiation during meiosis, the molecular consequences of *Prdm9* mutation could also directly alter

the transcriptional program of meiotic prophase. The PRDM9 protein selects and binds to specific genomic sites, known as hotspots, which are subsequently recognized by the SPO11 protein for formation of the DNA double strand breaks (DSBs) that initiate the molecular events of meiotic recombination. In the absence of functional PRDM9, DSBs are misplaced to other genomic sites, primarily, but not exclusively, gene promoters (Brick et al., 2012). The ectopic DSBs fail to be repaired, which is likely a part of the cause for meiotic arrest in *Prdm9* mutant spermatocytes. It has also been shown that *Prdm9*$^{-/-}$ mutant spermatocytes fail to inactivate the sex chromosomes (Hayashi and Matsui, 2006). With its robust characterization, the *Prdm9* mutant is an informative model in which to study the coupling of cellular differentiation and transcriptomic programs in the context of an arrested developmental program. Here, our goal was to determine how the transcriptomic phenotype in *Prdm9*$^{-/-}$ germ cells relates to well-characterized meiotic substages and meiotic arrest of *Prdm9*$^{-/-}$ spermatocytes. We studied the initial wave of spermatogenesis in newborn mutant and wild-type mice in order to compare similar populations of cells in both. We documented both cellular and transcriptomic states in the same cell populations. This strategy revealed that the cytological and transcriptomic programs become uncoupled during abnormal meiotic progression, with progression of the transcriptomic program in spite of disruption and arrest of the cytological program of differentiation. This observation suggests a complex degree of independent regulation of co-occurring programs of differentiation, a conclusion that underscores the importance of anchoring transcript abundance profiles to their cellular context in order to understand both processes.

2.2. Methods

2.2.1. Experimental design

2.2.1.1. Sample acquisition

All genotypes of mice collected for this study were bred from *Prdm9*^{tm1Ymat} heterozygous mice, which are nearly congenic C57BL/6J (B6) mice, with < 10% 129P2/OlaHsd remaining, and obtained from The Jackson Laboratory (Bar Harbor, ME). Samples were collected at 8, 12, and 16 days post partum (dpp); Ten samples were collected at each time point, comprising various proportions of *Prdm9*$^{+/+}$, *Prdm9*$^{+/-}$, and *Prdm9*$^{-/-}$ mice. At each time point, five mice were *Prdm9*$^{-/-}$. The testes of each mouse were pooled prior to germ cell enrichment. An aliquot of enriched germ cells was analyzed for the substage proportion in each sample through cytological methods, while the rest of the germ cells were used for our transcriptome analysis using RNA sequencing. All mice were maintained following protocols approved by the Jackson Laboratory (JAX) Institutional Animal Care and Use Committee (IACUC).

2.2.1.2. Germ cell enrichment

Interstitial cells were removed with collagenase from seminiferous tubules, which were subsequently enzymatically digested to yield dispersed cells. Germ cells were isolated from this population by size filtration. Details of the procedure are outlined in (Ball et al., 2016; La Salle et al., 2009). The resulting cell populations were relatively pure, with no more than 10% of each sample comprised of somatic cells (recognized as not expressing SYCP3, phosphorylated histone H2AX, STRA8, or H1T and showing a distinct DAPI staining pattern) (Figure 2.1B).

2.2.2. Cytological methods

2.2.2.1. Chromatin spread preparation and immunostaining of spread chromatin

Briefly, as previously described (Ball et al., 2016), spread chromatin of germ cells were fixed in 1% paraformaldehyde (PFA) in H_2O containing 0.015% SDS and 0.02% Photo-Flo (Kodak, Rochester, NY, USA) for 1 hour at room temperature and cells were further fixed in 2% PFA in H_2O with or without 0.03% SDS for 3 minutes each. For immunostaining, prepared chromatin preparation was incubated with 10 % Antibody Dilution Buffer (ADB) blocking solution (PBS containing 2% BSA and 0.05% Triton-X 100), and immunostained with rat anti-SYCP3 antibody (1:1000 dilution, Handel lab), mouse anti-phosphorylated histone H2AX antibody (1:200 dilution, 05-636, Millipore, Billerica, MA, USA), rabbit anti-STRA8 antibody (1:1000 dilution, ab49405, Abcam, Cambridge, England), guinea pig anti-H1t antibody (1:500 dilution, Handel lab); and with Alexa Fluor® (488, 594, or 647) conjugated secondary antibodies (1:500 dilution, Molecular probes, Thermo Fisher Scientific, Waltham, MA, USA); nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI). Images were observed using a Zeiss AxioImager.Z2 epifluorescence microscope equipped with a Zeiss AxioCam MRm CCD camera (Carl Zeiss, Jena, Germany). Approximately 450 germ cells were staged per sample. Due to their similar frequency patterns, as well as to match the previously published dataset (Ball et al., 2016), late-leptotene and zygotene were combined, as well as late-pachytene and diplotene, for all analyses.

2.2.3. RNA methods

2.2.3.1. Isolation of RNA and sequencing library preparation and RNA sequencing

As outlined previously (Ball et al., 2016), cells were resuspended and homogenized before RNA was purified from each sample. The quality of the isolated RNA was assessed and then the mRNA sequencing libraries were prepared and subsequently tested for quality; 100-base paired-end reads were sequenced and filtered by quality. Technical replicates were run in different lanes and merged for the final samples used for later analyses.

2.2.4. Computational methods

2.2.4.1. Data and Code availability

All code used to produce major findings for this manuscript can be found at https://github.com/AFine1/Uncoupling-in-Prdm9KO. All transcriptomic data are available through Gene Expression Omnibus, accession GSE110703. An R Shiny web app is available at https://shinyapps.jax.org/d86f1ec60d6c596dfc1eab16e5d68aca for the visualization of gene expression levels from our dataset.

2.2.4.2. Alignment and expression

The transcripts from each RNA-seq sample were aligned and quantified on a custom built pseudotranscriptome, comprising the mm10 reference transcriptome of Ensembl Genome Reference Consortium, build 38, release 75 (Flicek et al., 2014), NONCODE v4 lncRNA (Xie et al., 2014), and piRNA precursor transcripts (Li et al., 2013). The sequences and genomic positions of the piRNA precursors were acquired from Ball *et al*. (2016). NONCODE and piRNA transcripts were defined on mm9 and

converted to mm10 coordinates using liftOver (Fujita et al., 2011). Alignment of our RNA-seq samples was performed using Bowtie 1.0.0 (Langmead et al., 2009) and the estimation of expression was calculated using RSEM (Li and Dewey, 2011). Transcript expression was quantified as log, base 2, of transcripts per million (TPM) from RSEM, $\log_2(\text{TPM}+1)$. Transcripts were excluded from further analysis if the expression was less than 1 for all samples, i.e., we required ComBat-adjusted (Johnson et al., 2007b) $\log_2(\text{TPM}+1) \geq 1$ in at least one sample. From the expression of *Prdm9*, it appeared that the genotype of two samples at 12 dpp had been mislabeled, so neither sample were included in the analyses.

2.2.4.3. RNA-seq sample integration

To increase the number of samples at each time point, we also utilized published data from previously collected *Prdm9*$^{+/+}$ samples (Supplemental Table S1) that had been collected with the same protocols as above (Ball et al., 2016) (GSE72833). These mice were all C57BL/6J mice obtained from The Jackson Laboratory (Bar Harbor, ME). Substage proportion and transcriptome profiles had been collected. Three samples were utilized from 8 dpp and five samples were utilized from 12 and 16 dpp respectively. To distinguish the origin of each sample, samples that were collected for the purpose of this study are referred to as being in the Prdm9 Dataset, while samples collected previously (Ball et al., 2016) are a part of the Background Dataset.

2.2.4.4. Principal component analysis

To visualize the transcriptomic variation between samples, a Principal Component Analysis (PCA) was run by performing Singular Value Decomposition (SVD) on transcript expression data, where each transcript's expression was centered and scaled.

The PCA was performed using the function svd(x) using the R statistical framework (R Core Team, 2015).

2.2.4.5. ComBat adjustment

We used ComBat (Johnson et al., 2007b) to adjust for known systematic variation in our dataset. The systematic variation between our samples, as visualized by PCA, showed that samples segregate by dataset (Baseline vs Prdm9, Figure 2.2A), as well as by litter (Figure 2.2B-D). Since this variation can be directly attributed to these known cofactors, we used ComBat in R from the package sva (Leek et al., 2017) to adjust for these batch and litter effects. First, we ran the expression data through ComBat to adjust for dataset variation (Prdm9 vs Baseline), keeping variation that could be attributed to genotype and age. Then, for each time point, we corrected for variation by litter. Genotype, dataset, and litter information can be found in Supplemental Table S1. The litter information for samples from the Baseline dataset was unknown, so these were considered to be a part of a single litter at each time point.

2.2.4.6. Differential expression analysis

We used a regression model to identify genes that were differentially expressed in *Prdm9*$^{-/-}$ germ cells, compared to *Prdm9*$^{+/+}$ germ cells. We ran the function lm(x) from the stats package in R (R Core Team, 2015). This linear model fits variation in gene expression across samples to given variables of interest. For this analysis, the variables that we were interested in were genotype, age, and genotype-by-age. We then used the function contrast(x) (Kuhn et al., 2016) to identify the significance of changes in gene expression between conditions at each time point. P-values were corrected for multiple testing using p.adjust(x) (R Core Team, 2015). Transcripts with an FDR < 0.01 and a Log

Fold Change > 0.5 were considered to be to be significantly differentially expressed. Differential expression coefficients, un-adjusted p-values, as well as FDR values are in Supplemental Table S2.

2.2.4.7. Epigenetic integration

Published ChIP-seq data (Brick et al., 2012) were downloaded and integrated into our analysis. We downloaded DMC1 peaks from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO, GSE35498). Using DMC1 as a marker for DSBs in germ cells, we identified which DSBs in *Prdm9*$^{-/-}$ germ cells were found at promoters. DSBs were defined as the central 1 Kb region of DMC1 peaks and promoters were defined as 1 Kb upstream or downstream of transcription start sites (TSS). We used Fisher's exact test to test the enrichment of DSBs at the promoters of differentially expressed genes (R Core Team, 2015)

2.2.4.8. Permutation-based Maximum Covariance Analysis (PMCA)

We assigned transcripts to substages of prophase in the *Prdm9*$^{-/-}$ samples with Permutation-based Maximum Covariance Analysis (PMCA) (Ball et al., 2016). This robust statistical method identifies common patterns across measurements of coupled samples and provides a false discovery rate (FDR) for each association. In this study, PMCA identified transcripts whose abundances followed the same pattern as the proportions of a substage's cells across time points, thus identifying substage-specific gene sets. We applied PMCA to the Prdm9 dataset and identified substage-specific gene sets for the Prdm9 mutant, with FDR ≤ 0.01 for all substages. Substage assignment for transcripts is indicated in Supplemental Table S2. Next, we compared these gene sets to

previously identified wild-type substage-specific gene sets, similarly identified with PMCA in a previously published study (Ball et al., 2016).

2.2.5. Bioinformatic methods

2.2.5.1. Gene Ontology (GO) analysis

GO term enrichment analysis was implemented on sets of differentially expressed genes at each time point using Gene Ontology enRIchment anaLysis and visuaLizAtion (GOrilla) (Eden et al., 2009; Harris et al., 2004). Because the GO analysis we used does not take into account direction or magnitude of differential expression, differentially expressed genes (DEGs) at each time point were analyzed separately based on the directionality of their differential expression and provided to GOrilla as an unranked list of genes. We used the list of all genes expressed in our dataset as the background set, to avoid over-representation of germ cell and meiosis related terms.

2.2.5.2. Pathway analysis

To further assess the functionality of DEGs, we used Ingenuity Pathway Analysis (IPA, QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis) (2014). This program uses curated gene annotations, associations, and functions to provide pathways that are enriched in a given gene list. We used IPA to analyze our DEG lists from 12 and 16 dpp respectively. We initially used FDR < 0.05 for our significance threshold for DEGs for these analyses; however, due to the number of DEGs at 16 dpp, a large number of pathways were enriched that made the interpretation of these results challenging. Therefore, we reran this analysis after further limiting our

DEGs at 16dpp to FDR < 0.01 in order to identify only the most relevant pathways enriched.

2.2.5.3. Transcription factor analysis

Transcription factors (TFs) were identified for gene lists using iRegulon, version 1.3 (Janky et al., 2014) in Cytoscape, version 3.4.0 (Shannon et al., 2003). For all analyses, identified TFs were limited to those that were expressed in our dataset and had a normalized enrichment score (NES) greater than or equal to 4, corresponding to an FDR less than 0.01. TFs identified in this study were compared to previously published TFs (Ball et al., 2016).

2.3. Results

2.3.1. Cytological staging sets parameters of the *Prdm9* mutant phenotype and provides context for concurrent transcriptomic analyses

We characterized the cellular and molecular effects of loss of PRDM9 function in male germ cells undergoing meiosis. Germ cells were enriched from testes collected from *Prdm9* wild-type (WT), heterozygous (*Prdm9*$^{+/-}$), and mutant (*Prdm9*$^{-/-}$) male mice at 8, 12, and 16 days post partum (dpp) (Methods) (Figure 2.1A; Supplemental Table S1), a time period when the first wave of differentiating germ cells progresses through meiotic prophase I substages. A small portion of each sample of enriched germ cells was used for cytological staging, with the remainder used for RNA-seq (Figure 2.1A).

Cytological substage-specific protein markers were used to characterize cellular morphology and determine the relative proportions of meiotic substages in each sample of *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ germ cells (Figure 2-1B and C). Spermatogonia (SP), and

multiple sub-stages of spermatocytes - pre-leptotene (PL), early-leptotene (EL), late-leptotene/zygotene (LL/Z), early-pachytene (EP), and late-pachytene/diplotene (LP/D) - were scored by combinatorial application of antibodies recognizing previously established marker proteins (Methods) (Ball et al., 2016). As the juvenile mice age toward puberty, the time period when the first wave of germ cells develop, subsequent waves of spermatogenesis are also continuously initiated, resulting in presence of each of substages preceding the most developed substage at each time point (Figure 2.1C). While all stages listed above were represented in the WT B6 samples, no cells in the *Prdm9*$^{-/-}$ samples met the cytological criteria for mid-to-late-pachytene spermatocytes (e.g., full synapsis detected by labeling with SYCP3 antibody, γH2AX restricted to the XY chromosome pair, and presence of histone H1t). Those cells exhibiting only some of these features, e.g., only partial synapsis, with a diffuse labeling pattern of γH2AX labeling that is indicative of accumulated DSBs, were defined as "pachytene-like" (P-like) cells (Figure 2.1B). These observations are consistent with previous reports of the *Prdm9*$^{-/-}$ phenotype (Hayashi and Matsui, 2006; Hayashi et al., 2005). Each time point scored captured a key aspect of the mutant phenotype in *Prdm9*$^{-/-}$ mice.

The substage proportion differences between the WT and mutant germ cell populations were not severe at 8 dpp but became much more apparent by 12 dpp, with the greatest differences observed at 16 dpp. Thus, at 8 dpp, the germ cells present in both WT and mutant samples are in early meiotic prophase, and do not exhibit any apparent morphological phenotype in *Prdm9*$^{-/-}$ testes (Figure 2.1C, left). However, by 12 dpp, there is apparent delay in differentiation of *Prdm9*$^{-/-}$ germ cells, reflected in relatively lower proportions of later prophase spermatocytes compared to WT at the same age

(Figure 2.1C, middle). Finally, at 16 dpp, when WT testes have many mid- to late-prophase spermatocytes (EP and LP/D), the germ cells in *Prdm9*$^{-/-}$ testes appear arrested, with no progress beyond a pachytene-like stage (described above), reflected by both appearance of the spermatocytes and diminished proportion of stages represented in the samples collected at this time (Figure 2.1C, right). These cytological characterizations and determination of proportions of substages, which reflect prior knowledge of the phenotype, inform the interpretation of germ-cell transcriptome data (below).

2.3.2. Specific gene signatures reflect known mutant phenotypes

RNA-seq was used to obtain genome-scale transcriptome states that correspond with the cellular states obtained at each time point (above). We sought to use an unbiased method to account for multi-dimensional variation our RNA-seq data in order to reveal transcriptomic differences between *Prdm9*$^{+/+}$, *Prdm9*$^{+/-}$, and *Prdm9*$^{-/-}$ samples. We first identified and removed variability in the transcriptome data stemming from batch and litter differences (Methods, Figure 2.2 & Figure 2.3) (Johnson et al., 2007b), and then performed Principal Component Analysis (PCA) on the residual variation (Figure 2.4A). The first principal component (PC) separated samples by age, and the second PC separated samples by genotype (at 16 dpp), both parameters being expected determinants of transcriptome states. Further, PC3 segregated the 12 dpp samples by their *Prdm9* genotype (Figure 2.3A). Despite *Prdm9*$^{+/-}$ mice being fertile, they generally displayed intermediate cytological (Figure 2.3B) and transcriptomic phenotypes (Figure 2.4A, Figure 2.3A, C-E). Differential expression analysis (Methods) between WT and *Prdm9*$^{-/-}$ transcriptomes at each time point revealed a number of differentially expressed genes (DEGs), with some downregulated (Figure 2.4B), and some up regulated (Figure 2.4C).

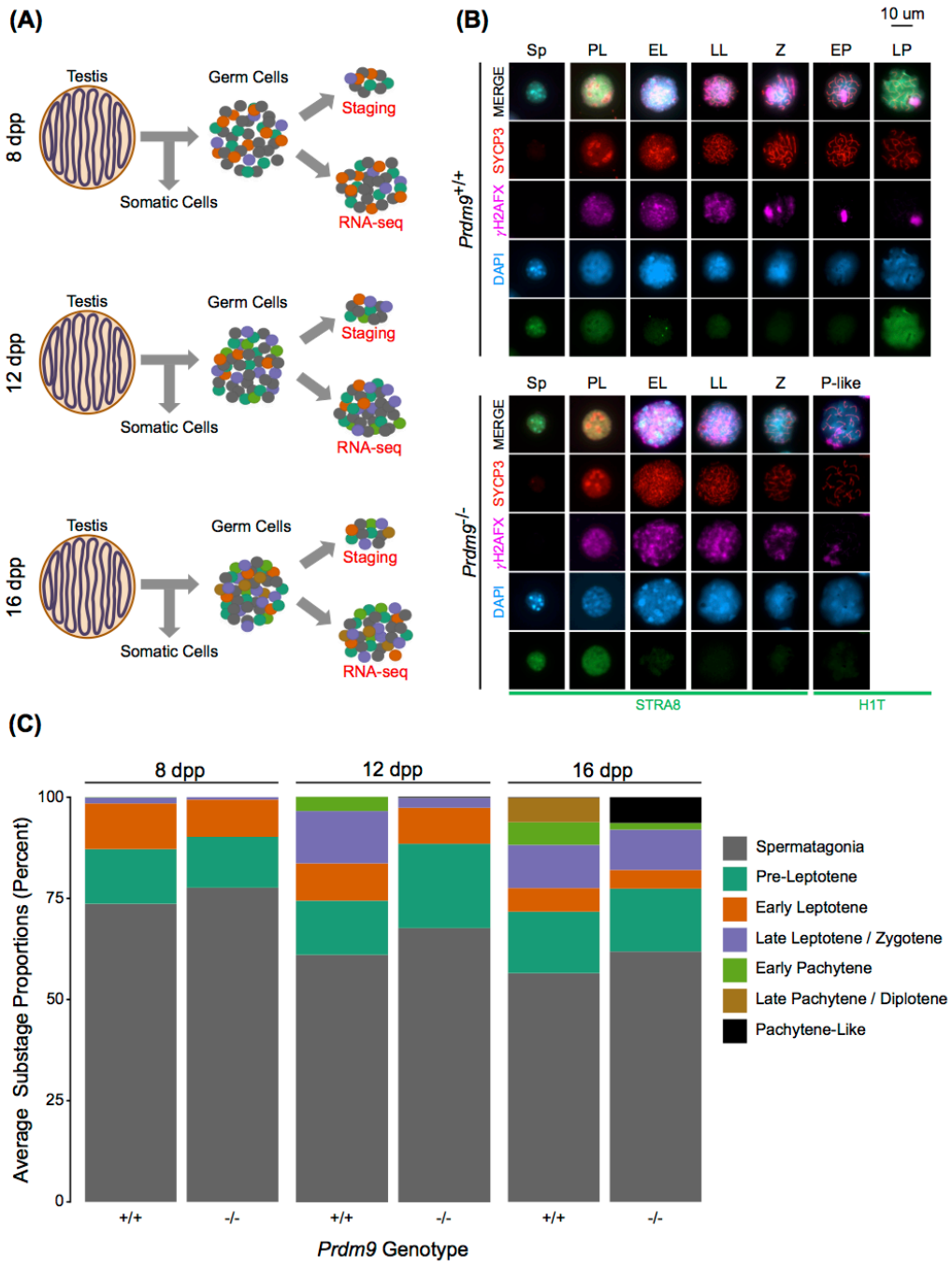

Figure 2.1. Cytological phenotypes of *Prdm9*$^{-/-}$ spermatocytes reflect meiotic arrest. (A) Cytological analyses and RNA sequencing were performed on germ cells enriched from testes at 8, 12, and 16 dpp. (B) Representative images of spermatocytes in each meiotic substage across *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ samples. Germ cells were immunostained for combinatorial arrays of marker proteins that are well established for cytological characterization of meiotic prophase substages in spermatocytes. (C) Quantification of average frequencies of spermatogenic and meiotic prophase substages represented at each time point in the samples of germ cells retrieved from *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ testes.

The number of DEGs detected increased with age in both WT and mutant samples (Figure 2.4B and C), reflecting the increase in complexity of cellular composition with increasing age.

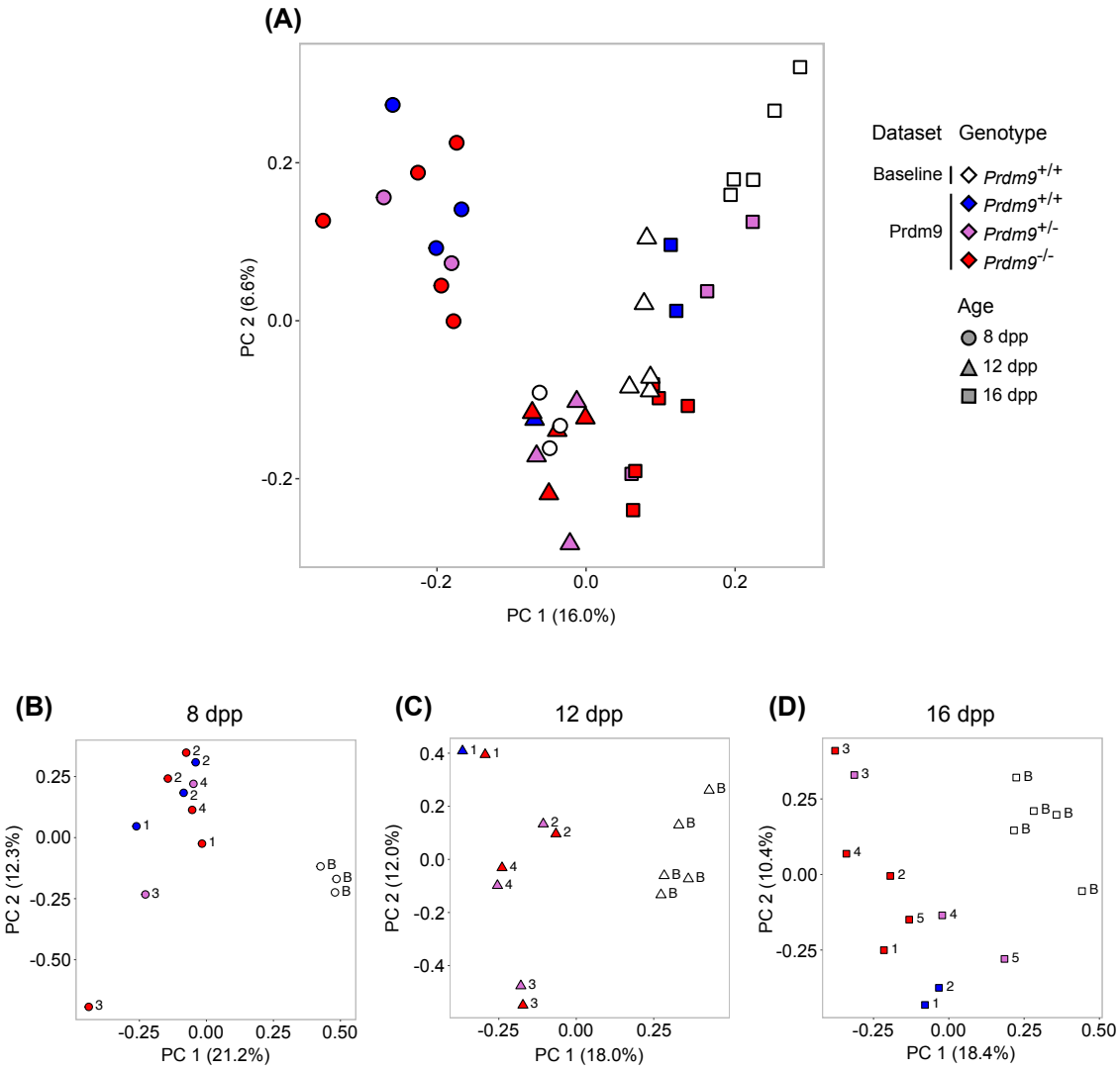

Figure 2.2. Systematic variance is apparent across batches and litters.
Colors correspond to genotype and shapes correspond to sample age, as indicated. Dataset information refers to the incorporation of a previously published dataset (baseline) into our dataset (Prdm9). (A) Principal component 1 (PC1) versus principal component 2 (PC2) of PCA run on all samples. (B,C,D) Principal component 1 (PC1) versus principal component 2 (PC2) of PCA run on 8 dpp, 12 dpp, and 16 dpp samples respectively. Icon labels correspond to designations of litters.

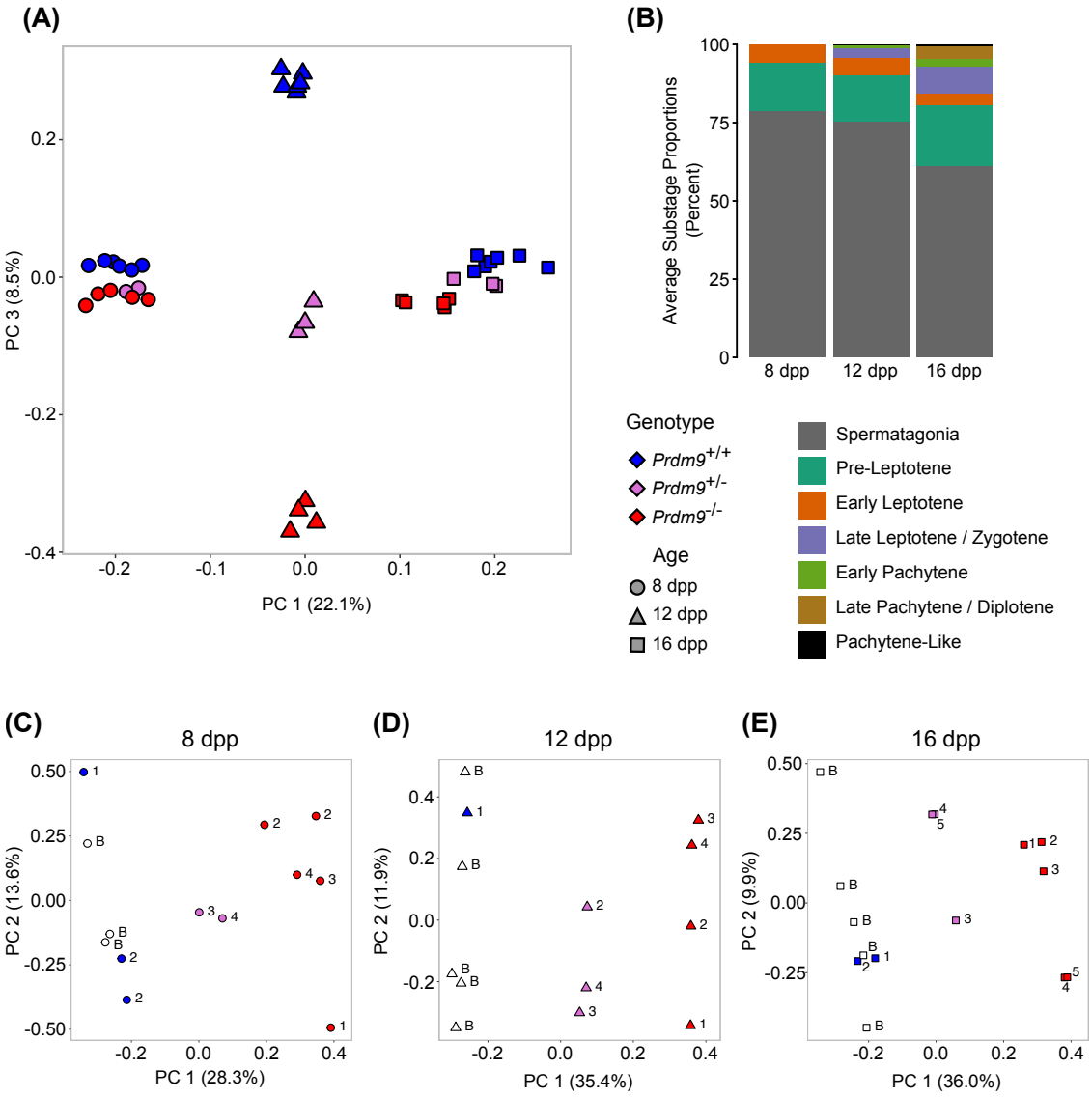

Figure 2.3. Differential expression between *Prdm9*$^{-/-}$ and *Prdm9*$^{+/+}$ samples is detected irrespective of batch effects.
Colors correspond to genotype and shapes correspond to sample age, as indicated. (A) Principal component 1 (PC1) versus principal component 3 (PC3) of PCA run on all ComBat-adjusted samples. (B) Quantification of average frequencies of spermatogenic and meiotic prophase substages represented at each time point in the samples of germ cells retrieved from *Prdm9*$^{+/-}$ testes. (C,D,E) Principal component 1 (PC1) versus principal component 2 (PC2) of PCA run on ComBat-adjusted data from samples at 8 dpp, 12 dpp, and 16 dpp respectively. Icon labels correspond to designations of litters.

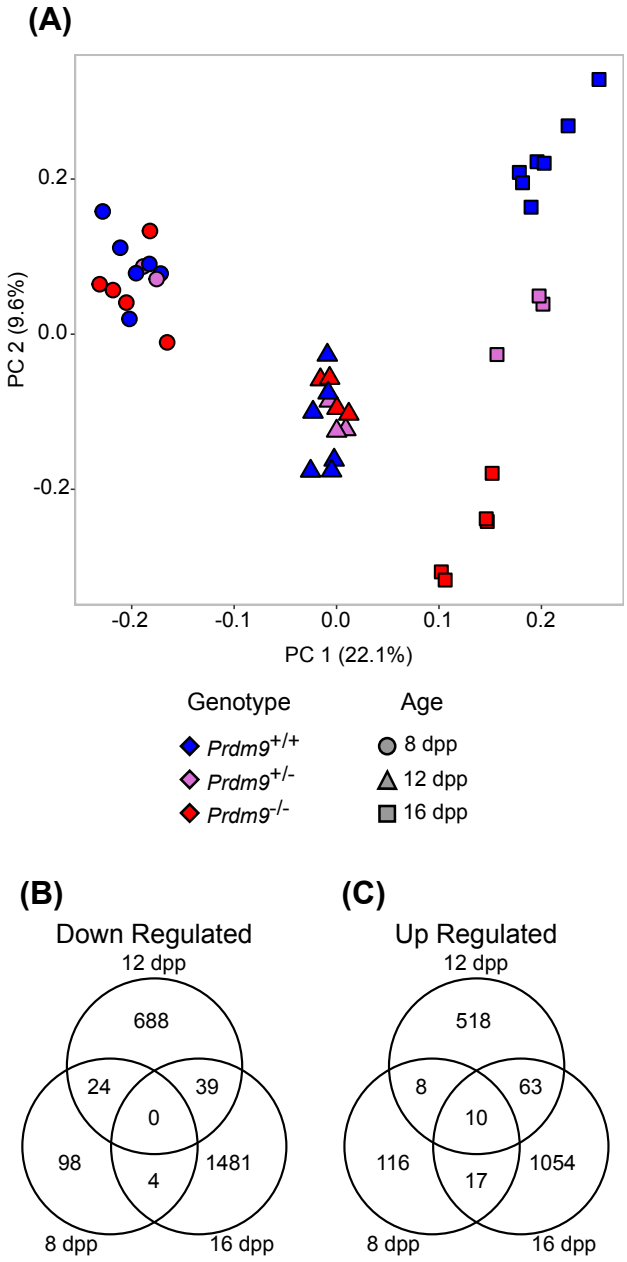

Figure 2.4. ComBat-adjusted data of gene expression across genotype and age conditions shows differential expression between *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ samples. (A) Principal component 1 (PC1) versus principal component 2 (PC2) from PCA of all ComBat-adjusted samples. Colors denote genotype and shapes denote sample age, as indicated. (B,C) Shared and unique differentially expressed transcripts with decreased or increased abundance in *Prdm9*$^{-/-}$ samples compared to *Prdm9*$^{+/+}$. FDR < 0.01 and LFC > 0.5 for B and C.

As expected, expression of the *Prdm9* gene in *Prdm9*$^{-/-}$ samples was significantly less than in WT samples (Figure 2.5A). Although low levels of *Prdm9* transcripts were detected in the *Prdm9*$^{-/-}$ samples, no reads mapped to the exons deleted in the mutant (Figure 2.5B), and PRDM9 has been shown to be absent from these mice (Sun et al., 2015). Other genes previously reported (Hayashi et al., 2005) to be differentially expressed in *Prdm9*$^{-/-}$ germ cells or key genes involved in processes expected to be disrupted in *Prdm9*$^{-/-}$ testes were validated in our data (*Morc2b*, *Hspa2*, and *Piwil1*, Figure 2.6A), as well as early meiotic regulatory genes not expected to change (*Dmc1*, *Spo11*, and *Stra8*, Figure 2.6B). As might be anticipated, given the decrease in *Piwil1* expression between *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ samples, Piwi-interacting RNA (piRNA) precursors were not expressed at wild-type levels in *Prdm9*$^{-/-}$ germ cells at 16 dpp (Figure 2.6C). In concordance with molecular patterns found in previous analyses of *Prdm9* mutants, we found that XY-linked genes were enriched in over-represented DEGs ($p < 2.2x10^{-16}$) from *Prdm9*$^{-/-}$ germ cells at 16 dpp, corresponding with the failure of cells to progress to a stage with a fully formed and transcriptionally inactivated sex-body (Figure 2.7A) (Hayashi and Matsui, 2006; Hayashi et al., 2005; Namekawa et al., 2006). Autosomal transcripts were not similarly over-represented (Figure 2.7B). Taken together, these transcriptomic data reflect both previous reports on the *Prdm9* mutants (Hayashi et al., 2005), and known temporal patterns of gene expression in spermatocytes (Deng and Lin, 2002).

2.3.3. Transcriptomic changes precede cytological phenotypes in *Prdm9*$^{-/-}$ testes

We annotated the DEGs at each time point to compare how well mutant transcriptomic phenotypes coincide with cytological phenotypes. To determine if there

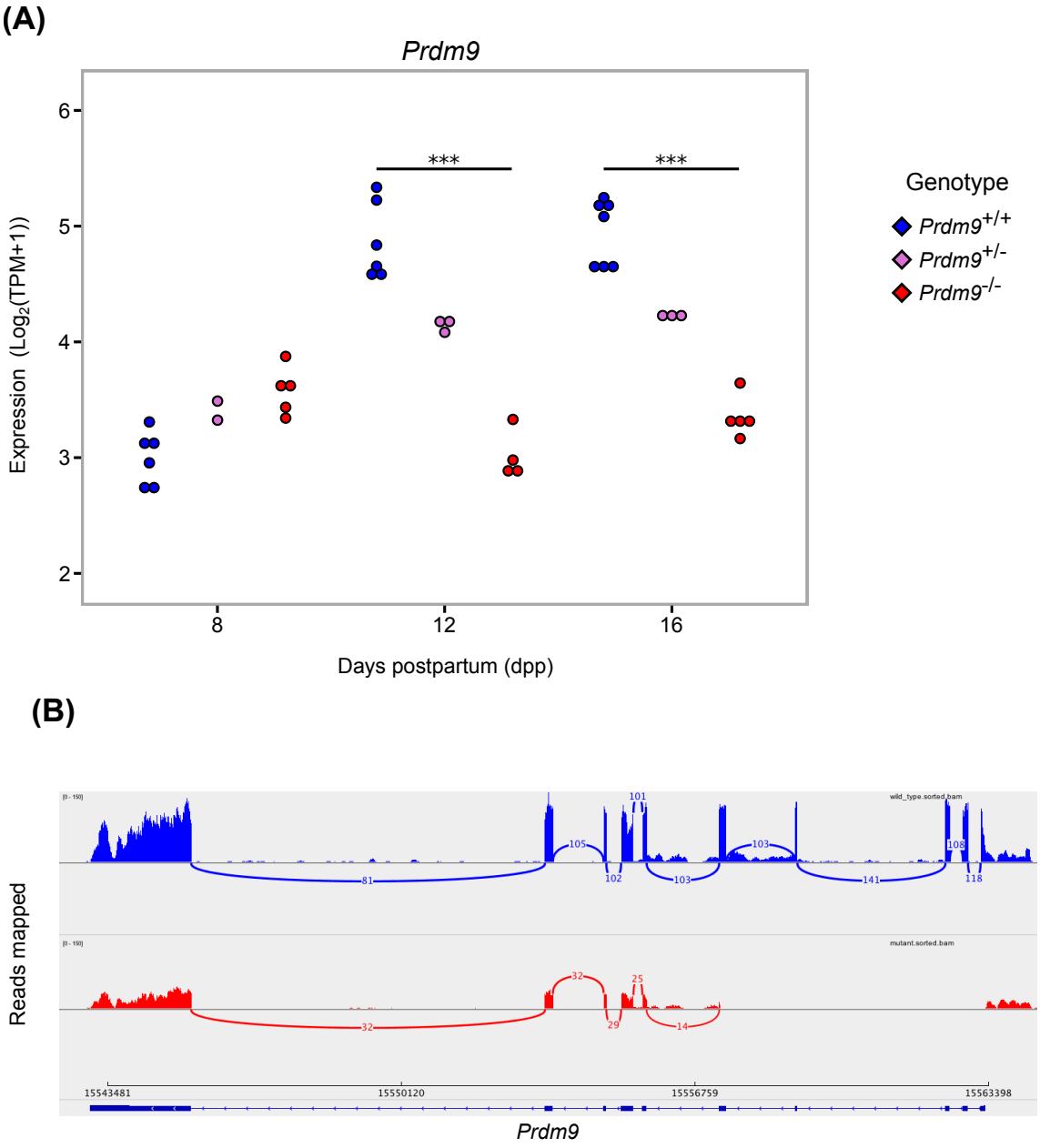

Figure 2.5. *Prdm9* transcript abundance reflects the genetic mutation.
(A) Log_2(TPM+1) expression of *Prdm9* at 8, 12, and 16 dpp. *** represents FDR < 0.0001. (B) Sashimi plot of transcript alignment to exons of *Prdm9* in $Prdm9^{+/+}$ (blue) and $Prdm9^{-/-}$ (red). The x-axis represents the length of the *Prdm9* gene and the y-axis represents the number of reads aligned to a given region. Horizontal, curved lines represent the mRNA reads bridging an intron, with the number of reads spanning the intron denoted in the middle of the line.

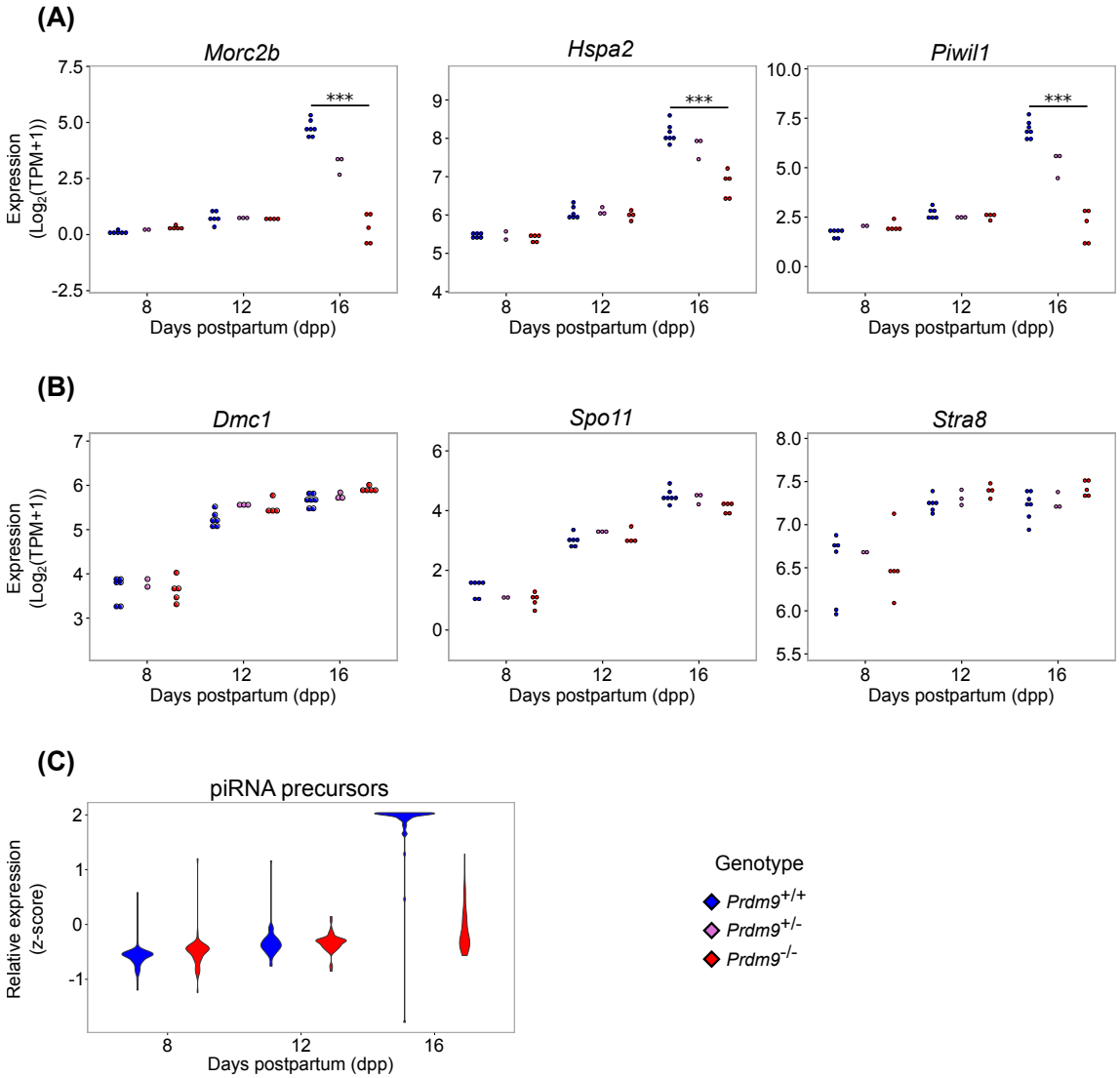

Figure 2.6. Changes in expression of specific meiotic gene reflect abnormalities and meiotic arrest in *Prdm9*$^{-/-}$ germ cells.
Colors denote genotypes, as indicated. (A) Log_2(TPM+1) expression of *Morc2b*, *Hspa2*, and *Piwil1* at 8, 12, and 16 dpp. (B) Log_2(TPM+1) expression of *Dmc1*, *Spo11*, and *Stra8* at 8, 12, and 16 dpp. (C) Relative expression of piRNA precursors at 8, 12, and 16 dpp. *** represents FDR < 0.0001.

are early transcriptomic signatures of the *Prdm9*$^{-/-}$ cytological phenotype, we conducted Gene Ontology (GO) enrichment analyses and Ingenuity Pathway Analysis (IPA) on the DEGs at each time point (Methods). At 16 dpp, when the mutant testes exhibit the most drastic phenotype, genes annotated to spermatogenesis-related GO terms, specifically genes related to late spermatogenesis, were enriched in the down-regulated DEG lists,

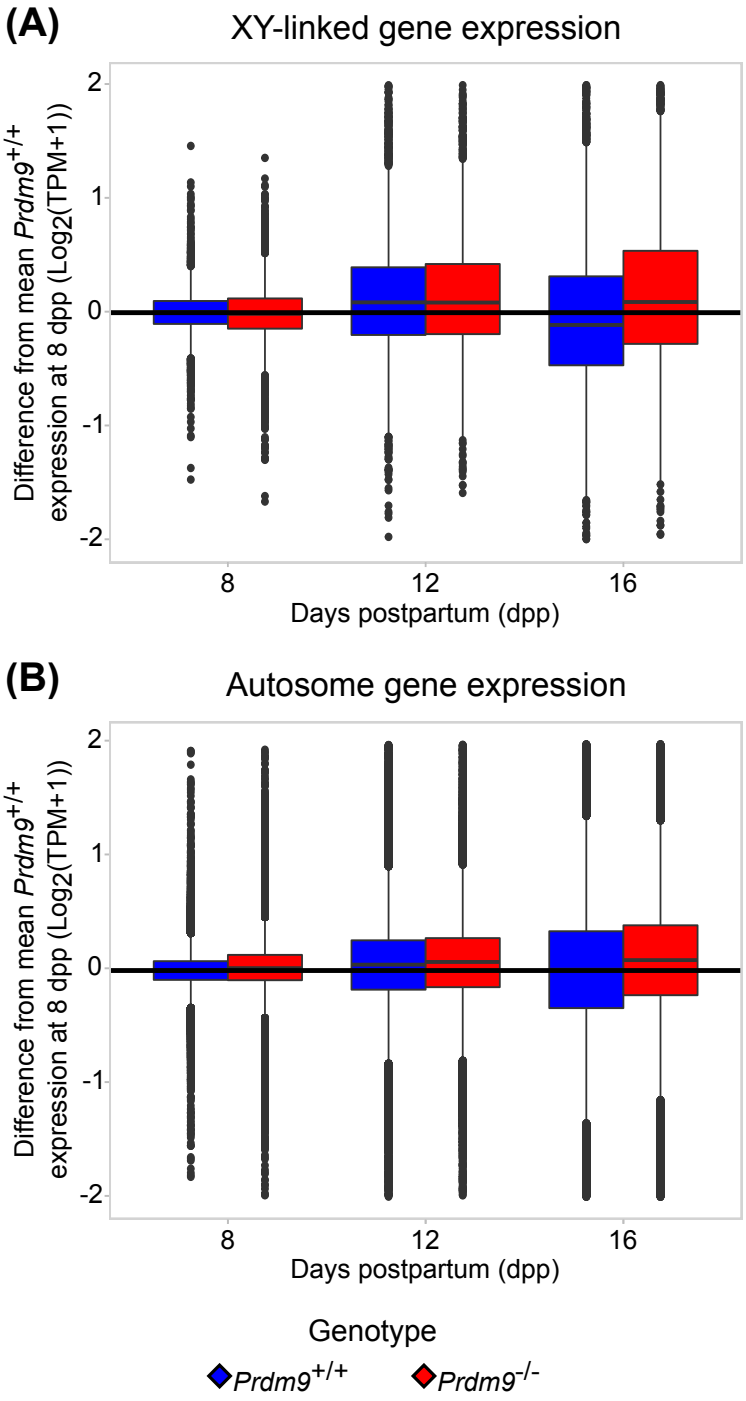

Figure 2.7. Sex-chromosome gene expression reflects impaired meiotic sex-chromosome inactivation (MSCI) in *Prdm9*$^{-/-}$ samples.
(A) Expression of each XY gene at 8, 12, and 16 dpp, relative to mean expression at 8 dpp in *Prdm9*$^{+/+}$ samples. (B) Expression of each autosomal gene at 8, 12, and 16 dpp, relative to mean expression at 8 dpp in *Prdm9*$^{+/+}$ samples. Colors denote genotypes in A and B, as indicated.

and terms related to defense and immune responses were highly enriched in the up-regulated DEG lists (all FDR <0.05; Supplemental Table S3). (Although this analysis could be biased by increased representation of sex-chromosome transcripts, limiting GO enrichment analysis to autosomal DEGs did not substantially change the repertoire of enriched terms.) IPA identified 'Cyclins and Cell Cycle Regulation' ($p = 8.52x10^{-4}$) as an enriched pathway in the DEGs at 16 dpp, following 'Gluconeogenesis I' ($p = 8.19x10^{-5}$) and 'LPS/IL-1 Mediated Inhibition of RXR Function' ($p = 4.84x10^{-4}$). 'Sperm Disorder' was an enriched function in this gene set as well ($p = 8.10x10^{-5}$). These results build upon the GO term enrichment analysis to suggest that much of the signal seen at 16 dpp is due to the loss of late-prophase subtypes, as well as cell cycle arrest.

To identify transcriptomic changes that precede either the appearance of EP cells or the mutant phenotype of meiotic prophase arrest in the mutant, we analyzed DEGs at 8 and 12 dpp. No significantly enriched GO terms were found in either the up-regulated or the down-regulated gene sets at either 8 or 12 dpp (both time points are before detection of any cytological anomalies). We did find an enrichment of the pathways 'Cell Cycle: G2/M DNA Damage Checkpoint Regulation' ($p = 5.16x10^{-3}$) and 'EIF2 Signaling' ($p = 1.04x10^{-2}$) among the top 10 most significant pathways. The G2/M DNA Damage Checkpoint would logically be activated in these cells with unrepaired DSBs, and EIF2 Signaling, a translational regulation program, could indicate post-transcriptional regulation as a component of the molecular response to absence of PRDM9. The highest scoring network in this DEG list is 'Cell Cycle, Cell Death and Survival, Endocrine System Disorders'. These findings demonstrate a transcriptomic signal for cell death even before the appearance of the EP cells where cell death may be manifest.

Differential expression of genes, especially down-regulated transcripts, could be due to a general response to genome-wide DNA damage in spermatocytes, or altered expression specifically of genes that undergo the ectopic DSBs that occur in *Prdm9*$^{-/-}$ germ cells (Brick et al., 2012). To discriminate between these alternatives, we compared the location of promoters of genes expressed in *Prdm9*$^{-/-}$ germ cells to the genomic locations of DSBs, using previously published data (Brick et al., 2012) on localization of DMC1, which is a widely accepted surrogate for sites of DSBs. Most genes with DMC1 peaks within their promoters in *Prdm9*$^{-/-}$ testes (Brick et al., 2012) were expressed in our dataset (74%); however, these genes were not biased toward being differentially expressed at any time point (Fisher's Exact Test, $p = 1$). Within those genes exhibiting a promoter DMC1 peak, the magnitude of the peak (frequency within the sample) was not correlated with the coefficient of differential expression for the gene (Spearman's rho = 0.001, $p = 0.98$). This result suggests that promoter-localized DSBs do not contribute significantly to the observed gene expression changes in *Prdm9*$^{-/-}$ germ cells. Instead, DEGs could be a part of a general response to DNA damage and/or due to changes in frequencies of specific meiotic substages in the cell populations (Figure 2.1), given meiotic arrest. Because these data revealed that transcriptomic changes in germ cells precede the onset of the characteristic cellular phenotype, we next sought to determine if cellular arrest was coupled with arrest of the spermatogenic transcriptomic program.

2.3.4. Transcriptomic progression is uncoupled from cellular progression in *Prdm9*$^{-/-}$ germ cells

Cytological arrest in a perturbed (mutant) system may *a priori* be expected to occur coincidentally with, or as an immediate consequence of, transcriptional arrest. To

test this expectation, we compared the expression of substage-specific genes in WT versus *Prdm9*$^{-/-}$ spermatocytes. Transcripts detected in the RNA-seq analysis derive from mixed pools of germ cells at different substages of meiotic progression. Therefore, any differences in transcript abundance due to *Prdm9* genotype could stem from either changes in the relative proportion of cells in different meiotic substages (Figure 2.1C) or from intracellular changes in expression within individual cells at the same substage. Permutation-based Maximum Covariance Analysis (PMCA, Methods) (Ball et al., 2016) utilizes two measurements made on the same sample to identify co-varying modules. Here, spermatogenic substage frequencies and RNA abundance were processed by PMCA to identify substage-specific lists of transcripts. PCMA-based identification and analysis of substage-specific transcripts in both WT and mutant samples allowed us to detect changes in substage-specific transcriptomic progression that are independent of changes in meiotic substage proportions in each germ-cell preparation.

We used previously defined lists of WT substage-specific genes (Ball et al., 2016) derived by PMCA (Methods) to compare the expression patterns of substage-specific genes between WT and *Prdm9*$^{-/-}$ germ cells (Figure 2.8A; Figure 2.9A and B), thereby determining the impact of the *Prdm9*$^{-/-}$ phenotype on developmentally unfolding transcriptomic program. Most differences in expression patterns of WT substage-specific genes in mutant germ cells reflected the differences in the relative substage frequencies in the mutant cell populations (Figure 2.1C) with only a few substage-specific transcripts being differentially expressed (Supplemental Table S4). However, a striking exception to this generalization is with respect to genes assigned by PCMA to WT EP and LP/D substages. Even though there was notable loss of these meiotic prophase substages

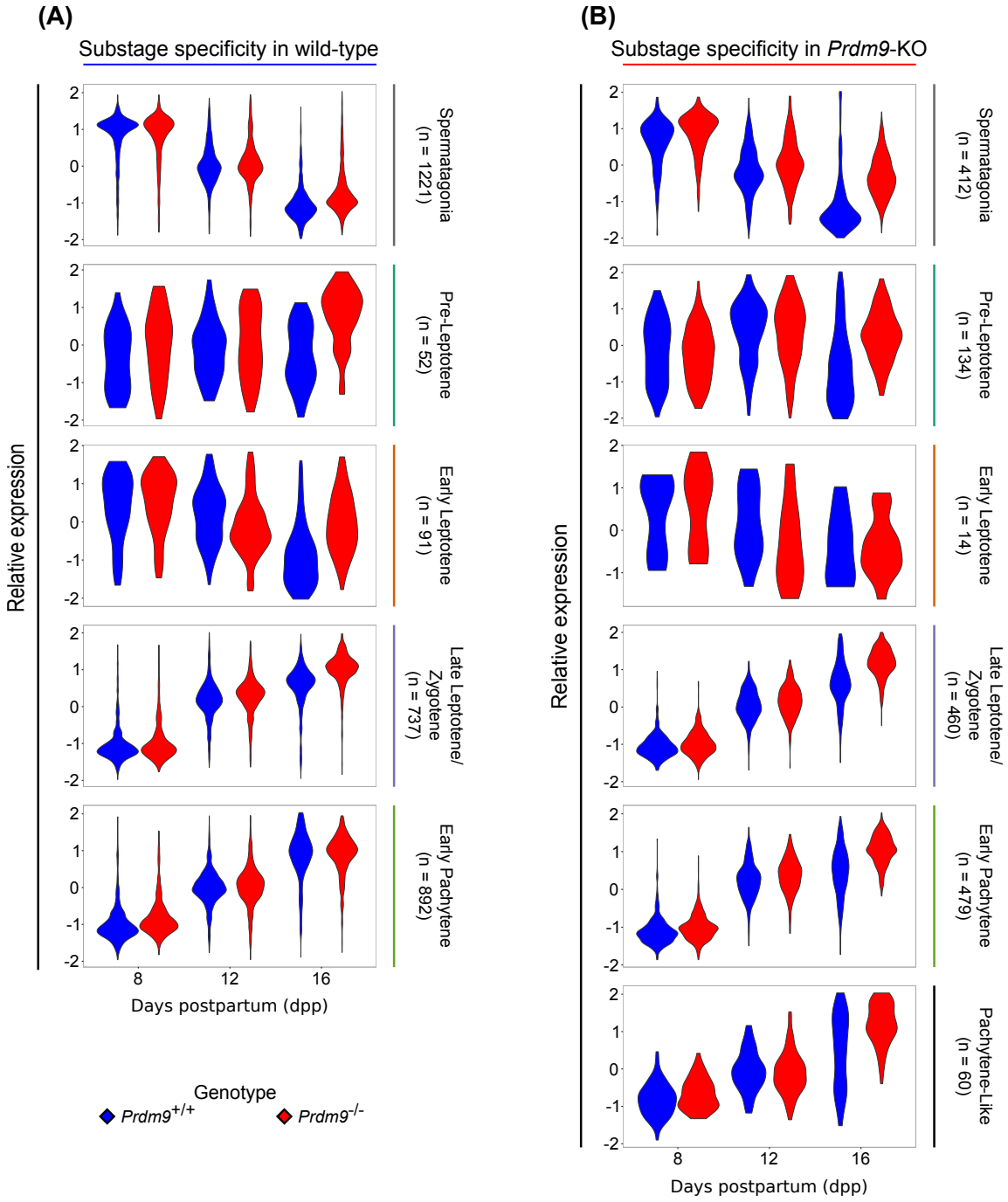

Figure 2.8. PMCA identifies substage-specific transcripts in wild-type and *Prdm9*$^{-/-}$ samples.
(A) Relative expression of transcripts assigned to each substage based on their wild-type expression patterns. (B) Relative expression of transcripts assigned to each substage based on their *Prdm9*$^{-/-}$ expression patterns.

among *Prdm9*$^{-/-}$ germ cells, there was no equivalent loss of expression of genes normally specific to these substages (Figure 2.9A; Figure 2.8A). Indeed, despite the notable decrease in abundance of EP cells at 12 and 16 dpp *Prdm9*$^{-/-}$ germ cells (Figure 2.1C), only approximately 17% of EP transcripts were differentially expressed between *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ samples (Figure 2.8A and Supplemental Table S4). And although there was a decrease in relative average expression of LP/D transcripts in 16 dpp *Prdm9*$^{-/-}$ samples (Figure 2.9A), only 33% of transcripts were differentially expressed (Supplemental Table S4). This was surprising because there are no cells cytologically characterized as LP/D cells in these samples (Figure 2.1C), and suggests that a spermatogenic transcriptomic program is being executed independently of the normally co-occurring cytological differentiation program. In general, the LP/D-specific transcripts detected in mutant spermatocytes exhibited either of two divergent expression patterns. Some 16 dpp LP/D transcripts in *Prdm9*$^{-/-}$ germ cells were at the same level as at 12 dpp. As mentioned above, only 33% of LP/D transcripts were differentially expressed between *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ germ cells at this time point. These transcripts could simply reflect the loss of pachytene spermatocytes by 16 dpp. However, some LP/D-specific transcripts in *Prdm9*$^{-/-}$ samples exhibited increase in expression level from 12 dpp to 16 dpp, with approximately 20% (n=771) of LP/D transcripts being differentially expressed between in *Prdm9*$^{-/-}$ germ cells 12 and 16 dpp (FDR < 0.01), trending similarly to their pattern in the WT transcriptomic progression (Figure 2.9A). This aberrant expression of pachytene- and diplotene-specific transcripts in *Prdm9*$^{-/-}$ cell population, despite dramatic decrease in the abundance of these cell types, suggests uncoupling of transcriptomic progression and cytological progression in *Prdm9*$^{-/-}$ germ cells.

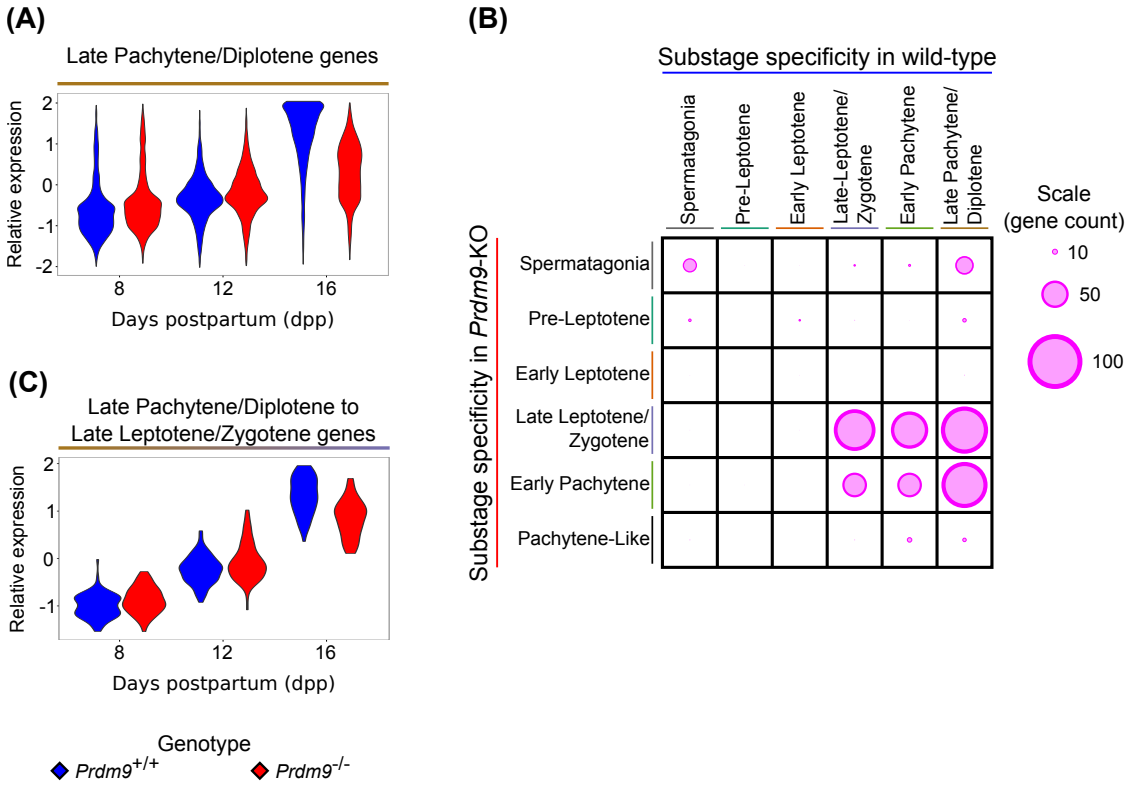

Figure 2.9. Substage specificity of transcripts determined from PMCA is different in Prdm9$^{-/-}$ germ cells than in wild-type germ cells
(A) Relative expression of transcripts assigned to Late-Pachytene/Diplotene substage based on their wild-type expression patterns. (B) Relative number of genes shared between substages assigned in wild-type and *Prdm9*$^{-/-}$ samples. The size of the circle represents the relative number of genes shared between two substage assignments. (C) Relative expression of transcripts assigned to Late-Pachytene/Diplotene in wild-type samples but to Late-Leptotene/Zygotene in *Prdm9*$^{-/-}$ samples.

To further investigate this apparent discrepancy between cytological stage and gene expression, we applied PMCA to assign transcripts to the specific substages detected by cytological markers in the *Prdm9*$^{-/-}$ germ-cell samples (FDR < 0.01, Figure 2.8B). We compared these *Prdm9*$^{-/-}$ substage-specific gene lists to those derived from PMCA analysis of WT data. First, relatively few transcripts from the mutants were assigned to early prophase substages, due to low variation in the cytological frequency of those substages across biological samples (Figure 2.1C); this had been reported previously for WT substage-specific transcription (Ball et al., 2016). This resulted in few

genes overlapping between PL and EL. Many transcripts in *Prdm9*$^{-/-}$ cell populations were identified with the same substage as they were in WT, despite the genetic perturbation (Figure 2.9B). However, a substantial number of genes that had been annotated to EP (n=68) and LP/D (n=85) in the original WT analysis were assigned to an earlier substage, LL/Z, in the PCMA analysis of the mutant transcriptomes (Figure 2.9C). This unanticipated observation is evidence for transcriptomic progression of the *Prdm9*$^{-/-}$ germ cells despite their apparent arrest or delay at the LL/Z substage. These transcripts contribute to the subset of LP/D genes that trend towards WT expression levels at 16 dpp in *Prdm9*$^{-/-}$ germ cells (Figure 2.9A). We compared these late meiotic prophase-specific genes expressed in mutants to genes expressed in more extensively purified WT pachytene spermatocytes (Ball et al., 2016). Forty-three of 68 EP transcripts that exhibit LL/Z specificity in the *Prdm9*$^{-/-}$ germ cells, and all 85 LP/D transcripts that instead exhibit LL/Z specificity in the mutant, were represented in the purified WT pachytene spermatocyte transcriptome. These analyses reveal that while cytology identifies a stage-specific meiotic arrest by EP in *Prdm9*$^{-/-}$ germ cells, some aspects of the spermatogenic transcriptomic program move forward unabated.

We used a bioinformatic approach to determine if the expression of LP/D genes in LL/Z *Prdm9*$^{-/-}$ germ cells might be driven by the same regulators that normally control the repertoire of LP/D genes, or if the evidence suggested a more stochastic response to the mutant phenotype. To identify candidate factors that might regulate the uncoupled molecular program in *Prdm9*$^{-/-}$ germ cells, we used iRegulon (Janky et al., 2014) to identify shared transcription factor (TF) binding sites among substage-specific genes. Analyses revealed enriched motifs for E2F1, REL, and YY1 at genes expressed in

Prdm9$^{-/-}$ EL, while ZFP143, ETV6, and ETV5 motifs were identified by genes specifically expressed in LL/Z and EP (Figure 2.10). In the mutant data, NR3C1 and QSOX1 potential binding sites were enriched near genes from the P-like list (Figure 2.10), so these could potentially be driving some of the transcriptional and meiotic arrest. Analysis of the mutant data led to identification of TF motifs that had previously been annotated in WT data as gene-expression regulators for specific substages, particularly the early prophase substages (Ball et al., 2016), as well as potential novel regulators of

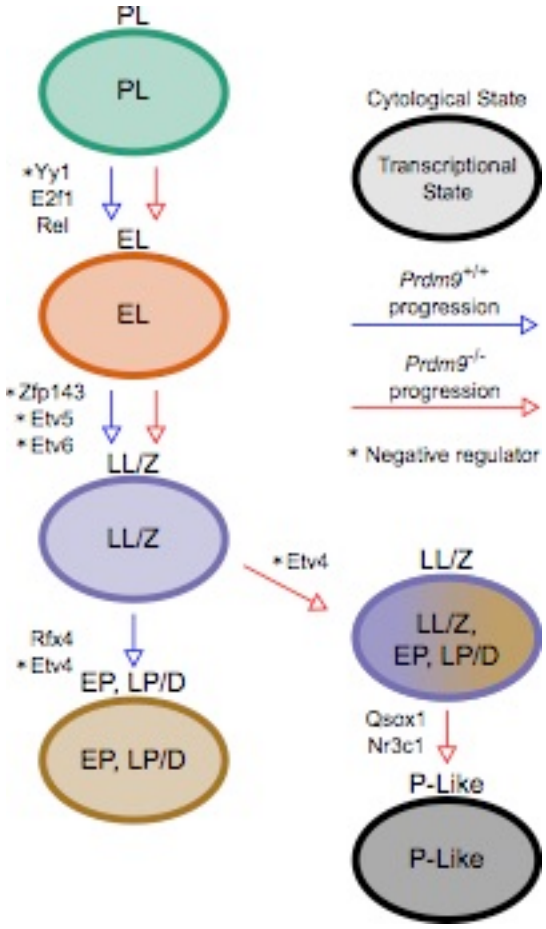

Figure 2.10. Summary model of cellular and molecular progression in *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ germ cells.
Arrows represent meiotic progression, colored by genotype as indicated. Text labels adjacent to the arrows indicate the transcriptional regulators of the genes expressed in the cell substage following the arrow.

the P-like stage. To identify prospective mechanisms underlying the uncoupling of transcriptomic programs from a cytological differentiation program, we conducted a TF motif analysis on the subset of genes that are specific to LP/D cells in WT testes, but

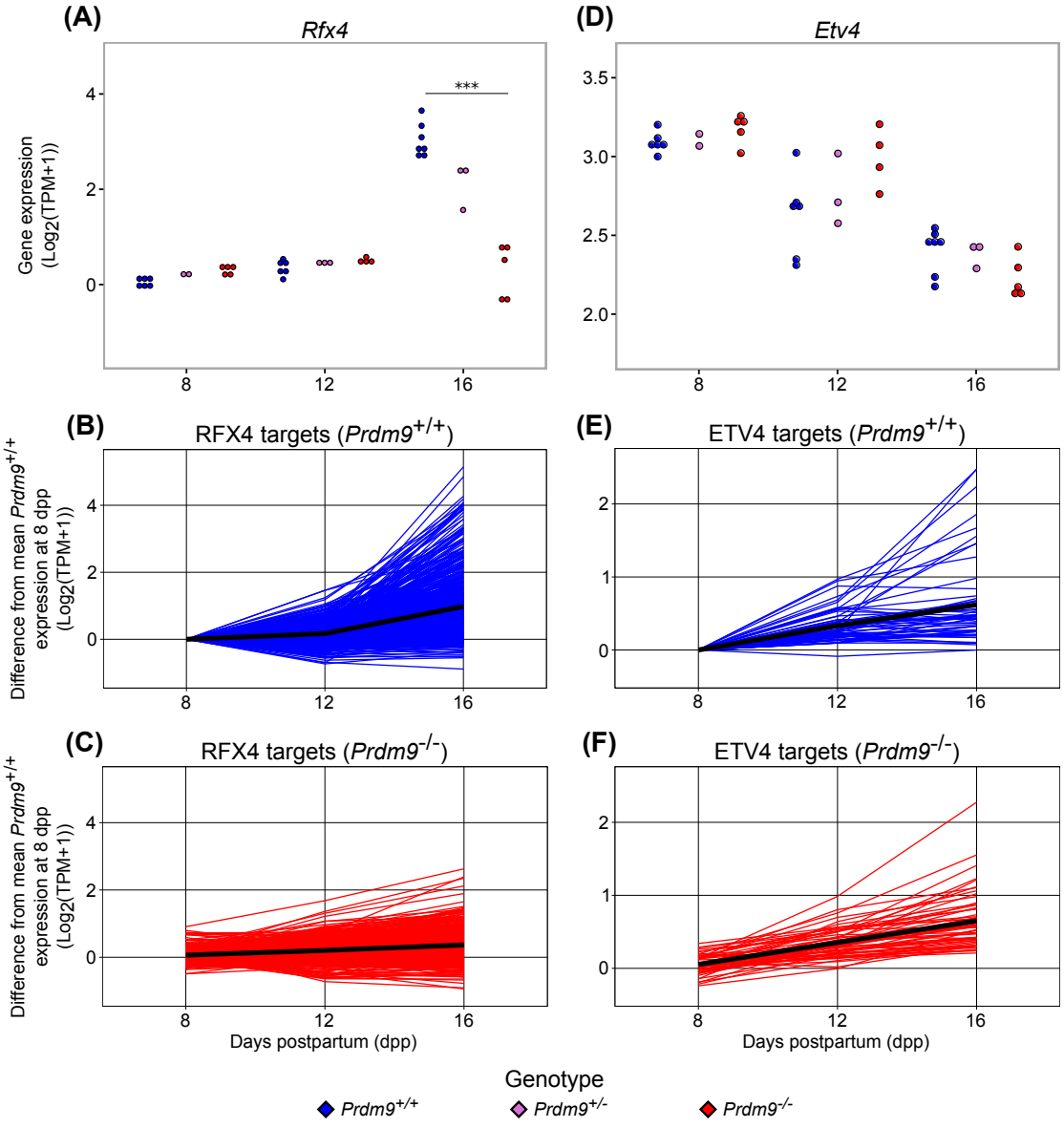

Figure 2.11. Upstream regulators of Late-Pachytene/Diplotene-specific genes show divergent expression changes in *Prdm9*$^{-/-}$ germ cells.
(A) Log_2(TPM+1) expression of activating transcriptional regulator *Rfx4* at 8, 12, and 16 dpp. (B,C) Relative expression of RFX4 targets in *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ samples, respectively. (D) Log_2(TPM+1) expression of repressive transcriptional regulator *Etv4* at 8, 12, and 16 dpp. (E,F) Relative expression of ETV4 targets in *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ samples, respectively. *** represents FDR < 0.0001.

aberrantly expressed in LL/Z cells in *Prdm9*$^{-/-}$ testes. Among the 56 TF motifs enriched at these genes was that of ETV4, which has been previously annotated as a regulator of the expression of LP/D genes (Ball et al., 2016). Interestingly, unlike *Rfx4*, encoding another regulator of LP/D, *Etv4* is not differentially expressed in *Prdm9*$^{-/-}$ germ cells compared to WT (Figure 2.11). This suggests that while some TFs that regulate LP/D transcription, such as RFX4, are not activating their targets, another subset of TFs, including ETV4, continue to regulate transcription even though the *Prdm9*$^{-/-}$ germ cells do not reach the LP/D substage. This mixed and seemingly uncoordinated response to meiotic arrest suggests that the uncoupling of molecular and cellular pathways in the *Prdm9*$^{-/-}$ mutant is not tightly regulated (Figure 2.10).

2.4.Discussion

In this study, we used a mutant spermatogenesis phenotype as a model to determine the correspondence between molecular and cellular differentiation programs when the developmental process is perturbed and/or abrogated. The *Prdm9* mutation is a robust model for this analysis because of the well-characterized morphological and cytological phenotypes caused by the absence of PRDM9, a key meiotic regulator protein. Normally, the spermatogenesis transcriptomic program and the meiotic gene expression programs are closely associated, running in parallel. However, by identifying transcriptomic signatures that match each of the known cytological substages and mutant phenotypes of *Prdm9*$^{-/-}$ germ cells, we found that the transcriptomic and cytodifferentiation programs of meiotic progression are partially uncoupled in mutant germ cells, most notably in two fundamental aspects. First, transcripts reflective of meiotic arrest and germ-cell death phenotypes were detectable before the relevant

characteristic cytological phenotypes appear. Second, mutant germ cells expressed transcripts typical of late prophase substages in spite of being arrested before the onset of these substages; thus, a "spermatogenic transcriptome program" moves forward in spite of abrogation of the unfolding "meiotic cytodifferentiation program." Overall, these two features yield stage-specific transcriptomes that are disassociated from their cell differentiation contexts. These findings not only further characterize the *Prdm9*$^{-/-}$ model of infertility, but are generally applicable to studies of developmental processes of differentiation and demonstrate the value of a combination of computational and cytological tools to address the challenges of assigning transcripts to cell types of origin, particularly in cases where single-cell RNA-seq is not feasible.

A problem that frequently impedes biological interpretation of mutant versus wild-type transcriptomic analyses is the inability to differentiate the causal differences that propel the mutant phenotype by either primary or secondary effects of the mutation. This problem is exacerbated in heterogeneous cell populations, such as in testes, where it is difficult to purify cell stages in a developmental lineage. We used computational analysis to relate transcriptomic data to cellular differentiation in a heterogeneous cell population where the proportions of component cell stages are known. Specifically, we applied Permutation-based Maximum Covariance Analysis (PMCA) (Ball et al., 2016) to assign transcripts to their cells of origin, both in wild-type and mutant cell populations. By thus analyzing transcriptomic differences at the substage-specific level, we determined expression differences that were molecular phenotypes of specific cell types, rather than attributable to sample-level variation in the proportions of specific substages. Together these methods allowed interpretation of transcriptomic changes as preceding or

following the observable phenotypes, thereby contributing to building a hierarchy of molecular and cytological phenotypes.

We showed that the transcriptomic and cytodifferentiation programs of meiotic progression are dis-associated in the *Prdm9*$^{-/-}$ germ cells. We identified two distinct points where this asynchrony, or uncoupling, of the molecular and cytological programs, is apparent. First, at 12 dpp in our samples, transcriptomic signatures of DNA damage checkpoint, activation, and repair appear in *Prdm9*$^{-/-}$ germ cells, preceding the specific cytological phenotypes that indicate unrepaired DNA damage (e.g., prevalence of the pH2AFX proteins indicative of DSBs). This seems biologically relevant, because we might expect that a future phenotype would be prefigured at the level of gene expression. The second example of asynchrony of transcriptomic and cytological phenotypes is not an anticipatory event prefiguring a phenotype, but instead is the unfolding of a transcriptome characteristic of later stages that never appear in the mutant. This is reflected in the shift in transcriptome profiles in *Prdm9*$^{-/-}$ germ cells as they exit the zygotene substage. Normally, this is followed cytologically by full chromosome synapsis, defining the pachytene substage. To the contrary, very few *Prdm9*$^{-/-}$ germ cells reach the early-pachytene substage and none reach the mid-pachytene substage. Instead, cells delay at late-leptotene/zygotene substages, and subsequently enter an abnormal pachytene-like state, with subsequent cell death. As expected, the transcriptomic signature for cell death coincides with the appearance of the pachytene-like germ cells. Additionally, however, the mutant germ cells delayed at the late-leptotene/zygotene substages express transcripts typical of the late spermatogenesis genes expressed in WT late-pachytene/diplotene cells (Ball et al., 2016), representing dis-association of the ongoing transcriptome program

from the delayed cytodifferentiation program. Indeed, the late-spermatogenesis transcriptomic signature of arrested spermatocytes suggests that regulatory programs promoting this transcription are at least partially functional in these germ cells, regardless of delay and arrest in cytological differentiation. The fact that the spermatogenic transcriptomic program is uncoupled from the differentiation program, and seemingly running on an independent clock, would not have been apparent without this two-fold transcriptomic and cellular analysis of the meiotic arrest mutant phenotypes.

Transcriptional-factor analyses reveal that the late-spermatogenesis transcripts expressed in *Prdm9*$^{-/-}$ germ cells appear to share common transcription factor binding sites. This suggests that these transcripts are co-regulated and that their WT-like expression in *Prdm9*$^{-/-}$ germ cells may be caused by the programmed processes, rather than random chance. One transcription factor with an enriched binding motif among these genes is ETV4, which is negatively correlated with the expression of its targets. Therefore, the developmentally premature decline in *Etv4* expression could lead to the observed up-regulation of spermiogenic transcripts. In contrast, the late-spermatogenesis transcripts that fail to be expressed (or are down-regulated) in the mutant spermatocytes are enriched for the binding motif of RFX4, which is positively correlated with the expression of its proposed targets. In this case, the root regulatory event may be failure to express the transcription factors in the mutant germ cells, but here, that failure also leads also to the down-regulation of its targets. Together, these observations suggest that transcription factor down-regulation can incongruently lead to both up-regulation and down-regulation of targets, explaining the apparent asynchrony between cellular and transcriptomic stages in mutant germ cells undergoing meiotic arrest. Thus, these

findings put forth a model of meiotic arrest in which there is asynchrony of transcriptomic and cytological differentiation programs, each revealing independent autonomy.

In addition to the biological implications, divergent or asynchronous programs in phenotypes of developmental arrest present challenges for biological data interpretation. The premature expression of late-spermatogenesis transcripts in inappropriately early meiotic substages in *Prdm9*$^{-/-}$ germ cells is obviously not sufficient to avoid meiotic arrest or propel spermatocytes to the cellular stage appropriate for the transcript expression. Although it is not known if the prematurely expressed transcripts are translated in *Prdm9*$^{-/-}$ spermatocytes, it is not likely that their translational state could rescue meiotic arrest, especially since some of the transcripts are destined for post-meiotic translation. Nonetheless, an unsolved problem that transcriptomic analyses cannot address is the degree of developmental autonomy and synchrony between cytological differentiation and the programs of protein translation and activation. Moreover, finding transcriptomic signatures at variance with cytological stage has implications for the interpretation of bulk RNA-seq from other mutants, especially those with less well-characterized cytological phenotypes. As this study reveals, where transcriptomic data may not reflect corresponding cellular changes, it is essential to have a complement of classical cytological measures and a method for integration of the two types of data. Much of the benefit of single cell RNA-seq (scRNA-seq) relies on the ability to infer the cellular context of a cell based on its transcriptomic signature, but our study suggests that uncoupled molecular and cellular processes would complicate such inferences, absent a cytological characterization to accommodate the results. Determination of cellular

context and autonomous transcriptomic programs in a developmental mutant might be facilitated with scRNA-seq data; however, the possibility of cellular and molecular uncoupling should be considered even when analyzing single-cell data. In particular, cell-type-specific molecular markers may become unreliable identifiers of cellular states when standard transcriptional programs fragment. This phenomenon is unlikely to be unique to spermatogenesis and warrants investigation in other developmental and differentiation contexts.

2.5. Acknowledgements

We thank the Handel and Carter laboratories for discussions, and gratefully acknowledge Sabrina Petri for animal care. We also thank Drs. J. Trowbridge, A. Yee, C. Cowan, and S. Munger for feedback on the project and comments on the manuscript. This work was supported by the NIH/National Institute of General Medical Sciences (NIGMS) grant P01 GM099640 (GWC and MAH), and by a fellowship from the Eunice Kennedy Shriver National Institute of Child Health & Human Development grant T32 HD007065 (ADF).

2.6. Contributions

This project was conceived and planned by myself with Drs. Carter and Handel. Yasuhiro Fujiwara performed the cell preparations and cytological analeses. Robyn Ball contributed to the data preparation and computational analyses. I led the majority of the computational analyses, the figure preparation, and the writing of the manuscript.

CHAPTER 3: Modeling the multiple zinc finger protein PRDM9 binding affinity with Affinity-seq meiosis

Arat, S.*, Fine, A. D.*, Walker, M., Billings, T., Paigen, K., Petov, P. M. & Carter, G. W.

To be submitted.

3.1. Introduction

Zinc-finger (ZF) proteins were first identified in the 1980s (Brown et al., 1985; Gibson et al., 1988; Miller et al., 1985; Vrana et al., 1988) and have since been annotated to play major roles in numerous biological systems and diseases (Cassandri et al., 2017; Ladomery and Dellaire, 2002). ZF proteins have been found to be crucial regulators of cellular and molecular processes, including orchestration of development, maintenance of healthy adult systems, and success of gametogenesis (Carballo et al., 1998; Elton et al., 2015; Hayashi et al., 2005; Kim et al., 2002; Lomniczi et al., 2015; Nakamura et al., 2004). Therefore, it is unsurprising that misfunction of ZF proteins can lead to a number of diseases, including cancer, neurological disorders, and hematological diseases, among others (Arnaud et al., 2010; de Castro-Catala et al., 2017; Gustafsson Sheppard et al., 2012; Mastrangelo et al., 2000; van Haaften-Visser et al., 2017). Despite the fact that ZF proteins are implicated in multiple biological processes, their diversity and complexity have made identifying their specific functions and mechanisms challenging (Cassandri et al., 2017; Fedotova et al., 2017).

The defining characteristic of ZF proteins is their binding domain, which can bind to a number of molecules in a cell, including DNA, RNA, and proteins (Cassandri et al., 2017; Fedotova et al., 2017; Fu and Blackshear, 2017; Pavletich and Pabo, 1991). ZF domains are independently folded domains of a protein coordinated by zinc ions, which bind specifically based on their sequence (Choo et al., 1994; Klug, 2010; Laity et al., 2001; Pavletich and Pabo, 1991). ZF proteins typically fall into one of the following major classes of ZFs: C2H2, RING, PHD, and LIM (Cassandri et al., 2017). However, ZF proteins can contain additional functional elements, that diversify their function. For

example, Krüppel associated box (KRAB) domains, which facilitate binding to other proteins, are commonly found in ZF proteins, but other domains can facilitate other functions like perform epigenetic modifications (Lupo et al., 2013; Urrutia, 2003). With these additional functions, a common role for ZF proteins is gene expression regulation by directly regulating transcription or translation (Cassandri et al., 2017; Klug, 2010; Pavletich and Pabo, 1991). While the alternative domains of ZF proteins play a major role in this regulation, it is still the binding sequence of the ZF array that determines its localization, and therefore its specific function. Identifying where ZF proteins bind is therefore an important aspect of deciphering their function, and thereby anticipating how their dysfunction will affect a tissue.

It has long been known that the amino acid sequence of ZF arrays determines their DNA binding location and affinity. The crystal structure of a short ZF-to-DNA interaction was identified in 1991, revealing the basic principles of ZF binding (Pavletich and Pabo, 1991). Namely, each finger recognizes a 3-bp sequence in DNA. However, with further investigation, it became apparent that adjacent fingers can affect binding specificity, and that a single finger likely binds based on overlapping 4-bp sequences (Isalan et al., 1997; Isalan et al., 1998). Despite decades of understanding how a ZFs binds to DNA, predicting the specific DNA-binding sequences for long ZF arrays has remained a challenge in the field. Currently, tools exist that can predict ZF binding with limited, varying accuracy (Gupta et al., 2014; Kaplan et al., 2005; Persikov and Singh, 2014). Generally, short arrays can be predicted fairly well; however, longer ZF arrays are less easily interpreted.

Understanding how ZF proteins select binding sites can help predict the effect of their dysfunction and enable utilization of their molecular functions to regulate gene expression exogenously. There are two main mechanisms by which variation could alter ZF function, by which a greater understanding of ZF binding would allow us to predict ZF dysfunction. First, models of ZF binding would increase our ability to interpret local variation in the binding domain of ZF proteins. If a variant alters where a ZF protein binds, the novel binding sites could be informative of phenotypic changes. Second, variation in distal binding sites that could affect ZF binding would add to this understanding. ZFs can be crucial for proper expression of key genes, especially in development, so disruptions to any of their target sites could have significant consequences to cellular functions. Apart from interpreting natural variation, ZF proteins have been proposed as machinery to exogenously regulate gene expression. There have been numerous examples of utilization of zinc-finger proteins to alter gene expression in a controlled manner (Choo et al., 1994; Jamieson et al., 2003; Klug, 2010; Reynolds et al., 2003). However, this work has been limited by an incomplete understanding of where these proteins will bind, risking off target effects (Klug, 2010; Moore et al., 2001; Perez et al., 2008). While this can be assessed at a single-protein level (Walker et al., 2015), computational tools that could predict binding sites of ZF arrays would enable this to work even better.

PR/SET domain 9 (PRDM9) is an intriguing model for ZF binding. PRDM9 has a long zinc-finger array, a KRAB protein-protein interaction domain, and a SET histone methyltransferase domain. It interacts with other proteins to activate homologous recombination hotspots, at least partially through the deposition of crucial histone

modifications. Understanding the binding specificity and affinity of PRDM9 is important because not only is its function required for successful gametogenesis in many mammals, but also its ZF array is among the most common subtype of ZFs, C2H2 ZF proteins. Predicting its binding patterns would inform both fertility research and ziZF biology. Previously, the binding affinity of $PRDM9^{Dom2}$ was assessed in the C57BL/6J background with Affinity-seq, which measures binding frequencies of a DNA binding domain to naked genomic DNA, thus eliminating complications such as chromatin state. Utilizing this assay for other genomic backgrounds allows introduction of natural variation that can subsequently be incorporated into a model of long ZF binding affinity. Here, we perform Affinity-seq on $PRDM9^{Dom2}$ in the CAST/EiJ background and integrate those findings with the previous analysis to build a computational model of long ZF binding.

3.2. Methods

3.2.1. Data collection

3.2.1.1. Affinity-seq & sequence analysis

We performed Affinity-seq using the $PRDM9^{Dom2}$ allele in the CAST/EiJ genome according to the previously published methods (Walker et al., 2015). Briefly, we allowed HA-tagged zinc-finger array from $PRDM9^{Dom2}$ to bind to naked DNA. We used this tag to isolate regions of the genome that were bound by the zinc-finger array and sequenced at 100-bp reads using Illumina HiSeq 2500. Peaks were called using MACS2 ($p < 0.01$). Binding sites were interpreted to be a 36-bp segment, to reflect the length of the $PRDM9^{Dom2}$ zinc-finger array, that corresponded to the annotated motif for this allele of

PRDM9. Additional Affintiy-seq data was obtained from NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61613) (Walker et al., 2015).

3.2.1.2. Sample normalization

All peaks were put onto a common coordinate system. Log transformed C57BL/6J and CAST/EiJ PRDM9^{Dom2} Affinity-seq peaks were normalized between these genomes to have mean-centered distrbutions with a scale factor of 1.68166.

3.2.2. Computational methods

3.2.2.1. Nucleotide frequency prediction

Binding sites were predicted using previously published methods for zinc-finger binding sire prediction, based on expectation maximization (Kaplan et al., 2005), support vector machine (Persikov et al., 2009), and random forest (Gupta et al., 2014) approaches.

3.2.2.2. Linear models

We performed linear regressions on this dataset using the lm() function in the stats package in R (R Core Team, 2015). This model fit variation in binding affinity to the sequence variation across the PRDM9^{Dom2} binding site, as well as some interactions between nucleotides at different positions.

3.2.2.3. Iterative Random Forest

To determine the interactions, we utilized an iterative random forest (iRF) model, which trains a feature-weighted random forests to identify "stable" high-order

interactions with accurate prediction in a supervised fashion. We built an iterative random forest (iRF) model using iRF package in R (Basu et al., 2018), with the following options: n.iter = 5, ntree = 500, n.bootstrap = 20. We then took the intersection of the stable interactions that show up in all 5 iterations (86 interactions) and the intersection of the stable interactions that show up in the 2nd, 3rd, 4th and 5th iterations (9 interactions). There are 94 2-way interactions, and 1 3-way interaction, G8:A30:T31. Finally, to determine if these 94 2-way interactions were statistically significant, a two-sample t-test was performed by comparing the interaction effect with the additive effect. We then performed a Benjamini-Hochberg FDR correction. If the FDR-corrected p-value is < 0.05, then the interaction is considered statistically significant; i.e. significant interaction means the interaction effect is significantly more (or less) than the additive effect. Overall, there are 27 significant "stable" interactions that can help explain PRDM9 binding preference.

3.2.2.4. DNA shape

DNAShape is a high throughput prediction method that uses a 5-mer sliding window and all-atom Monte Carlo simulations to derive the structural features of DNA sequences: minor groove width (MGW), helix twist (HelT), propeller twist (ProT) and Roll (Zhou et al., 2013). Since these DNA shape features play an important role in protein-DNA binding specificity, we used DNAShapeR package, an R package for DNAShape features, to predict DNA shape of the DNA sequences that PRDM9 binds with default setting (Zhou et al., 2013).

3.3. Results

3.3.1. PRDM9 binding specificity is poorly explained by existing models

We first characterized the *in vitro* binding affinity of the long zinc-finger (ZF) protein, PRDM9. We used previously published data for binding frequency of the $PRDM9^{Dom2}$ ZF array to naked DNA from C57BL/6J (B6) mice as a surrogate for $PRDM9^{Dom2}$ binding affinity (Walker et al., 2015). We used the previously identified 31,770 Affinity-seq PRDM9 binding sites, trimmed these sites to the central 36 bp region, to reflect the 12 ZFs of $PRDM9^{Dom2}$ and its 1:3 ZF-to-bp binding pattern. Previously, a motif analysis had been performed on $PRDM9^{Dom2}$ based on these data, as well as a quantification of the frequency with which a SNP in the CAST genome created a novel binding site for $PRDM9^{Dom2}$, which is endogenous for the B6 genome (Figure 3.1A) (Baker et al., 2015). Importantly, these Affinity-seq binding sites have been shown to correspond to *in vivo* binding sites of PRDM9 (Walker et al., 2015). These binding sites and their associated affinity can now be used as a model of ZF binding.

We compared the motif and SNP frequency to the percentage of each nucleotide at each base position (Figure 3.1B). As fully deterministic binding based on 36 bases would be incredibly rare across the genome, it was no surprise that there was sequence variability between our PRDM9 binding sites. That said, we identified five key bases that housed a specific nucleotide with a frequency of greater than 90%, henceforth called 'anchor positions' (Figure 3.1B). These anchor positions comprise a G at position 8, a G at position 11, a T at position 13, a C at position 15, and a T at position 16. These are distinct to $PRDM9^{Dom2}$, as the binding sites of the $PRDM9^{Cst}$ allele contained a different binding motif (Baker et al., 2015; Baker et al., 2014). Outside of the anchor positions, the

region surrounding position 31 was particularly interesting, as this region seemed relatively inconsequential by the motif analysis, but was capable of creating many new binding sites by single mutations (Figure 3.1A). Despite indeterminate frequencies of nucleotides at various bases across the binding site, no location had a zero SNP count, indicating that binding preference is aggregated across all 36 bases in this binding motif, not just the anchor positions.

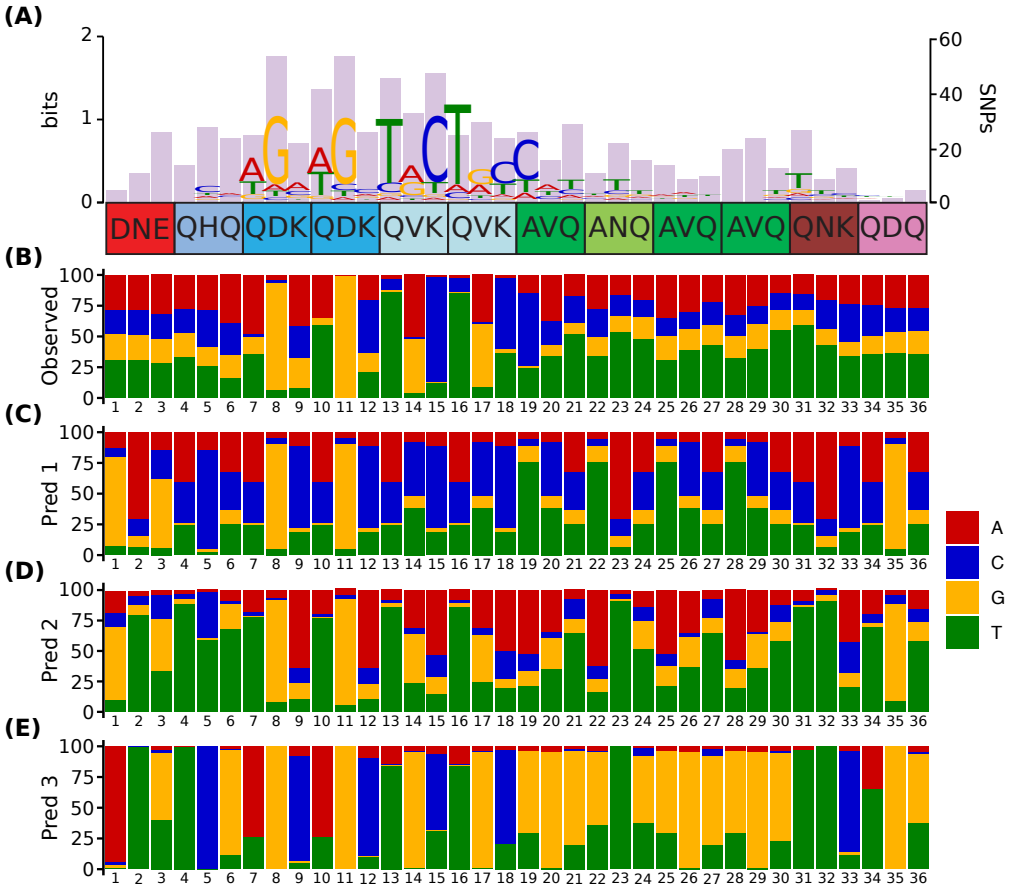

Figure 3.1. Current tools fail to predict PRDM9^{Dom2} binding sites.
Observed and predicted binding sites for PRDM9^{Dom2}. (A) Binding motif for PRDM9^{Dom2} in the B6 genome (letters, left scale) and SNP frequency across that motif in novel binding sites found in the CAST genome (bars, right scale). (B) Observed frequencies of each nucleotide found across the PRDM9^{Dom2} affinity-seq sites. (C-E) Predicted frequencies of each nucleotide for PRDM9^{Dom2} based on published computational tools.

To attempt to resolve this, we tested the ability for existing bioinformatic methods to predict these PRDM9 binding sites. We compared the measured nucleotide frequencies (Figure 3.1B) to computationally predicted nucleotide frequencies (Figure 3.1C-E) (Gupta et al., 2014; Kaplan et al., 2005; Persikov et al., 2009). While these methods demonstrated an improvement over background nucleotide frequencies for some base positions, none of them could fully recapitulate the PRDM9^{Dom2} binding site. These results demonstrate the need for a more robust model for the binding affinity of long ZF arrays.

3.3.2. Single-base model of PRDM9 binding corresponds to SNP frequencies but poorly correlates with observed binding affinity

Next, we wanted to introduce natural variation to our peakset to assess the importance of given nucleotides at certain base-pair positions. We performed Affinity-seq again for PRDM9^{Dom2}, but this time, in naked DNA from CAST/EiJ (CAST) mice. There are SNPs in the CAST genome at approximately every 150 bp, compared to the B6 genome. For Affinity-seq peaks shared between the B6 and CAST genomes, 14841 peaks were identified at genetically identical binding sites, while 2083 peaks were found to contain at least one SNP between the two. By comparing the genetically identical Affinity-seq sites, were able to scale the CAST sites to have a more similar range of binding affinities (Figure 3.2). These two datasets now allow us to model PRDM9^{Dom2} binding *in vitro*, as well as test the power of that model in a second genetic context.

Next, we modeled the effect of having any nucleotide present across the binding motif for PRDM9^{Dom2}. We ran a linear regression on our B6 dataset to fit variance in read counts for each site, as a surrogate for binding affinity, to each nucleotide at each base

across all binding sites (Methods). We used the coefficients from that regression to dictate the estimated effect of that nucleotide on binding affinity (Figure 3.3). Three nucleotides are shown for each base, with the hidden bases contributing to the baseline estimate (Intercept = -1.720036). Interestingly, some positions show a strong preference for a single nucleotide (ie. T at position 31), while others showed a strong preference against a single nucleotide (ie. C at position 11). Notably, this model demonstrates value

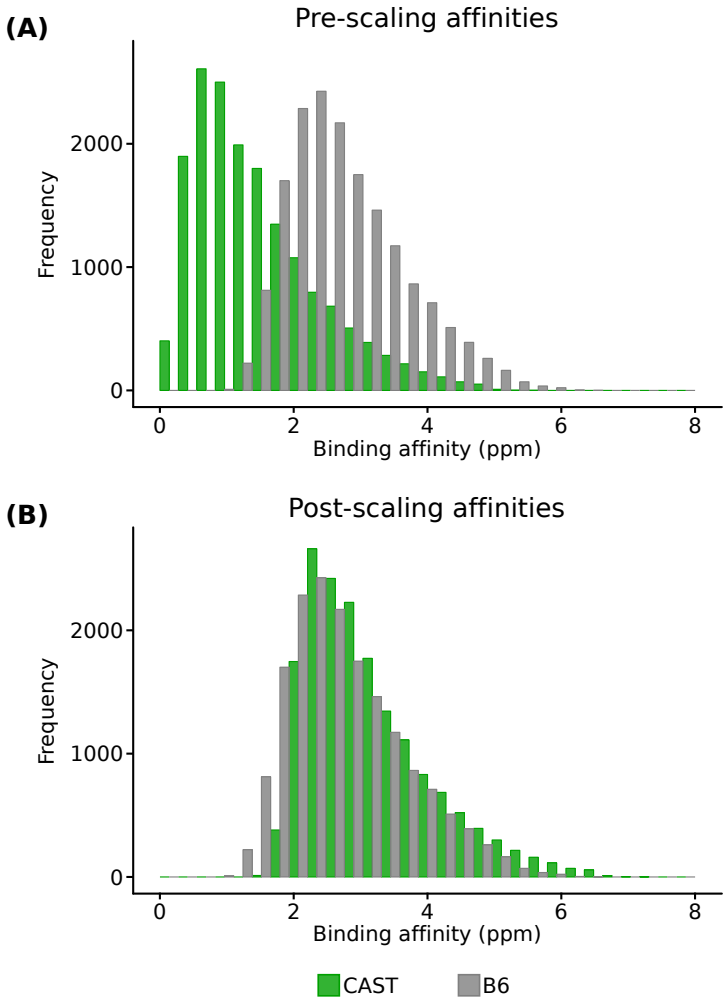

Figure 3.2. Binding affinity were scaled to be comparable between CAST and B6. (A) Distributions of PRDM9^{Dom2} binding affinities between matching genetic sequences in CAST and B6 were initially difference. (B) The application of a scale factor between these two datasets corrected for range differences.

beyond the frequency-based motif analysis, as nucleotides with relatively low representation in the motif (ie. C at position 7) showed similar estimated effects on binding affinity in our model. Because of their relative invariability, it is unsurprising that some anchor, particularly 16, has high standard error for each estimate. Overall, this model identified testable estimates for nucleotide effects on PRDM9^{Dom2} binding affinity.

To test our model, we calculated its ability to predict PRDM9^{Dom2} binding affinities in the B6 and CAST genome. First, we identified the PRDM9^{Dom2} binding sites that had identical genetic sequences between B6 and CAST (matched sites, n=14841). While not perfect, there was a high degree of correlation between the observed binding affinity for this subset of binding sites (Figure 3.4, $r^2 = 0.7582$, $p < 0.0001$). Therefore, it

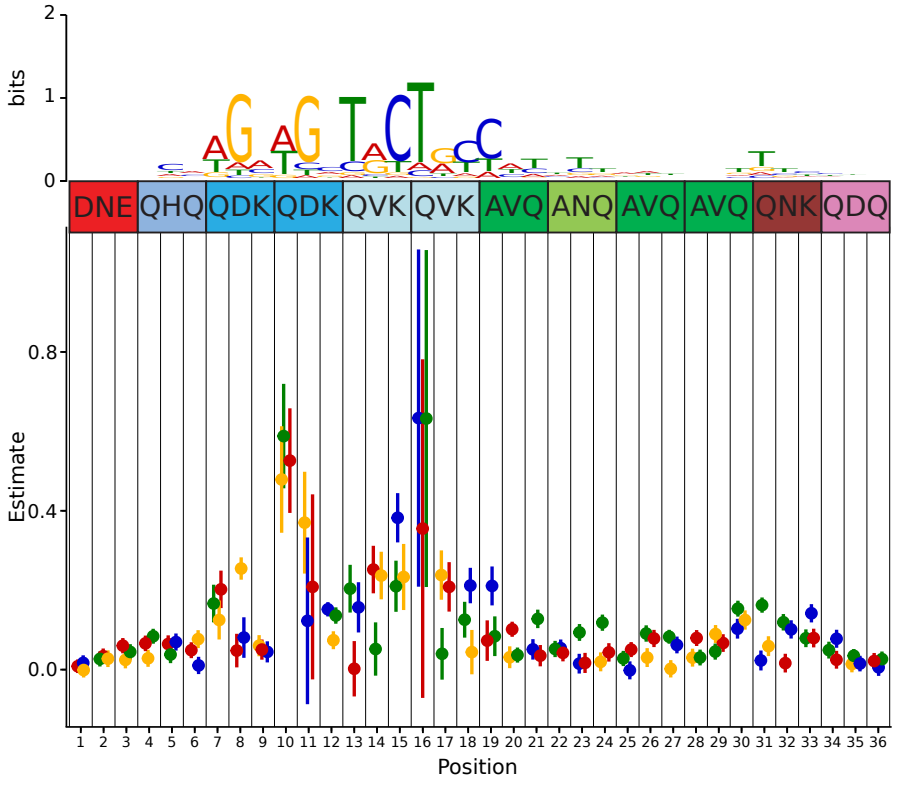

Figure 3.3. Single base-pair model of PRDM9^{Dom2} binding affinity.
Estimated effects on binding affinity of nucleotides across the binding motif for PRDM9^{DOM2}. Each point represents a change from the lowest ranked nucleotide.

was unsurprising that our single base-pair model predicted the binding affinity for the B6 ($r^2 = 0.06923$, $p < 0.0001$) and CAST ($r^2 = 0.0461$, $p < 0.0001$) matched sites similarly (Figure 3.5A,B). While slightly positively correlated, the range of predicted affinities was drastically less than the observed affinities, limiting the ability for this model to successfully predict PRDM9^{Dom2} binding. This was also true of the binding sites with genetic variation between the B6 ($r^2 = 0.08536$, $p < 0.0001$) and CAST ($r^2 = 0.09845$, $p < 0.0001$) genomes (Figure 3.5C,D). These findings demonstrate that a single base-pair model of PRDM9^{Dom2} binding is insufficiently able to predict its ZF array's binding affinity to a given genetic sequence.

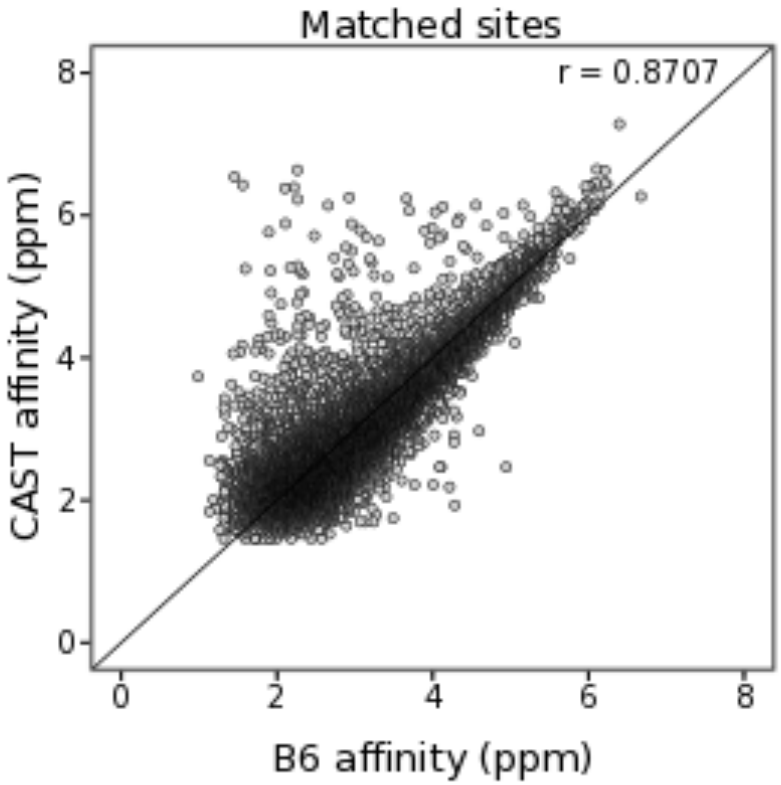

Figure 3.4. PRDM9^{Dom2} affinity on genetically identical binding sites in B6 and CAST genomes.
The matched B6 and CAST sites are highly correlated between B6 and CAST.

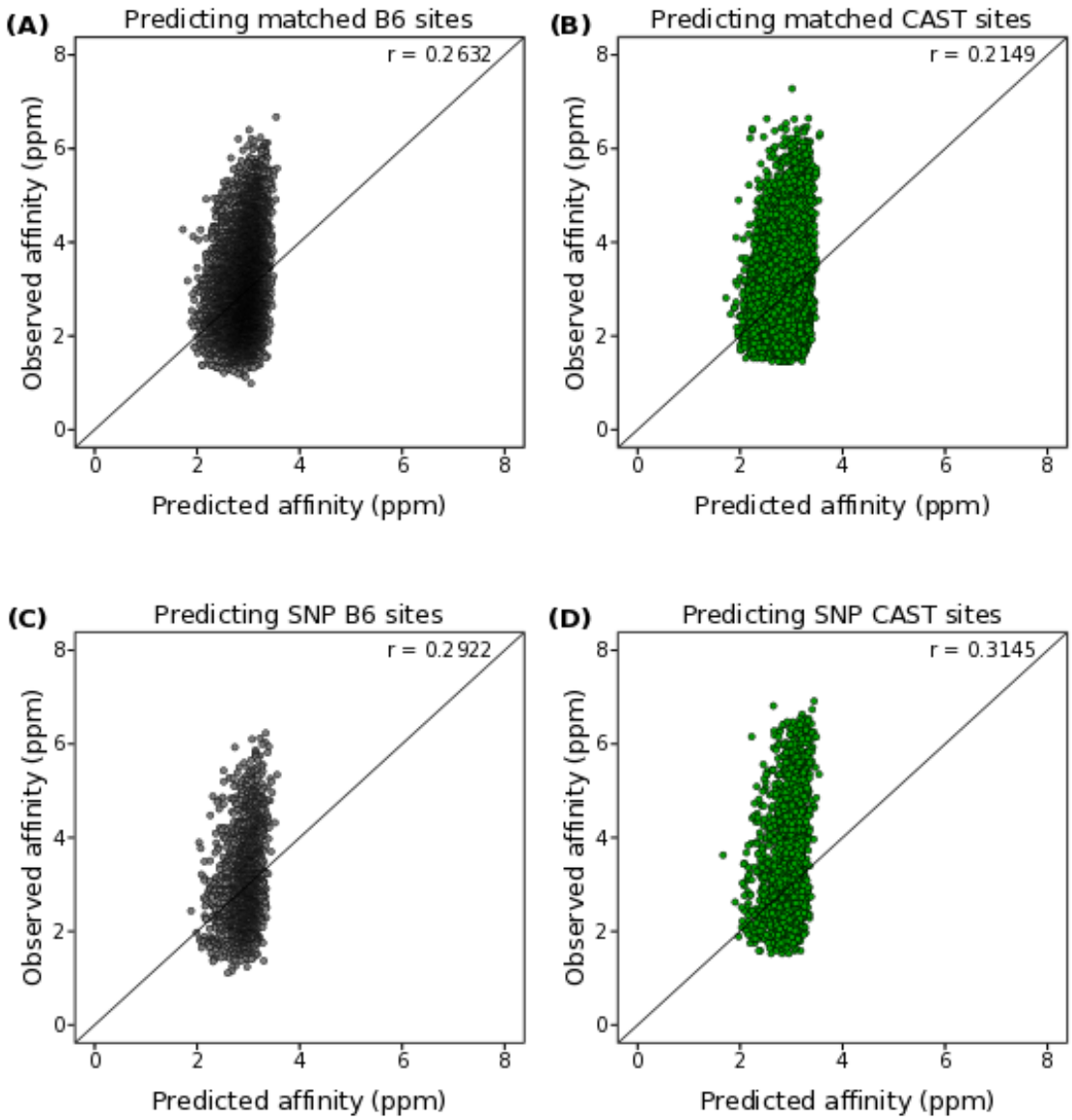

Figure 3.5. Predicted vs observed binding affinity of single base-pair model of PRDM9^{Dom2}.
The single base-pair model fails to predict binding affinity of PRDM9^{Dom2}. (A,B) Predicting the binding affinity of PRDM9^{Dom2} in the B6 (A) and CAST (B) genomes when there are identical genetic sequences in the binding sites for each. (C,D) Predicting the binding affinity of PRDM9^{Dom2} in the B6 (C) and CAST (D) genomes when there are SNPs found between the two genomes.

In addition to our model's ability to assess PRDM9^{Dom2} binding to a genetic sequence, we wanted to measure its ability to predict the effect of a SNP within the PRDM9^{Dom2} binding site on affinity. We took all sites containing one or more SNPs

between B6 and CAST and calculated the change in binding affinity predicted by our single base-pair model. We then compared this predicted to the observed change in binding affinity (Figure 3.6A, $r^2 = 0.1974$). Like the full-sequence binding affinity prediction, our predictions lacked the range of observed changes. However, there was a greater degree of correlation in our SNP effect prediction than the full sequence prediction. We were curious if some bases were better predicted by this model than others, so we examined each individual base on its own. We found a range of correlations across the positions in the binding site, with some having relatively high predictive power (Figure 3.6B, $r^2 = 0.3811$, $p < 0.0001$) and some having little predictive power (Figure 3.6C, $r^2 = 0.01921$, $p < 0.0001$). A common trend we noticed at the single-position level was SNPs that had varying observed effects on affinity, such as in Figure 3.6B. While each point represents a SNP within an otherwise identical genetic sequence, there was variation in the rest of the sequence between points. This suggests that there are interacting effects between bases that cause a single SNP to have varying effects on binding affinity across different genetic contexts. Taken together, these results demonstrate the need for a more complex model of ZF array binding beyond a single base-pair level.

3.3.3. Multi-base models of PRDM9 binding improves correlation with binding affinity

We generated a new model of $PRDM9^{Dom2}$ binding affinity based on each single base-pair, as well as all two-way interactions between nucleotides (Methods). Like with our single base-pair, model, we then tested how well this model could predict observed binding affinity in B6 ($r^2 = 0.4442$, $p < 0.0001$) and CAST ($r^2 = 0.323$, $p < 0.0001$) sites without any genetic variation (Figure 3.7A,B). Notably, this two-way interaction model

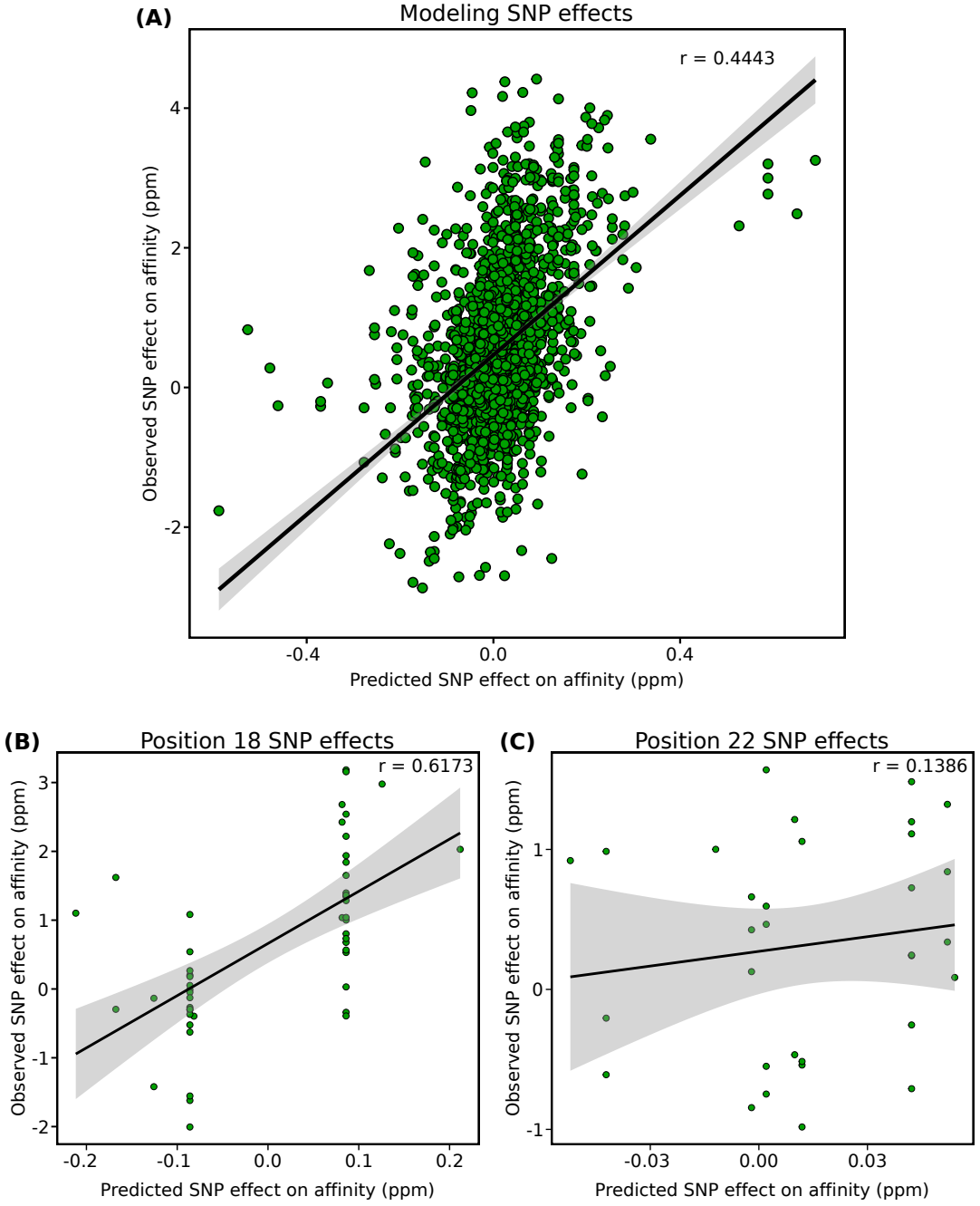

Figure 3.6. Single base-pair model predicts SNP effects between B6 and CAST genomes on PRDM9^{Dom2} binding.
PRDM9^{Dom2} Binding affinity in the CAST genome for sites with SNPs was predicted based on the single base-pair model and compared to observed binding affinities to those sites. (A) The predicted and observed effects on affinity for all positions. (B) Example of position 18, where the model's prediction was well-correlated with observed changes. (C) Example of position 22, where the model's prediction was not well-correlated with observed changes.

better predicts the binding affinity of PRDM9^{Dom2} to the matched B6 and CAST sites that the model was trained on compared to the single base-pair model (Figure 3.5A,B). However, linear models are prone to overfitting, and this model has over 4,000 terms. To test this, we evaluated our ability to predict the PRDM9^{Dom2} affinity to the subset of B6 ($r^2 = 0.4927$, $p < 0.0001$) and CAST ($r^2 = 0.0003157$ $p = 0.4176$) sequences that contained SNPs (Figure 3.7C,D, Figure 3.8). The B6 set of sequences were included in our training set, so it was unsurprising that the observed affinities were well correlated with our predictions (Figure 3.7C). The observed affinities for CAST sequences containing SNPs were not well correlated with our predictions, suggesting out model was over fitting these data (Figure 3.8). Even when we exclude sites that were predicted beyond the range of observed affinities, observed affinities were not as well correlated as in our training set (Figure 3.7D, $r^2 = 0.2659$, $p < 0.0001$). These findings demonstrate the need for a model of PRDM9^{Dom2} binding broader than the single base-pair model but without the number of terms that the two-way interactive model contains, in order to avoid over fitting our model.

3.3.4. Random Forest identifies relevant interactions across PRDM9^{Dom2} binding site

We performed Random Forest on these data to identify the important multi-base interactions for PRDM9 binding affinity. We classified 93 two-base interactions (Figure 3.9) and 1 three-base interaction (G8:A30:T31, effect = -0.08841). This model was not as predictive as all two-base interactions (Figure 3.7, Figure 3.10A,B) for B6 ($r^2 = 0.09154$) or CAST ($r^2 = 0.0638$); however, it did not appear to over fit to the training dataset (Figure 3.8, Figure 3.10C,D). Generally, effects were relatively small for any single- or two-base term, with 31 being the most impactful position, with varying effects based on

which nucleotide is present. The only positions to not participate in a modeled interaction were 11, 13, and 16, all of which are anchor positions. This is likely due to the underrepresentation of non-ideal nucleotides at those positions. Despite there being

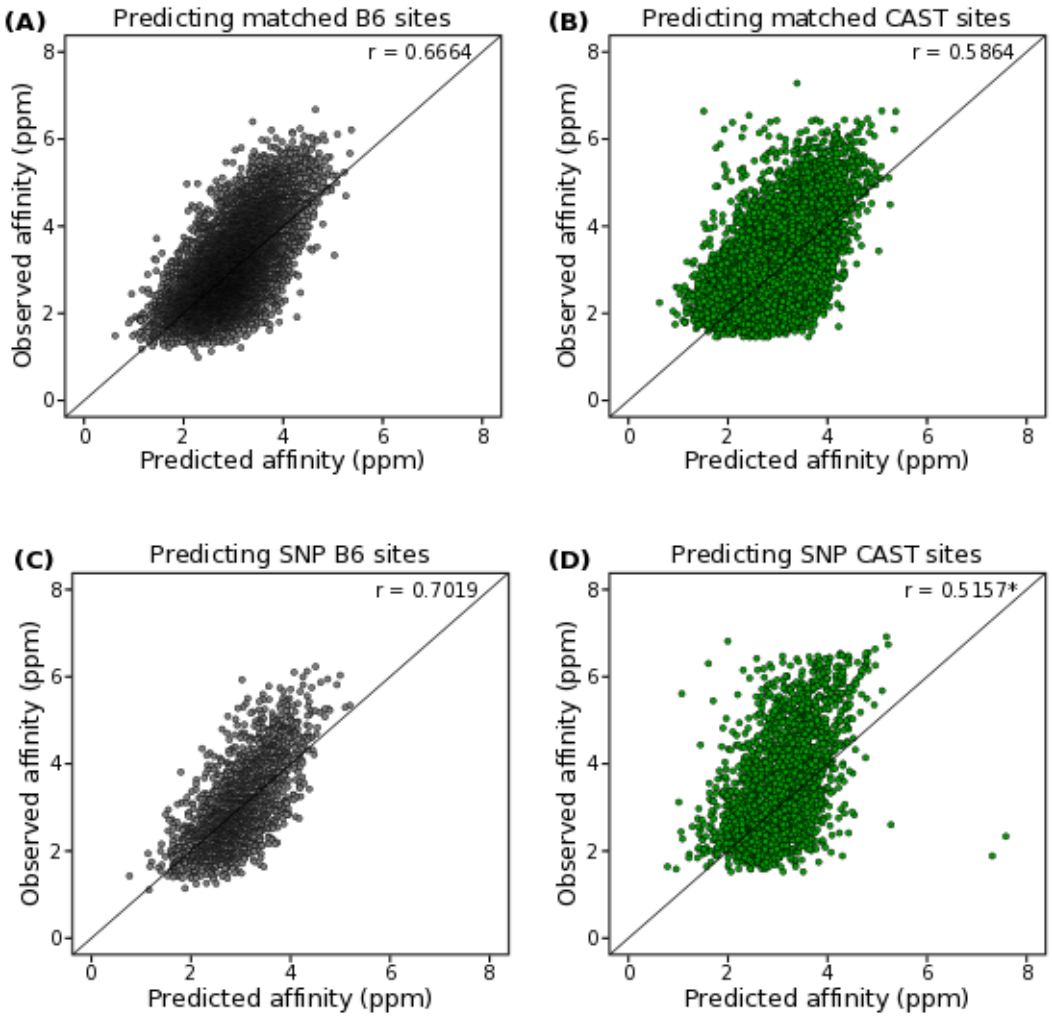

Figure 3.7. Predicted vs observed binding affinity of two-base interaction model of PRDM9^{Dom2}.
The two-way interactive model predicts binding affinity of PRDM9^{Dom2} to a limited degree. (A,B) Predicting the binding affinity of PRDM9^{Dom2} in the B6 (A) and CAST (B) genomes when there are identical genetic sequences in the binding sites for each. (C,D) Predicting the binding affinity of PRDM9^{Dom2} in the B6 (C) and CAST (D) genomes when there are SNPs found between the two genomes. The prediction of affinities for the SNP CAST sites were limited to predictions that fell within the affinity range of 0-8, and the correlation coefficient reflects that.

interactions across the entire binding site, almost all interactions were limited to those that included G8, A30, or T31 (Figure 3.9). G8 is an anchor position, and A30 and T31 fall in the important section of the binding site largely unexplained by standard motif analyses (Figure 3.1A). It's noteworthy that T31 alone has a negative effect on binding affinity, while in combination with many other nucleotides it has a positive effect. These findings suggest that specific nucleotides at key bases interact across the PRDM9^{Dom2} binding site to influence binding affinity.

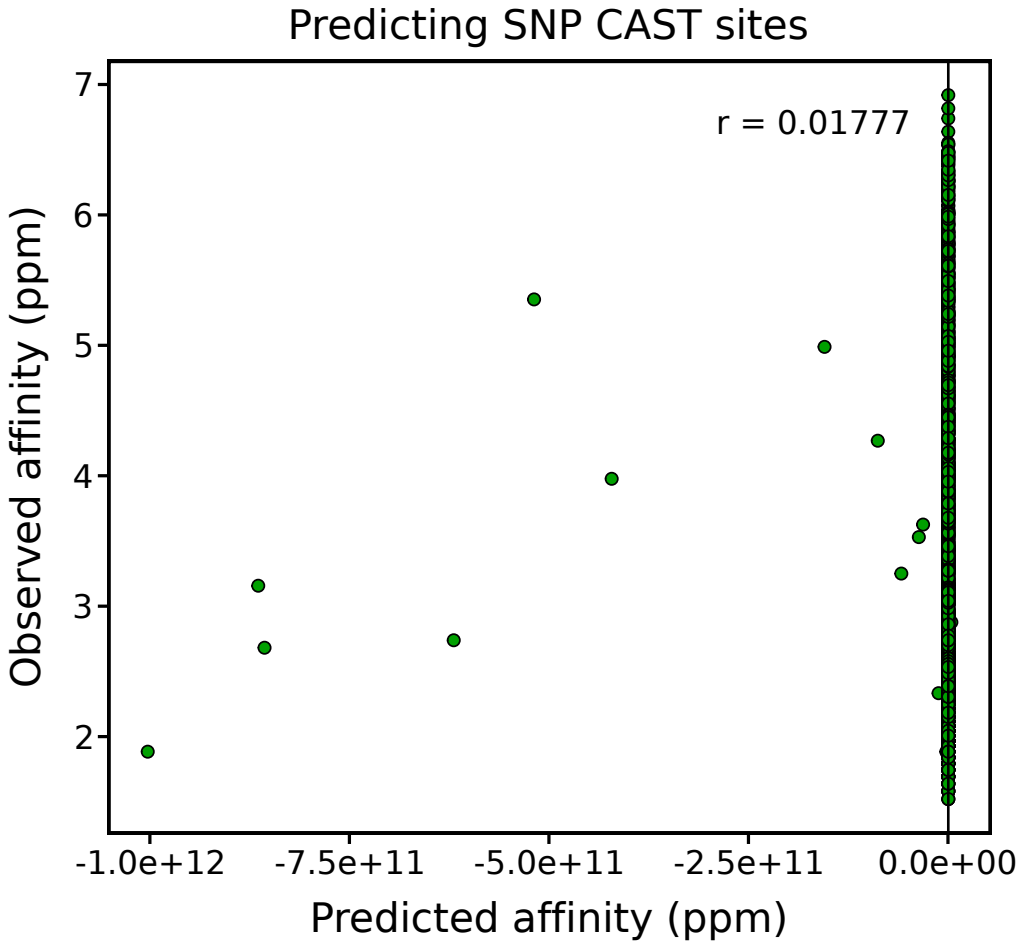

Figure 3.8. Predicted vs observed binding affinity of two-base interaction model of PRDM9^{Dom2} in CAST sites with SNPs.
Predicting the binding affinity of PRDM9^{Dom2} in the CAST genome when there are SNPs found between B6 and CAST at those sites.

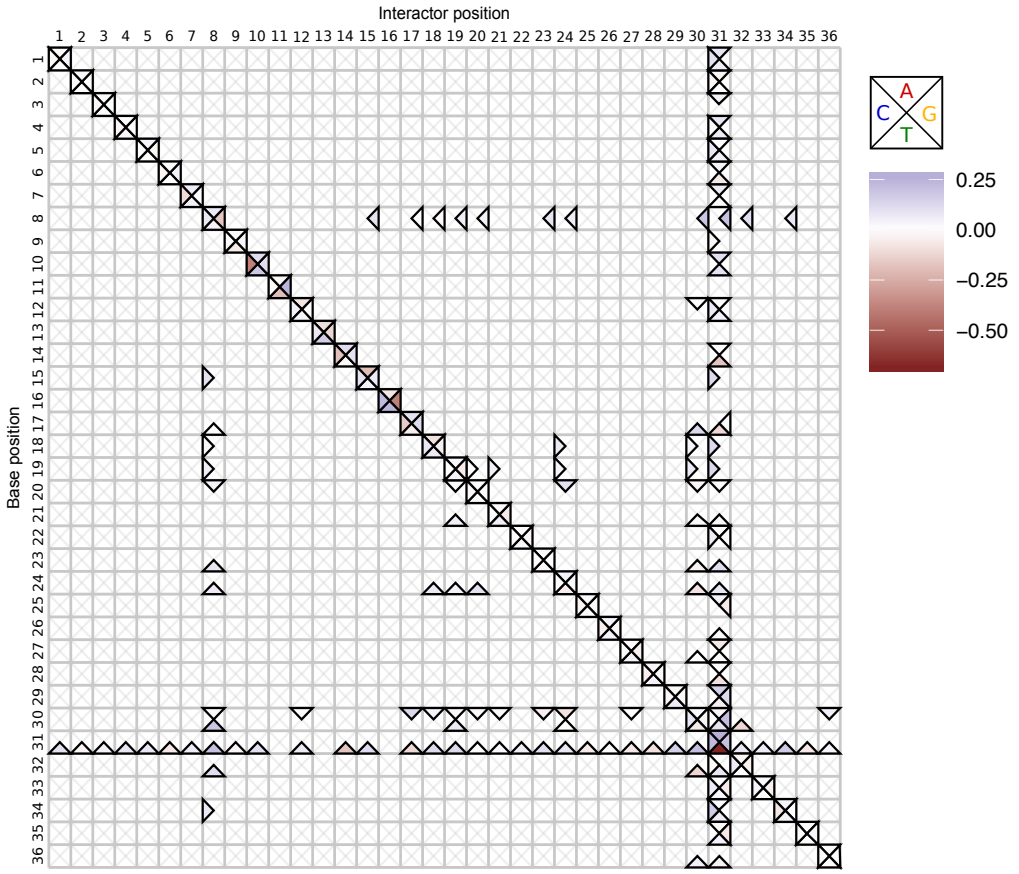

Figure 3.9. Single- and relevent two-base predicted effects on affinity across the PRDM9^{Dom2} binding site.
Single- and two-base interactions identified by Random Forest. Quartile within each square represents the nucleotide present at the base position along the y-axis. The value along the x-axis designates the interacting position. The fill of a quartile defines the predicted effect on binding affinity.

When we interrogated specific interactions, we observed that some had effects on binding affinity that differed from what would be expected (Figure 3.11). For each interaction, we classified each binding sequence into four categories: containing neither interacting nucleotide (base), containing one interacting nucleotide, containing the other interacting nucleotide, or containing both interacting nucleotides. We then summed the average difference from the base sequence in both sequences containing just one nucleotide (Figure 3.11, orange) and compared that to the observed average difference

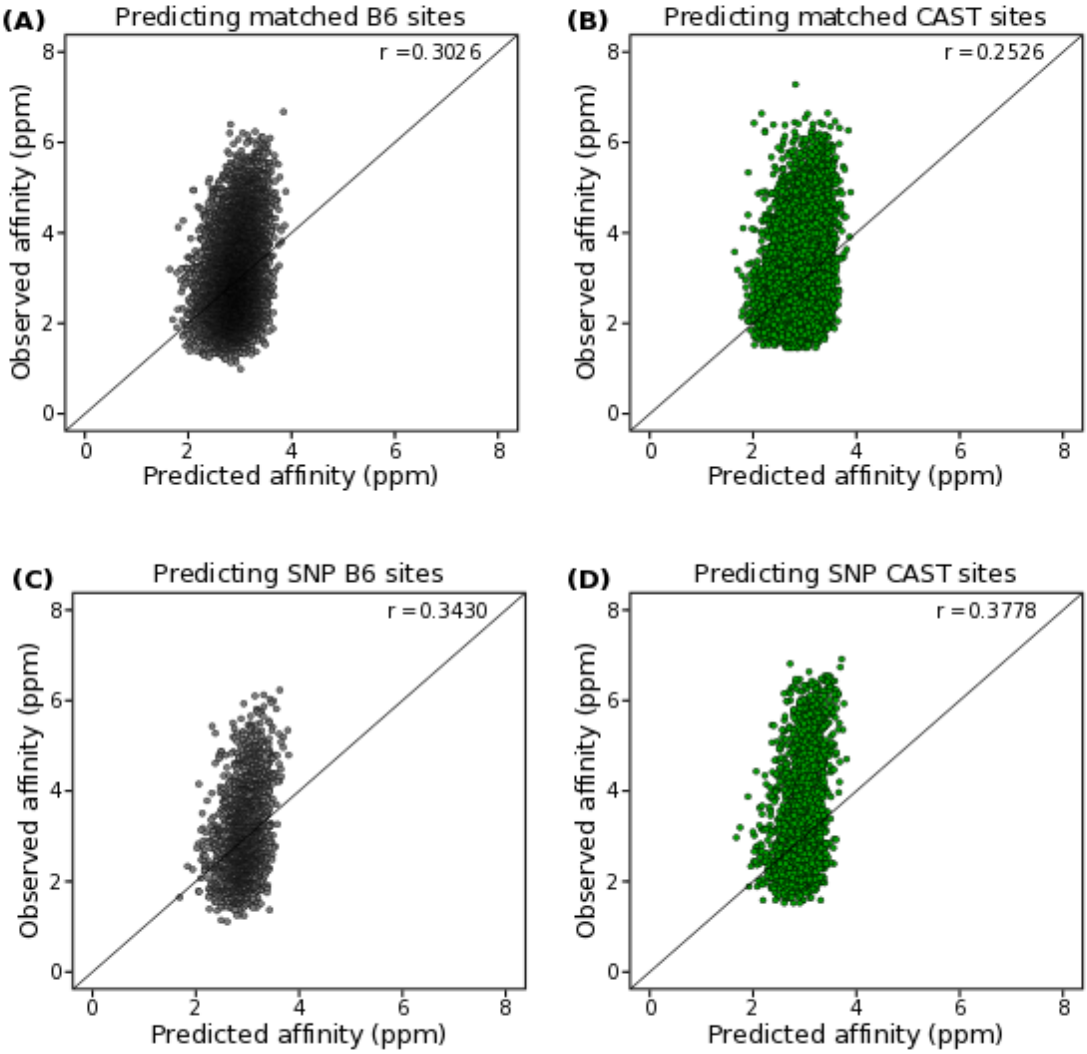

Figure 3.10. Predicted vs observed binding affinity of single- and two-base interaction model of PRDM9^{Dom2} identified by Random Forest.
The singe- and two-way interactive Random Forest model predicts binding affinity of PRDM9^{Dom2} relatively poorly. (A,B) Predicting the binding affinity of th PRDM9^{Dom2} in the B6 (A) and CAST (B) genomes when there are identical genetic sequences in the binding sites for each. (C,D) Predicting the binding affinity of PRDM9^{Dom2} in the B6 (C) and CAST (D) genomes when there are SNPs found between the two genomes.

from the base sequence in the sequences containing both interacting nucleotides. Some interactions with a positive effect score had a greater-than-additive effect on binding affinity (Figure 3.11A, FDR < 0.05). Alternatively, some interactions with a negative effect score had a less-than-additive effect on binding affinity (Figure 3.11C, FDR <

0.05). However, not all interactions had an effect that differed significantly from their additive estimation would suggest (Figure 3.11B). These epistatic effects partially explain the inability of our single base-pair model to predict binding affinity (Figure 3.5, Figure 3.6). Some nucleotides contribute to affinity in a way that is dependent on nucleotides at other positions in the binding site. Overall, while this model is not incredibly successful at predicting PRDM9^{Dom2} binding, it does identify key interactions across the binding site that epistatically affect binding affinity beyond what is accounted for in a single base-pair model.

3.3.5. T31 acts to stabilize PRDM9^{Dom2} binding

Due to the prevalence of T31 across two-base interactions in our model, we were curious about its specific role in binding affinity. Further, it was shown to act in cooperation with two anchor positions, G8 and C15 (Figure 3.9). As the anchor positions are generally required for PRDM9^{Dom2} binding (Figure 3.1), we wanted to see if T31 was able to make up for binding sites that lacked one or more key nucleotide at an anchor position. Indeed, we found that any combination of missing key nucleotides at anchor positions were represented to a greater degree if that sequence also had T31 (Figure 3.12). The majority of binding sites contained all five key nucleotides at the anchor positions, with over 50% of those also containing T31. Binding sites lacking G8, T13, C15, or T16 were present in the thousands, but were more comment when the binding sites also contained T31. G11 appeared to be the most necessary nucleotide, as only 35 sequences lacked this anchor with all other anchors and T31 present. These data do suggest that while the presence of all anchors within a binding site are preferred, T31 can, to a degree, make up for missing anchors.

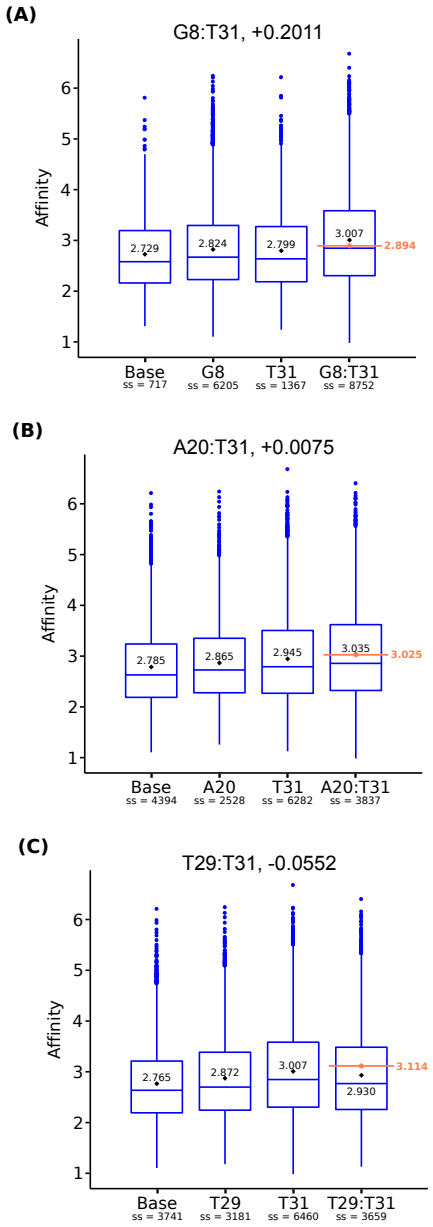

Figure 3.11. Example epistatic effects of Random Forest identified two-base interactions.
Measured affinity for sequences containing various combinations of nucleotides. (A) Binding affinity for sequences containing G8:T31, T31 but not G8, G8 but not T31, and neither. (B) Binding affinity for sequences containing A20:T31, T31 but not A20, A20 but not T31, and neither. (C) Binding affinity for sequences containing T29:T31, T31 but not T29, T29 but not T31, and neither. Black points represent observed means. Orange line and value indicate what a purely additive interaction would lead to for affinity differences.

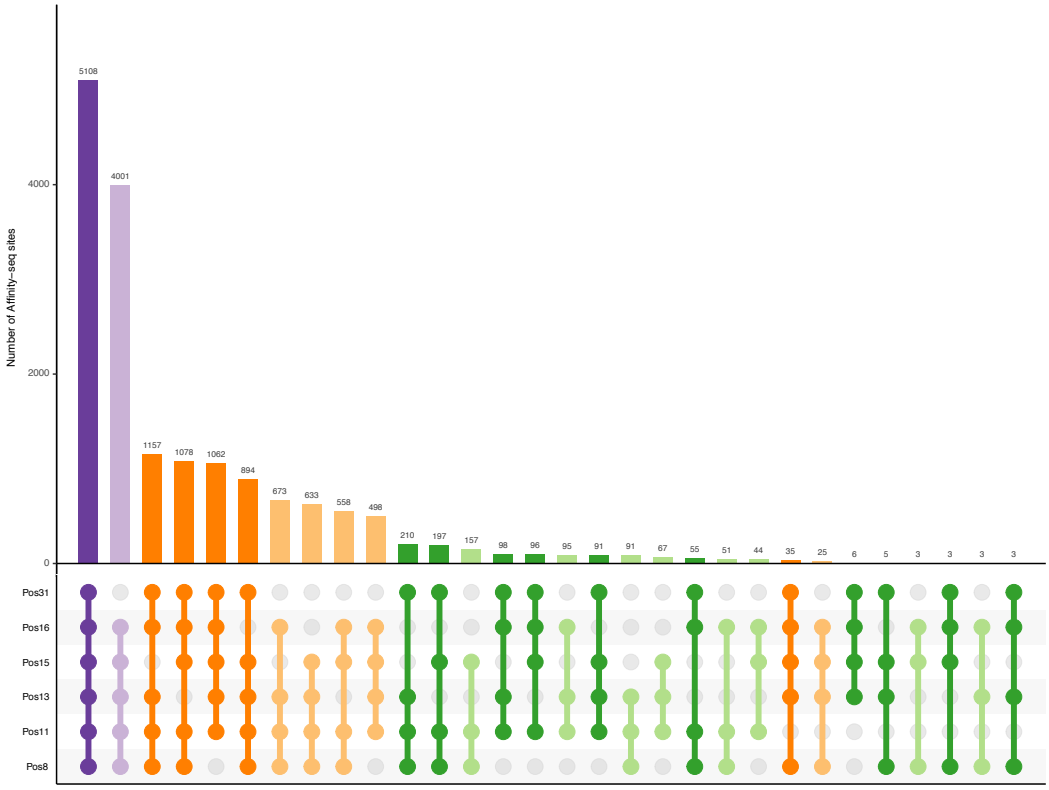

Figure 3.12. T31 supplements PRDM9^{Dom2} binding affinity when anchor sites lack key nucleotides.
Frequency of combinations of key nucleotides at anchor positions and position 31. Nucleotides included were T31, T16, C15, T13, G11, and G8. Bar plots designate the prevalence of nucleotide combinations in our dataset. Darker colors represent anchor position combinations with T31, lighter colors lack T31.

3.3.6. A potential mechanism for differential binding affinity.

We used the sequences of the 200 PRDM9^{Dom2} binding sites with greatest and least binding affinity to predict DNA shape parameters (Zhou et al., 2013). We assessed the minor groove width (MGW), propeller twist, roll, and helix twist estimations for each sequence (Figure 3.13A,B). These measures incorporate sets of three to five base-pairs, making them inherently consider nearby interactions. We found interesting patterns that suggested that the MGW is shorter around the anchor bases in the highest affinity sites than the lowest affinity sites, and that the MGW is more or less consistent in this section among the lowest affinity sites (Figure 3.13C,D). We also observed longer MGW

predictions around position 31 in the highest bound sites compared to the lowest (Figure 3.13C,D). These findings suggest that DNA shape, and specifically MGW, could be a part of the mechanism by which binding site sequence alters long ZF binding.

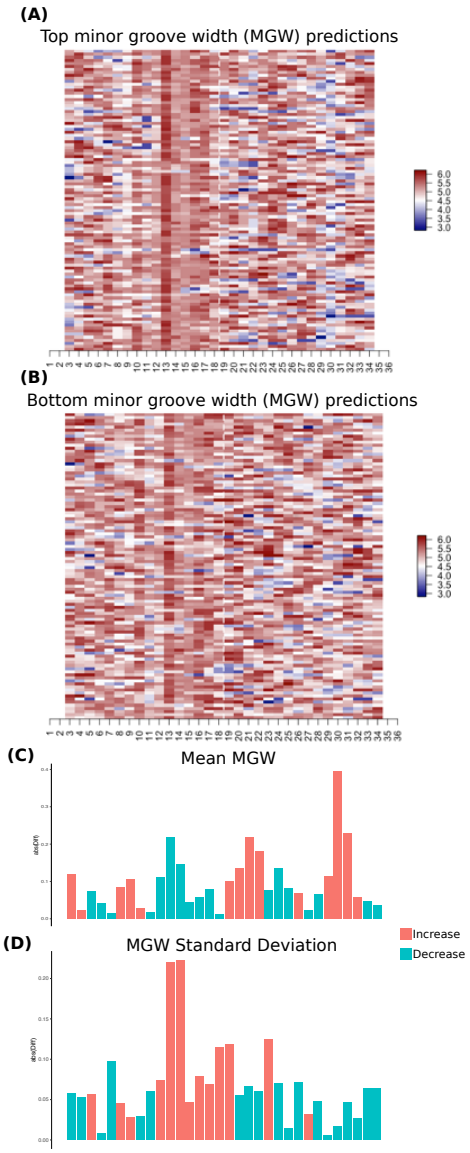

Figure 3.13. DNA shape at binding sites idicate possible mechanism for differential binding affinity.
DNA shape parameters predicted by the sequence of the highest and lowest affinity sites. (A) MGW predictions for the highest 100 sequences. (B) MGW predictions for the highest 100 sequences. Changes in the mean (C) and standard deviation (D) in predictions between the high and low affinity sites.

3.4. Discussion

We have utilized PRDM9^{Dom2} Affinity-seq in both C57BL/6J and CAST/EiJ genomes in order to model how genetic variation alters the binding of this long ZF protein. This dataset assessed long ZF binding in an unprecedented manner – within the context of genetic diversity, without chromatin limitations confounding binding site preference and affinity. We trained a linear model to fit variation in binding affinity to the SNPs within the binding site of PRDM9 in the B6 genome, and then tested its predictive power in the CAST genome. We then used Random Forest to identify significant interactions between nucleotides across the binding site and incorporated them into our model. These presented us with two major findings. First, single base-pair, additive models of long ZF binding are insufficient to predict how binding site genetic variation will alter affinity. Second, important nucleotide interactions almost always included specific nucleotides at a few positions within the PRDM9^{Dom2} binding site. These results inform us of how PRDM9 selects its binding sites, as well as how to model ZF binding at large.

An additive, single base-pair linear model of ZF binding was unable to model binding affinity in our system. This model predicted only a fraction of the range of affinities that we observed experimentally and the predicted affinities were poorly correlated with observed affinities. This prediction by this model of the effect of a single SNP within an otherwise constant sequence was better correlated with observed affinity; however, the model still failed to predict the range of effects observed. An issue with this model became apparent in this analysis – the same SNP at a given position was predicted to have a singular effect on binding affinity, but instead their surrounding sequence

modulated their effect. This likely limited our accuracy in predicting affinity for an entire sequence as well. This result was unsurprising, as at a minimum, the triplets of nucleotides that are bound by a ZF might interact to affect binding affinity, but the extent of these interactions, as well as the potential for cross-sequence interactions, were unknown. These findings clearly demonstrate the need for multi-base-pair models when predicting the binding affinity of long ZF arrays.

By adding two-base interactions to our model, we could predict binding affinity to a greater degree. The most important of these two=based interactions were largely limited interactions involving one of three nucleotides in specific positions: G8, A30, and T31. It is possible that some of the other anchors within the binding site would also be heavily represented in significant interactions, but the limited nucleotide variation at anchor positions likely reduced our ability to identify such interactions. It is intriguing to consider a model of ZF binding that relies on a few key bases, which all interact across the binding sequence. Interestingly, for G8 and T31, it appears that their relevance to binding affinity necessitated specific other nucleotides across the sequence. The effects of G8 and T31 alone were negative, while many of their interactions would allow for their presence in a sequence to have a positive effect on binding. Part of the relevance of positions like the anchors, 30, and 31 may be due to their effect on DNA shape. When comparing the sites with the strongest and weakest affinities, these two regions differed drastically, either in magnitude or variability. These changes in DNA shape could be the mechanism by which these positions differentially interact across the rest of the binding site to affect binding affinity.

These results can be used to inform how we study PRDM9 in its biological context. An accurate model of PRDM9 binding affinity can have utility for studying both genetics and reproduction. As the activator of hotspots for meiotic recombination, PRDM9 determines where genetic crossing over could occur between homologous chromosomes. A complete model would mean the ability to predict the possible recombination sites for a novel PRDM9 allele. Beyond better understanding of the genetics of meiosis, this would have significance within the field of reproductive biology. Homologous recombination is a crucial part of gametogenesis, and better understanding its mechanism and regulation will improve our capacity to study how it works and when it goes wrong. First, the selection of PRDM9 binding sites *in vivo* is limited to a subset of those that appear in the Affinity-seq analysis. This is likely due to cellular complexities that were specifically avoided in the Affinity-seq pipeline, such as epigenetic modifications and chromatin state. Studying why certain sites are selected *in vivo* will require a complete set of possible sites, e.g., those from Affintiy-seq. However, as this would be costly to perform on every genetic background for each PRDM9 allele, a computational model of PRDM9 binding would better support these research endeavors. Second, many meiotic mutants lead to errors in this process, from hotspot activation, to double-strand break formation, to recombination itself. Being able to computationally predict the location of proper PRDM9 binding sites could inform us of what stage of the process is going wrong. For both of these examples, PRDM9 binding sites have been well established in the C57BL/6J genome, but it is not known how SNPs in other genomes with $PRDM9^{Dom2}$ might perturb binding, nor the binding sequence for other alleles of PRDM9. Establishing a computational model of PRDM9 binding, for the *Dom2* allele

specifically, as well as a generalizable form for alternative alleles, would empower genetic and reproductive research, for example, perhaps clarifying mechanisms of hybrid sterility.

One limitation to this analysis lies in the range of the observed affinities, which could be partially remedied by greater genome-wide interrogation. Ultimately, all measured PRDM9^{Dom2} binding sites in this study were well-bound by PRDM9, as we deliberately set strict thresholds when calling peaks. This was done to limit false positives; however, PRDM9^{Dom2} likely does bind to other sites in the genome with sub-threshold affinity. Further, due to technical limitations of precipitations and DNA sequencing, false negatives are possible within our dataset. Therefore, it is possible that there are other sites in the genome to which PRDM9^{Dom2} binds. It could be informative to survey the rest of the B6 or CAST genomes for sites that were not called as Affinity-seq binding sites but contain a motif resembling that for PRDM9^{Dom2}. This could be compared to sub-threshold peaks in the Affinity-seq dataset, and those lacking Affinity-seq reads could be validated to confirm their true-negative status. True-negative sites in the genome could be utilized to confirm the necessity of specific nucleotides across the PRDM9^{Dom2} binding site, such as at anchor positions. Poor-affinity PRDM9^{Dom2} binding sites would further inform us of how long ZFs select their binding sites.

A generalizable model of long ZF binding would have benefits to other fields of biology as well, beyond PRDM9. ZF proteins play crucial roles in development, as well as the maintenance of healthy differentiated cells. This work demonstrates additional considerations that should be taken when modeling ZF binding. Singe-base models are inadequate for predicting binding sites of a long ZF array, as well as for modeling the

effect of genetic variants in a binding site on affinity. Multi-base models perform better, but should be carefully trained to not over-fit to their dataset. Our multi-base model incorporates interactions between nucleotides across the binding site of $PRDM9^{Dom2}$. Importantly, these interactions were not limited to nearby nucleotides that would be bound by the same ZF residue, so these types of long-range interactions should be considered in future models of ZF binding. Further, one of the most frequent interacting nucleotides in our model, T31, were not particularly overrepresented in our dataset, despite its relevance to $PRDM9^{Dom2}$ binding identified in our multi-base model. This further demonstrates the need for more complex models than those based on frequency across surveyed binding sites. Taken together, these data outline how to utilize Affinity-seq data to model long ZF binding and detail the types of multi-base interactions that should be considered when trying to predict how binding site sequence will affect differential ZF binding.

3.5. Contributions

This project was conceived and planned by myself with Drs. Arat, Carter, Paigen, and Petkov. Members of the Paigen and Petkov groups collected and prepared the data. Seda Arat and I shared the computational analyses. I generated figures and wrote the manuscript.

CHAPTER 4: Modeling the effect of genetic and epigenetic variation on gene expression in mouse hepatocytes

Fine, A. D.*, Tyler, A. L.*, Spruce, C.*, Ball, R. L., Pitman, W., Haber, A., Kursawe, R., Walker, M., Stitzel, M., Paigen, K., Petov, P. M., & Carter, G. W.

To be submitted.

4.1. Introduction

Gene expression is a facet of molecular biology that is intrinsically related to cellular function and organism health. Healthy cell identity is largely controlled by differential gene regulation; the expression of key genes plays a major role in differentiation, development, and cell maintenance. For example, expression of OCT4 can induce pluripotency in embryonic stem cells, while the expression of DNMT3A plays an important role in hematopoietic stem cell differentiation (Challen et al., 2011; Kim et al., 2009; Pan et al., 2002). Many cellular pathologies are caused by improper gene regulation. Too much or too little of a gene product can have drastic effects on cellular health, from PKD1 over-expression leading to kidney disease to BRM under-expression being a risk factor for lung cancer (Liu et al., 2011; Thivierge et al., 2006). Due to its importance, there are multilevel systems present in a cell to regulate gene expression. Understanding these mechanisms would allow for the prediction how a system perturbation would alter gene expression and, in turn, cell and organism health.

Many genetic diseases seem to be caused by improper gene regulation, as the majority of disease- and trait-associated single nucleotide polymorphisms (SNPs) fall in noncoding, regulatory regions of the genome (Cookson et al., 2009; Karczewski et al., 2013; Maurano et al., 2012; Nicolae et al., 2010). Over 75% of these SNPs fall in Deoxyribonuclease I (DNase I) hypersensitivity sites (DHSs), which are likely accessible due to their being regulatory elements of the genome (Degner et al., 2012; Maurano et al., 2012). These include promoters, enhancers, and repressors, which contribute to the regulation of gene expression largely through local epigenetic state and recruitment of proteins like transcription factors (Consortium, 2012; Degner et al., 2012; Shlyueva et al.,

2014). To interpret these biologically relevant SNPs, we must understand how the genetic sequence of regulatory elements determine their function.

There is a clear genetic basis for the determination of regulatory elements and their role in gene expression regulation. There is a measurable degree of genetic conservation in regulatory elements among species, suggesting that their genetic sequence is crucial to their function (Bejerano et al., 2004; Ngo et al., 2019; Ovcharenko et al., 2004; Plessy et al., 2005). Different regulatory elements, defined by genomic annotations or histone modifications, have common, underlying genetic motifs (Korkuc et al., 2014; Ngo et al., 2019; Whitaker et al., 2015). These include genetic elements identified in known regulatory regions, such as TATA boxes, which are found in many promoters (Dikstein, 2011; Lifton et al., 1978). There have also been genetic motifs computationally identified to be enriched in regions sharing specific histone modifications, which likely have shared functions in gene regulation (Ngo et al., 2019). For example, multiple motifs have been identified in the center of H3K27ac peaks, which when disrupted, lead to corresponding changes in histone modifications (Kasowski et al., 2013; Whitaker et al., 2015). Regulatory DNA, including distal elements such as enhancer DNA, are sufficient to regulate expression outside of their biological context (Banerji et al., 1981; Catarino and Stark, 2018; Shlyueva et al., 2014). Current methods for assessing enhancer function now include placing candidate enhancer sequences near genes and determining their effect on gene expression (Arnold et al., 2013). SNP disruptions to some regulatory elements have been associated with disease-related gene expression differences (Ngo et al., 2019; Pomerantz et al., 2009; Sagai et al., 2005),

indicating that these regulatory sequences are not just sufficient, but are necessary for their element function.

A major mechanism by which the genetic sequence in a regulatory element affects gene expression is through its chromatin state, being the epigenetic changes to a region of DNA. Largely through histone modifications, the chromatin around regulatory elements can differentially promote gene expression (Consortium, 2012; Shlyueva et al., 2014). These modifications are correlated with open chromatin to a varying degree, making them more or less accessible to activate gene expression, and also play a role in the recruitment of key proteins like transcription factors (Grewal and Jia, 2007; Huisinga et al., 2006). Substantial work has been done to define what regulatory elements are defined by varying combinations of histone modifications (Consortium, 2012; Karlic et al., 2010; Kellis et al., 2014; Zhang et al., 2015). For example, mono-methylation of lysine 4 on histone 3 (H3K4me1) and acetylation of lysine 27 of histone 3 (H3K27ac) are the canonical modifications that define active enhancers, while H3K4me1 and trimethylation of lysine 27 of histone 3 (H3K27me3) are the canonical modifications defining closed enhancers (Heintzman et al., 2007; Shlyueva et al., 2014). These histone modifications are often enriched for distinct genetic motifs, as stated earlier (Ngo et al., 2019). Taken together, these data provide a framework for how genetic sequences in regulatory elements can affect gene expression.

Despite all that is known about regulatory element function, their complex mechanisms for regulating gene expression have limited our ability to predict how they function, when they will dysfunction, and consequences of dysfuction. While there are genetic commonalities and motifs found among some regulatory elements, there are no

defining genetic sequences for any element type (Catarino and Stark, 2018; Ngo et al., 2019). For example, despite their relative prevalence, most human promoters lack a TATA-box (Smith et al., 2006). The genetic sequences that commonly underlie histone modifications are also variable (Whitaker et al., 2015). Further, while histone modifications are an accessible way to measure the activity of a regulatory element, enhancers can function without their canonical histone modifications (Catarino and Stark, 2018; Pollex and Furlong, 2017). These complications have made it difficult to predict when and how a SNP will alter the function of a regulatory element, and therefore gene expression. If these issues could be reconciled, we could better predict how disease-associated SNPs affect gene expression, making them targetable for health intervention.

Mice provide a unique resource as a model organism for studying genetic effects, such as SNP effects on regulatory element activity. Inbred strains of mice are homozygous at nearly every locus, allowing for reproducibility from mouse-to-mouse within a single strain. On the flip side, the diversity among multiple strains of mice can model the estimated degree of diversity within the human population, making inter-strain comparisons helpful analogs for human genetic variation. To study regulatory element activity, we utilized a panel of nine inbred strains of mice: A/J (AJ), C57BL/6J (B6), 129S1/SvlmJ (129), NOD/ShiLtJ (NOD), NZO/H1LtJ (NZO), CAST/EiJ (CAST), PWK/PhJ (PWK), WSB/EiJ (WSB), and DBA/2J (DBA). The first eight of these strains (all but DBA) have been used to generate a diverse, outbred population of mice (DO). DBA and B6 have also been utilized by the research community to generate a recombinant inbred line (BXD). A greater understanding of how genetic variation between these nine inbred strains alters the function of regulatory elements could not

only improve our interpretation capacity of disease- and trait-associated variants, but would also enable greater inferences to be drawn from DO and BXD mice as well.

To examine the relationships between genetics, epigenetics, and gene expression, we surveyed histone modifications and transcript abundance from hepatocytes of these mouse strains. Hepatocytes are a relatively homogenous and abundant differentiated cell type in the mouse, making them a simple model for interrogation of gene regulation. From this multi-omic dataset, we were able to build computational model to understand the relationships between noncoding variation, epigenetic regulation, and transcript abundance.

4.2. Methods

4.2.1. Experimental design

4.2.1.1. Sample acquisition

Samples were taken from 12-week female mice of nine inbred mouse strains: A/J, C57BL/6J, 129S1/SvlmJ, NOD/ShiLtJ, NZO/H1LtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ, and DBA/2J. These five strains include six standard laboratory strains and three wild-derived inbred strains. Eight of these strains comprise the vast majority of genetic variability in the eight Collaborative Cross/Diversity Outbred founders. The ninth strain is DBA/2J, which will facilitate the interpretation of existing and forthcoming genetic mapping data obtained from the BxD recombinant inbred strain panel. Mice were aged and processed in groups to maintain a steady sample preparation workflow. Mice were housed, born, and aged in the same mouse room, with uniformity in timing, diet, and all other possible conditions. Female mice were used for all experiments due to potentially

confounding effects from variation in testosterone among males that can affect liver gene expression, as well general experience that female expression is less variable than male in multiple tissues. This will also facilitate the analysis of maternal effects on offspring in later studies.

4.2.1.2. Liver perfusion and hepatocyte collection

Hepatocytes were collected from perfused livers according to published methods (Baker et al., 2019; Neufeld, 1997).

4.2.2. Genomic measurements

4.2.2.1. Histone Chromatin Immunoprecipitation (ChIP)

We performed H3K27ac ChIP-seq experiments as described in (Parker et al., 2013; Stitzel et al., 2010). Crosslinked chromatin was sheared on ice using a Branson 450 Sonifier (constant duty cycle, output 4, 12 cycles of 20 s sonication with 1 min rest between cycles) to a size of 200–1000 bp. Immunoprecipitation was performed using rabbit anti-H3K27ac antibody from Abcam (ab4729, Cambridge/UK). H3K27me3, H3K4me1, and H3K4me3 were collected as previously described (Baker et al., 2019).

4.2.2.2. RNA sequencing (RNA-seq) and ChIP sequencing (ChIP-seq)

The samples for ChIP-seq and RNA-seq were submitted to The Jackson Lab GES service for library preparation and sequencing. ChIP-seq samples were sequenced with 40 or more million reads per sample. RNA-seq samples were sequenced with approximately 30 million reads per sample. 100-base single-end reads were sequenced on the Illumina HiSeq 2500 and filtered for quality.

4.2.3. Computational methods

4.2.3.1. RNA-seq processing

Reads from each sample were mapped with Bowtie (Langmead et al., 2009) to strain-specific pseudogenomes that integrate known SNPs from each strain, created with EMASE (Raghupathy et al., 2018). Once the sequencing data is mapped to the custom genomes, EMASE was used to quantify transcripts. Transcripts with less than 1 CPM in two or more replicates were filtered to remove lowly expressed genes.

4.2.3.2. Principal component analysis

To visualize the gene expression variation by sample, a Principal Component Analysis (PCA) was run by performing Singular Value Decomposition (SVD) on RNA-seq data. Each transcript's expression was centered and scaled, and PCA was performed using svd(x) from the R statistical framework (R Core Team, 2015).

4.2.3.3. ChIP-seq processing

The raw sequencing data from ChIP-Seq was put through the quality control program FastQC. FastQC identifies problems or biases in either the sequencer run or the starting library material. The FastQC readout includes total number of reads, sequence quality, duplication level, and overrepresented sequences. All of our samples had comparable quality levels and no outstanding flags. Total number of reads was 40 million or more, with an average read length of about 100 bp. Quality scores were mostly above 28 (including their error bars), with the average above 38. Duplication level was reduced to <2 for about 95% of the sequences. Adapter content was trimmed using Trimmomatic 0.33.

For the sequence analysis, reads from each sample were mapped to strain-specific pseudogenomes, while the B6 samples were aligned directly to the reference mouse genome (GRCm38). Strain-specific sequence variation in transcripts can affect alignment quality and result in biased estimates of abundance. To counteract potential strain biases, sequencing data from each strain was aligned to a custom strain pseudogenome, allowing a more precise characterization of gene expression and histone binding. The pseudogenomes were created by incorporating strain-specific SNPs and indels into the reference genome, using g2gtools (https://github.com/churchill-lab/g2gtools). The resulting custom genomes are called pseudogenomes because they do not attempt to completely rebuild the entire genomic sequence from the scaffold up for each strain. Reads from each sample were aligned and mapped to the pseudogenomes using the Bowtie mapping algorithm, and then translated to the B6 (reference) coordinates using g2gtools. Unlike other liftover tools, g2gtools does not throw away alignments that land on indel regions.

4.2.3.4. ChIP peak calling

For H3K4me1, H3K4me3, and H3K27ac, we called narrow peaks with MACS2 (Zhang et al., 2008). We set a p-value cutoff of 1e-5 for each of these marks. For H3K27me3, we called broad peaks with EPIC (Xu et al., 2014). We used the default FDR cutoff (0.05) and increased the gaps allowed to 5.

4.2.3.5. Peakome determination

To define a common peakome across all strains, we ran a DiffBind Analysis on our samples for each histone modification separately (Ross-Innes et al., 2012; Stark and Brown, 2011). First, this analysis identifies peaks in each condition, which was strain in

our dataset. Then, DiffBind identifies peaksets in each strain, which we set to peaks that were in at least two samples in a condition, and otherwise standard inputs. Next, peaks in these peaksets were re-quantified in each strain, regardless of whether or not the peak had been called in that strain initially. This analysis returns a list of peaks for the entire dataset, as well as counts for those peaks in each sample. These counts were used for peakset analyses (Figure 4.3, Figure 4.4).

4.2.3.6. Chromatin state determination

ChromHMM was performed to identify chromatin states across the genome using standard inputs, including separating the genome into 200 bp chunks (Ernst and Kellis, 2012, 2017). We then used ChromHMM to designate those chunks into distinct states. We called a range of number of states, 4-8, and then decided on six states, as this was the state count that was the highest without having redundancy between states.

4.2.3.7. Chromatin state clustering

Chromatin states across strains were clustered using daisy in R from the cluster package (Maechler et al., 2019). We calculated similarity based on gower distances, as chromatin states were categorical, rather than numerical.

4.2.3.8. Multidimensional scaling

We used multidimensional scaling to calculate strain-specific values for each gene's state matrix. We first calculated the hamming distance between strains in the state matrix. We then used the function cmdscale in the stats package in R (R Core Team, 2015) on the distance matrix with k set to 1. This generated a single vector with one value per strain. We correlated these values to mean inbred expression.

4.2.4. Bioinformatic methods

4.2.4.1. Promoter identification

Promoter peaks were identified as being within 1 kb of TSSs, according to their position in mm10, as annotated in BioMart (Smedley et al., 2015). We used H3K4me3 peaks, as defined by DiffBind Analysis for strain-to-strain and sample-to-sample consistency in start and end positions.

4.2.4.2. Gene Ontology Term Enrichment Analysis

GO term enrichment analysis was applied to subsets of our gene list using Gene Ontology enRIchment anaLysis and visuaLizAtion (GOrilla) (Eden et al., 2009; Harris et al., 2004). We ranked genes by their correlation coefficient specific to the list they were a part of and compared them against all genes in the database (no background dataset).

4.3. Results

4.3.1. Gene expression corresponds to genetic background

We collected hepatocytes from 12-week-old mice from nine inbred strains: AJ, B6, 129, NOD, NZO, CAST, PWK, WSB, and DBA. These strains comprise the founders of the Diversity Outbred (DO) population of mice, with the addition of DBA. While all of these strains are genetically divergent, it has been well established that two wild-derived strains (CAST and PWK) are far more dissimilar from the rest of the strains, with a third wild-derived strain (WSB) being next.

We performed RNA-seq on each hepatocyte sample to quantify gene expression. To assess the variability across our dataset, we performed a principal component analysis (PCA) to see how the samples segregated in two dimensions. Looking at the first and

second components (PC 1 and PC 2), it was apparent that within-strain variation was low (Figure 4.1). Further, the two strains with the most distinct gene expression signatures were PWK and CAST, as PC 1 separates samples of these two strains from all of the other samples and PC 2 separates PWK samples from CAST samples. There is a slight distinction of WSB samples from the other strains along PC 1 as well. These differences in global gene expression generally correspond to the known genetic differences between these strains, demonstrating the relationship between genetics and gene expression.

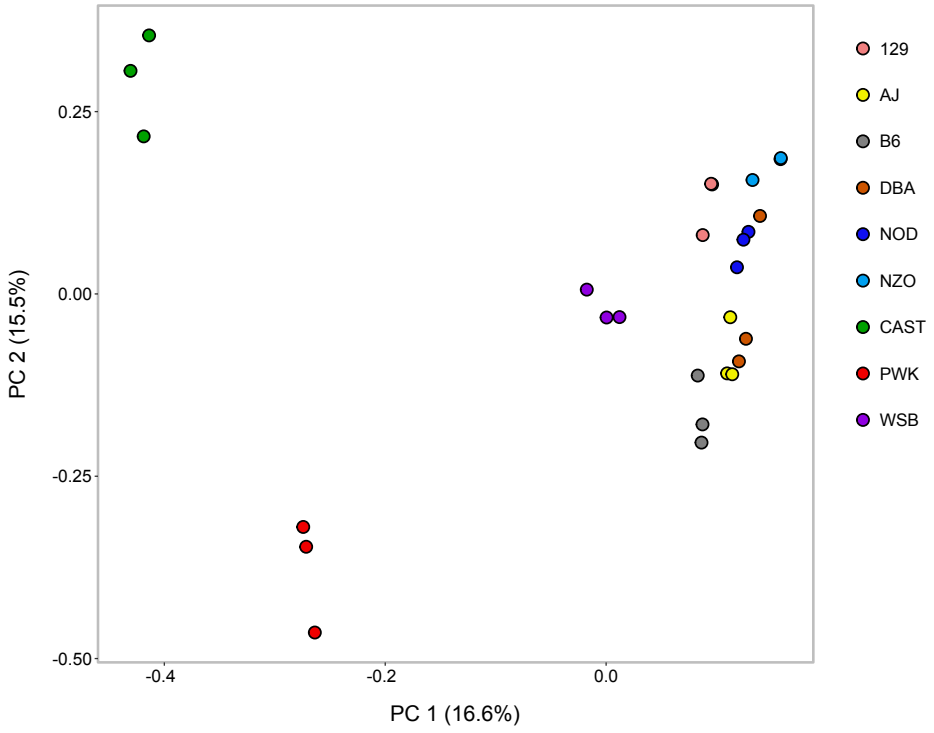

Figure 4.1. Hepatocytes have distinct gene expression signatures according to their genetic background
Principal components 1 and 2 from a principal component analysis (PCA) for RNA-seq data on hepatocyte samples. Points are colored by strain, in three replicates.

4.3.2. Local genetic variation correlates with gene expression

While it was apparent that gene expression differences between strains reflected genome-wide diversity, we were interested to identify how local variants alter gene

regulation. Simply using the inbred strains, it is impossible to distill local variation from distal variation. Therefore, we utilized public data from the DO population of mice, which eight of our nine inbred strains were used to generate. RNA-seq had been performed on livers collected from DO mice (Chick et al., 2016). These data were comparable to the gene expression data from inbred hepatocytes; however, the shuffled genome of DO mice allowed for the identification of local, allele-specific effects on gene expression via expression-based quantitative trait loci (eQTL) mapping. By leveraging this information against our inbred gene expression, we could assess the degree to which local genetic variation contributes to gene expression in our hepatocytes from inbred strains.

We used previously mapped eQTL data from livers of DO mice to model how local genetics contributes to gene expression in inbred mice (Figure 4.2). We selected the 6,732 genes that had a LOD score greater than 10, to limit our analyses to genes with statistically significant genetic contribution. To quantify allelic contribution, we took the local DO eQTL coefficient for each gene in this list (Figure 4.2A). This could be compared to the measured gene expression in each inbred strain of mouse (Figure 4.2B). For each gene, we calculated the degree of correlation between the DO eQTL coefficient and the measured, inbred gene expression (Figure 4.2C). We summarized the correlation coefficient for all genes with a LOD score greater than 10 to see how well local genetics corresponds to gene expression (Figure 4.2D). Approximately 48% of these genes had a correlation coefficient greater than 0.5. These results demonstrate that there is a generally positive relationship between local genetic variation and gene expression.

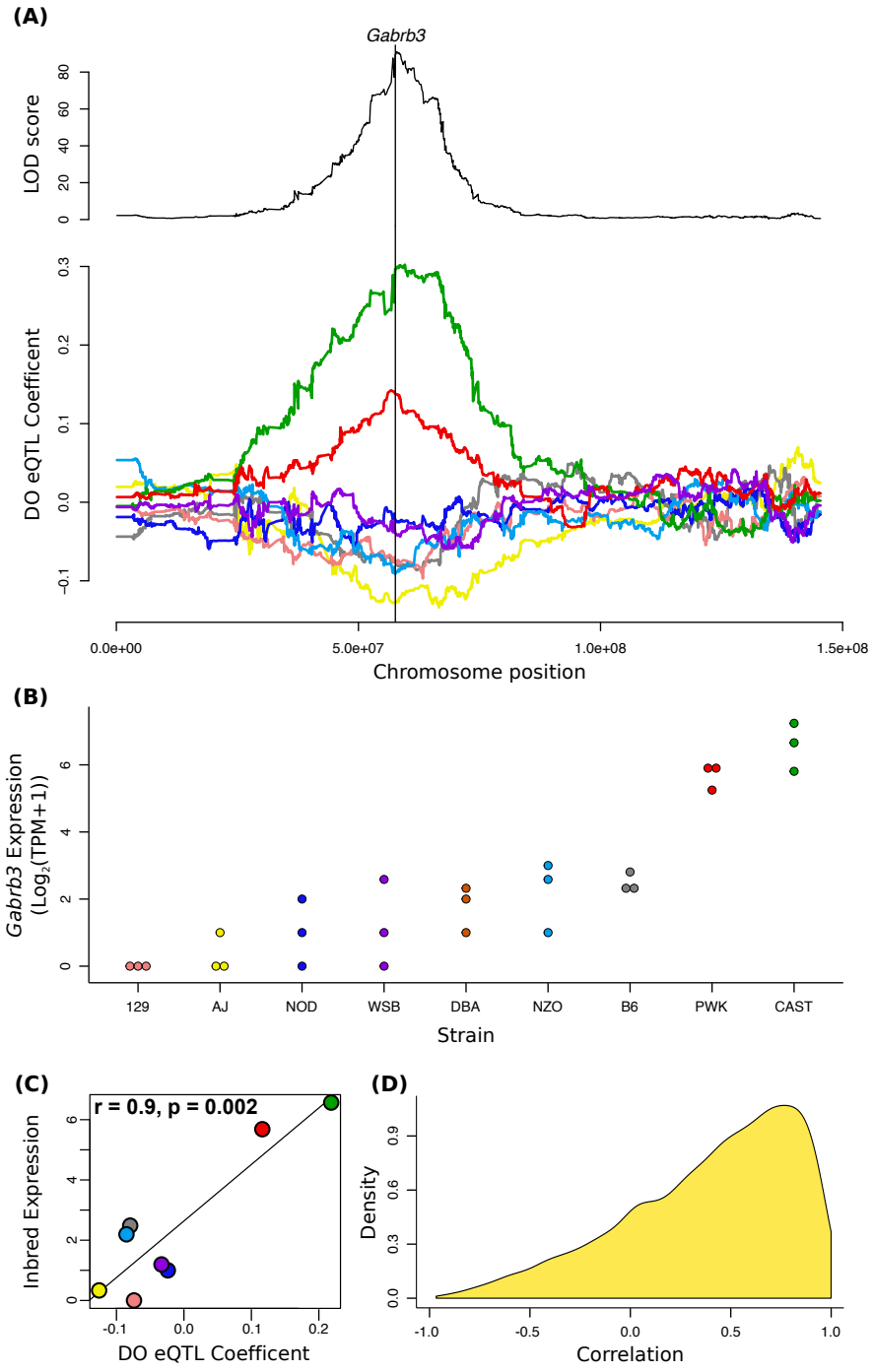

Figure 4.2. DO eQTL coefficients correspond to transcript abundance in inbred mice.
Representative analysis of local genetic variation correlating with gene expression in inbred mice. (A) DO eQTL coefficients for *Gabrb3* in the liver. (B) Expression of *Gabrb3* in inbred hepatocytes. (C) Correlation between DO eQTL coefficients and measured inbred expression for *Gabrb3*. (D) All correlation coefficients from genes with a LOD > 10.

4.3.3. Histone modifications correspond to genetic background

We performed ChIP-seq for multiple histone modifications in the same hepatocyte samples as our RNA-seq in order to assess how epigenetic state relates to gene expression. We collected data for four histone modifications: H3K4me3, H3K4me1, H3K27me3, and H3K27ac. These marks can be used to distinguish key regulatory elements in the genome. H3K4me3, H3K4me1, and H3K27ac are generally marks of active chromatin, while H3K27me3 is generally repressive. These are the key, known histone modifications that define the activity regulatory elements, thereby allowing us to model how epigenetic state in around genes alters their expression levels.

First, we wanted to assess the strain-to-strain variability for each histone modification. We called narrow peaks for each H3K4me3, H3K4me1, and H3K27ac sample ($p < 0.00001$) and broad peaks for each H3K27me3 sample ($p < 0.01$). Then, we performed DiffBind Analysis to generate peakomes for each mark – a defined list of genomic coordinates with the read counts quantified across each sample (Methods). These were then compared for similarity across each sample, by each mark (Figure 4.3). In general, samples clustered by strain, with only a few exceptions. Similar to gene expression, CAST and PWK samples were the most dissimilar from the rest of the strains, corresponding to their more distant genetic relatedness. With the exception of a single DBA H3K4me1 sample, WSB samples were the third most dissimilar, as seen in gene expression. Taken together, there is a noteworthy correspondence between genetic background and epigenetic state genome wide, and a likely relationship between epigenetic state and gene expression.

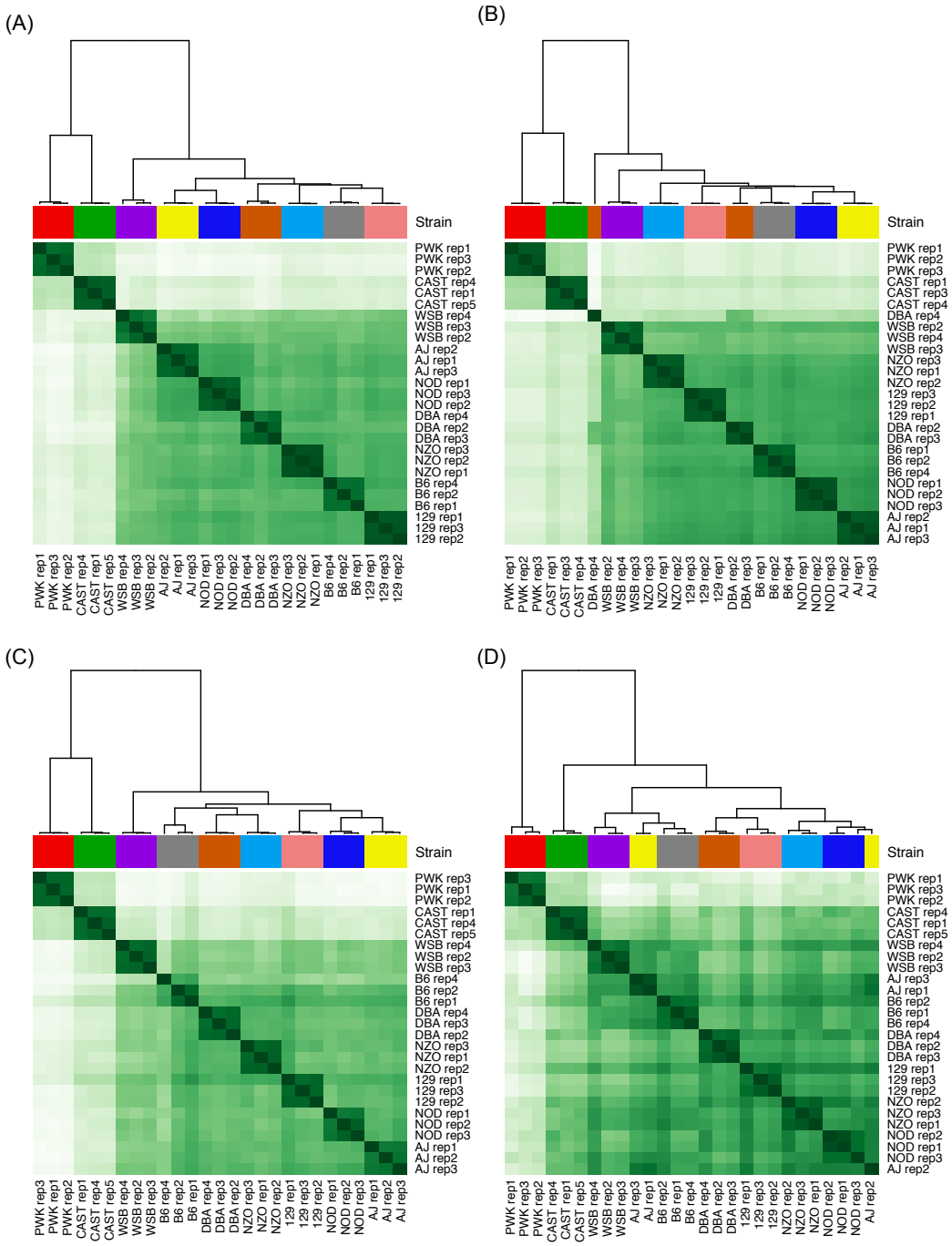

Figure 4.3. Histone modifications cluster by genetic background.
Heatmaps and clustering of peakomes by sample for each histone. Generally, H3K4me3 (A), H3K4me1 (B), H3K27ac (C), and H3K27me3 (D) all cluster by strain background.

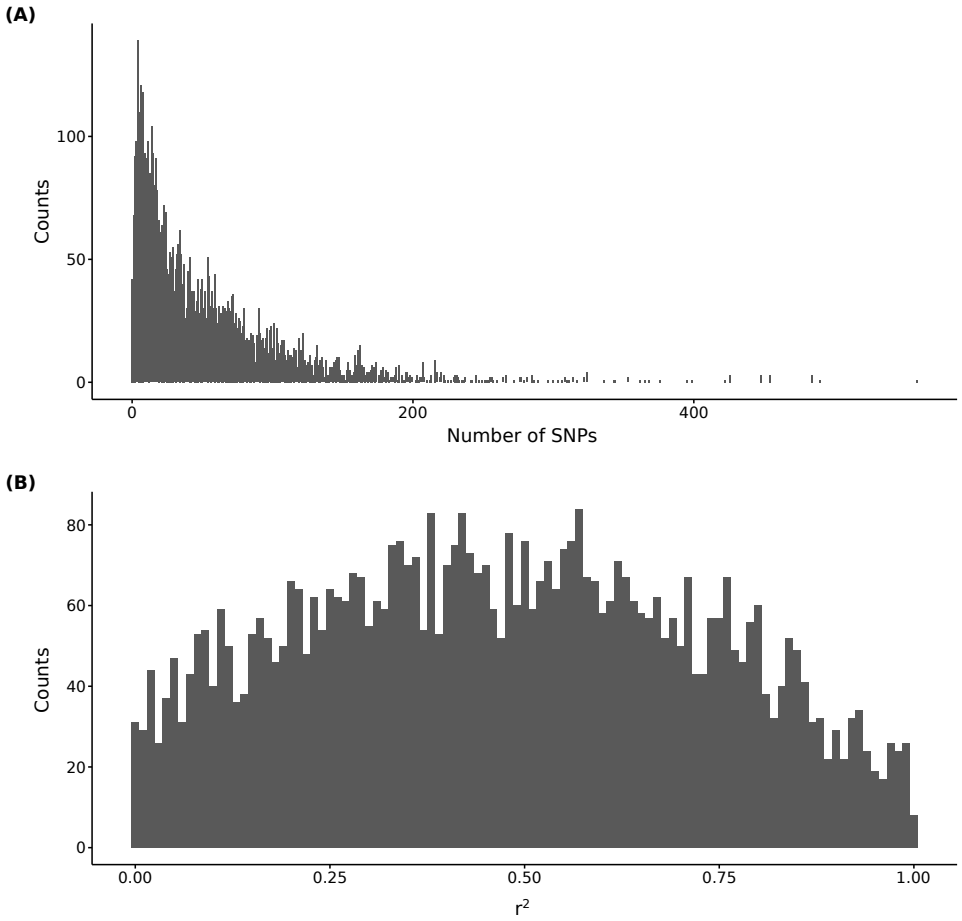

Figure 4.4. Genetic variants underlie chromatin peaks but fail to correlate with mark levels at most promoters.
Genetic variation could not be correlated with H3K4me3 abundance in promoters. (A) The number of SNPs identified in each H3K4me3 promoter peak. (B) The r^2 values for linear model fitting H3K4me3 abundance to underlying genetic variants.

4.3.4. Genetic variants underlie promoter activity

We wanted to explore if genetic variants underlying gene promoters could explain their strain-to-strain epigenetic variation. We identified H3K4me3 peaks within 1kb of transcription start sites and queried the genetic variants in those regions between our nine strains. Out of 5,380 peaks, the number of SNPS identified ranged from zero to 559, but only 42 had no called genetic variants across our nine strains (Figure 4.4A). We then

looked to see how well these variants correlated with the magnitude of the H3K4me3 peak they lay in. We fit the variation in H3K4me3 peaks to the called variants between strains and a broad spectrum of correlative power (Figure 4.4B). Generally, the local variants within promoter H3K4me3 could not explain the magnitude of said peak, suggesting that other factors, such as trans-regulation or distal histone methyltransferase binding sites, must be contributing to promoter epigenetic activity.

4.3.5. Chromatin state corresponds to genetic background

Next, we summarized epigenetic state in aggregate across H3K4me1, H3K4me3, and H3K27me3 measurements genome-wide with ChromHMM. We segmented the genome into 200 bp sections and measured the appearance of a histone modification in that range. We then classified these sections into one of fifteen chromatin states, based on their epigenetic marks (Figure 4.5A, left). We looked across each gene body in the genome and calculated how well chromatin state correlated with gene expression (Figure 4.5A, middle). We used these findings, as well as ENCODE annotations, to interpret the identity of each chromatin state (Figure 4.5A, right). We classified promoters, enhancers, and repressors by these marks. Since each individual mark had already been shown to have a genetic signal (Figure 4.3), we wanted to validate that that signal had not been lost with this transformation via ChromHMM. When we performed hierarchical clustering on genome-wide chromatin state, strains from a more similar genetic background generally had more similar chromatin states (Figure 4.5B). These findings demonstrate that these chromatin states correspond to known epigenetic signals, cluster by genetic background, and have differential correlations with gene expression, enabling us to go on to study how they might mediate the relationship between genetics and gene expression.

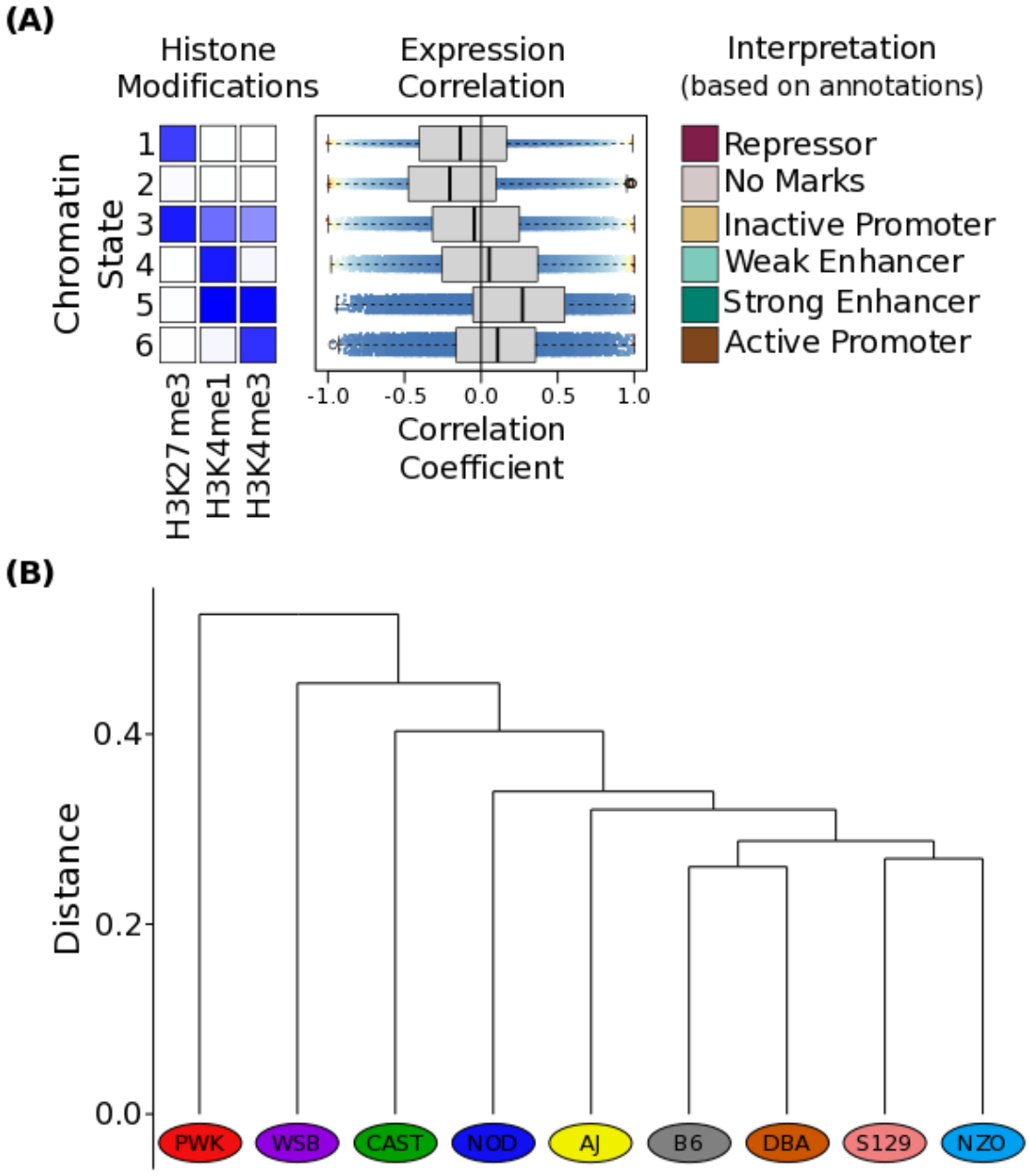

Figure 4.5. Genome-wide chromatin state between strains correspond with genetic similarity.
Chromatin state, as measured by ChromHMM analysis on H3K4me1, H3K4me3, and H3K27me3. (A) Enrichment of each histone modification across six distinct chromatin states (left), their correlation with gene expression (middle), and our annotation of their identity (right). (B) Clustering analysis on global chromatin states for each strain.

4.3.6. Local chromatin state correlates with gene expression

We next wanted to assess how the histone modifications around the gene body related to gene expression at a single-gene level. We considered the chromatin state across the gene body itself, as well as 1,000 bp up and downstream (Figure 4.6A). We assessed this for each strain in our dataset. These differential chromatin states around the gene were then scaled to a single dimension to make it easily comparable to gene expression (Methods). We queried gene expression for each individual gene across our inbred strains (Figure 4.6B) and compared that to the scaled chromatin state (Figure 4.6C). *Hsd3b1*, for example, had a local chromatin state that was fairly correlated with its expression ($r = 0.94$, $p = 0.00017$). We performed this analysis on all genes in our dataset and have plotted the distribution of correlation coefficients (Figure 4.6D). While approximately 34% of genes had local chromatin states that correlated with their expression with a coefficient greater than 0.5, this was fewer than observed in the genetics-to-expression correlation (Figure 4.2D). While local chromatin state had an observable relationship with some gene expression, our analysis of this local epigenetic control over gene expression explained less variation than local genetic variation.

4.3.7. Local genetics and chromatin both contribute to gene expression differences

For each gene, we compared the correlation between genetics and expression with the correlation between epigenetics and expression for all genes with a cis-eQTL LOD > 10 (Figure 4.7). The expression of 1,385 genes were correlated well with both local genetic and local epigenetic variation. However, many genes were only correlated with one of those underlying components. Genes whose expression levels were well correlated with underlying genetic variation (Figure 4.7, right) were enriched for GO terms relating

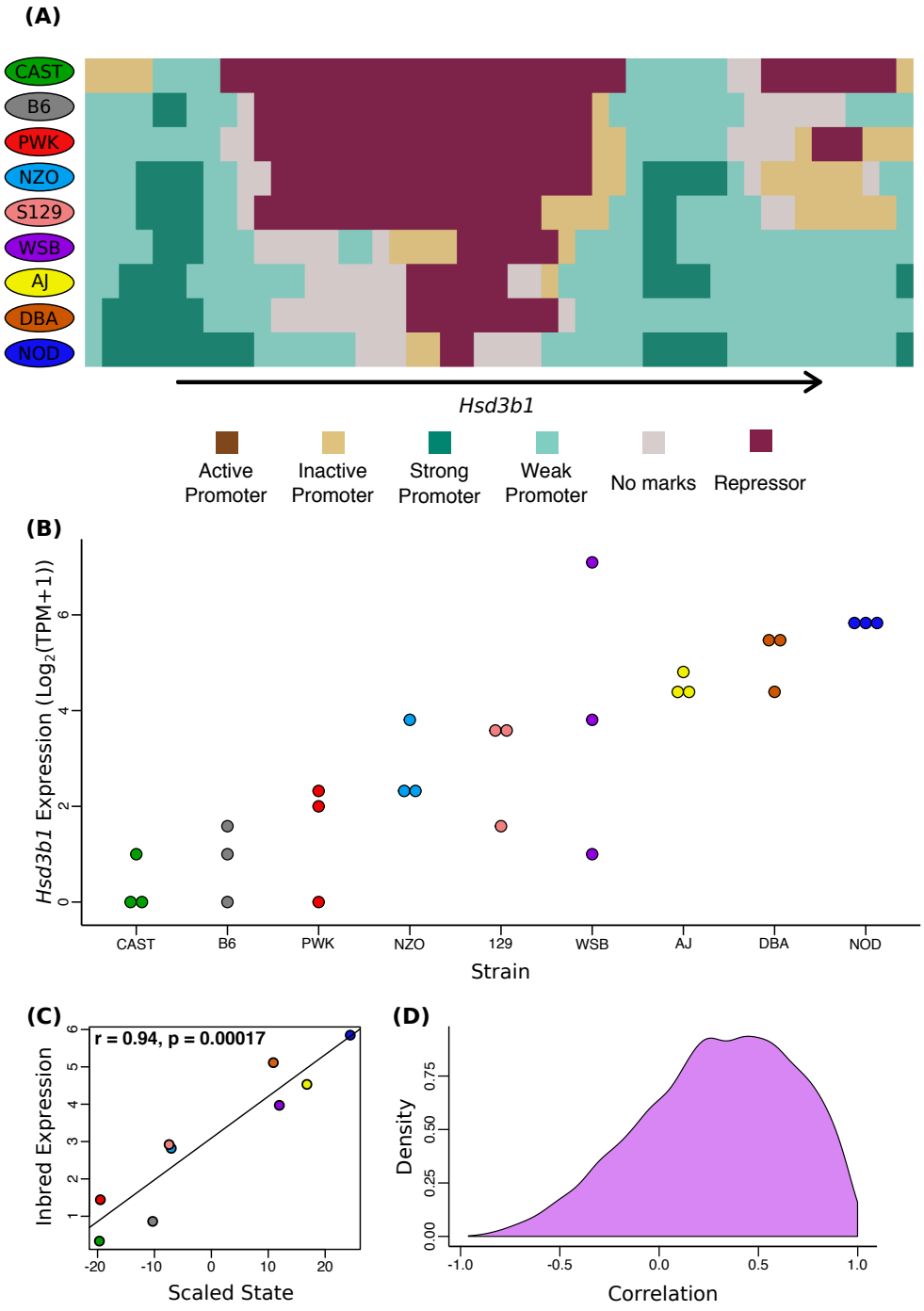

Figure 4.6. Local chromatin states correspond to transcript abundance in inbred mice.
Representative analysis of local chromatin variation correlating with gene expression in inbred mice. (A) ChromHMM-based chromatin state for *Hsd3b1* across hepatocytes from inbred mice. (B) Expression of *Hsd3b1* in hepatocytes of inbred mice. (C) Correlation between local scaled chromatin state and inbred expression for *Hsd3b1*. (D) All correlation coefficients from genes with a DO cis-eQTL LOD > 10.

to integral cellular components, such as components of the membrane, as well as some more dynamic cellular functions like monooxygenase activity and heme binding (FDR < 0.05). Genes whose expression levels were correlated with underlying epigenetic variation were not enriched for crucial, stable cell processes like membrane components. Instead, they were enriched for terms relating to potentially more dynamic processes, such as receptor regulator activity and iron ion binding (FDR < 0.05). These data imply that genes whose expression is tightly linked to their local genetics may be more stably expressed, housekeeping genes, whereas genes that are controlled by more dynamic, epigenetic regulation may participate in coinciding dynamic cellular needs. It is important to note that neither genetics, epigenetics, nor genetics and epigenetics correlate with expression levels for all genes (Figure 4.7). This could be due to trans-acting factors, as well as an incomplete model of chromatin state. Taken together, gene expression in hepatocytes is regulated by local genetics and epigenetics to a varying degree across the genome.

4.3.8. Distal chromatin state can be linked to gene expression

We investigated if distal chromatin could explain some of the gene expression differences that weren't well correlated with local epigenetic changes. Local genetics, which was far better correlated with gene expression in our model, was derived from DO eQTL data; however, those cis-eQTLs encompassed far larger sections of the genome than our local epigenetics analysis did (Figure 4.2). Therefore, it stands to reason that some genetically-influenced epigenetic changes outside of the gene body, but within the DO cis-eQTL, could improve the epigenetics-to-expression correlation at a gene level.

We expanded our scan of chromatin state to include distal chromatin changes that might influence gene expression.

To prioritize relevant distal chromatin for genes, we used a previously published list of liver-specific enhancers that were already associated with specific genes (Figure 4.8) (Shen et al., 2012). We measured the chromatin state at promoters for genes with

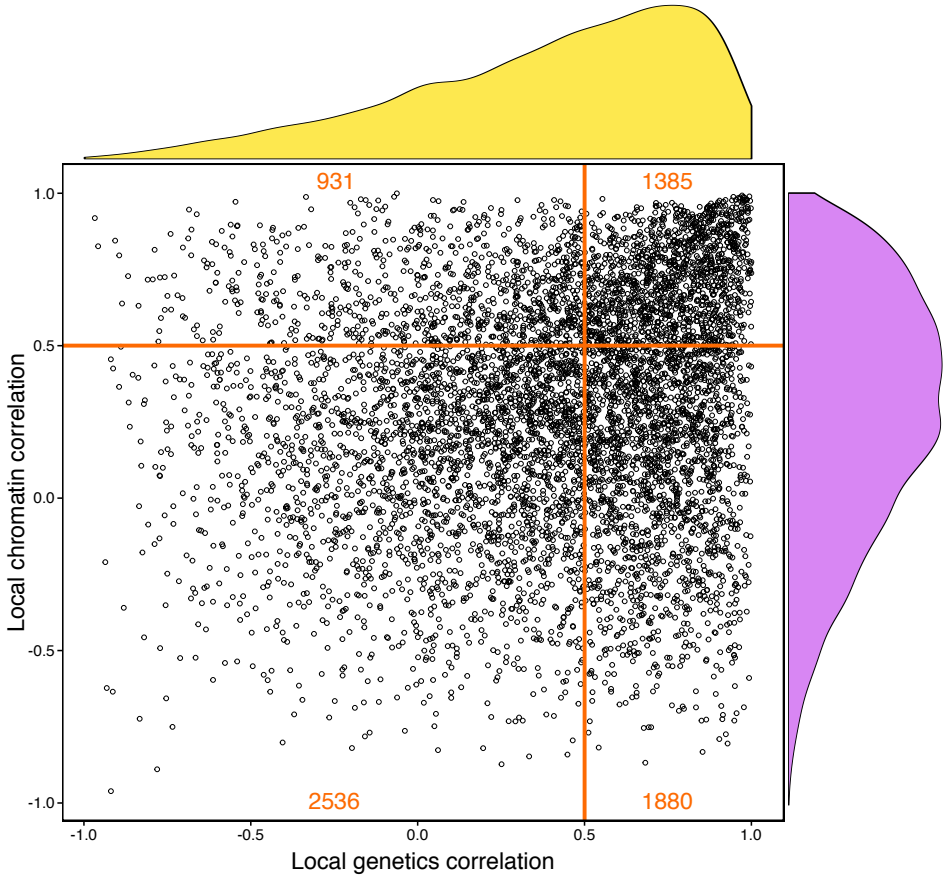

Figure 4.7. Local genetics and local epigenetics correlate with gene expression. Correlation coefficients of local genetics vs gene expression plotted against correlation coefficients of local epigenetics vs gene expression. Orange lines divide genes by their coefficients into four quadrants: gene expressions correlated with genetics and epigenetics (top right), gene expressions correlated with genetics but not epigenetics (top left), gene expressions correlated with epigenetics but not genetics (bottom right), and gene expressions not correlated with genetics nor epigenetics (bottom left).

liver-specific enhancers (Figure 4.8A), as well as promoters for genes with cortex-specific enhancers as a control (Figure 4.8B). In general, a greater percentage of the promoters with liver-specific enhancers had active marks than the promoters with cortex-specific enhancers (Figure 4.8A,B). Further, a greater percentage of genomic locations annotated as liver-specific enhancers had active chromatin than cortex-specific enhancers (Figure 4.8C,D). Taken together, enhancers specific to the liver are active in our hepatocytes and are associated with genes that are more likely to have active promoters.

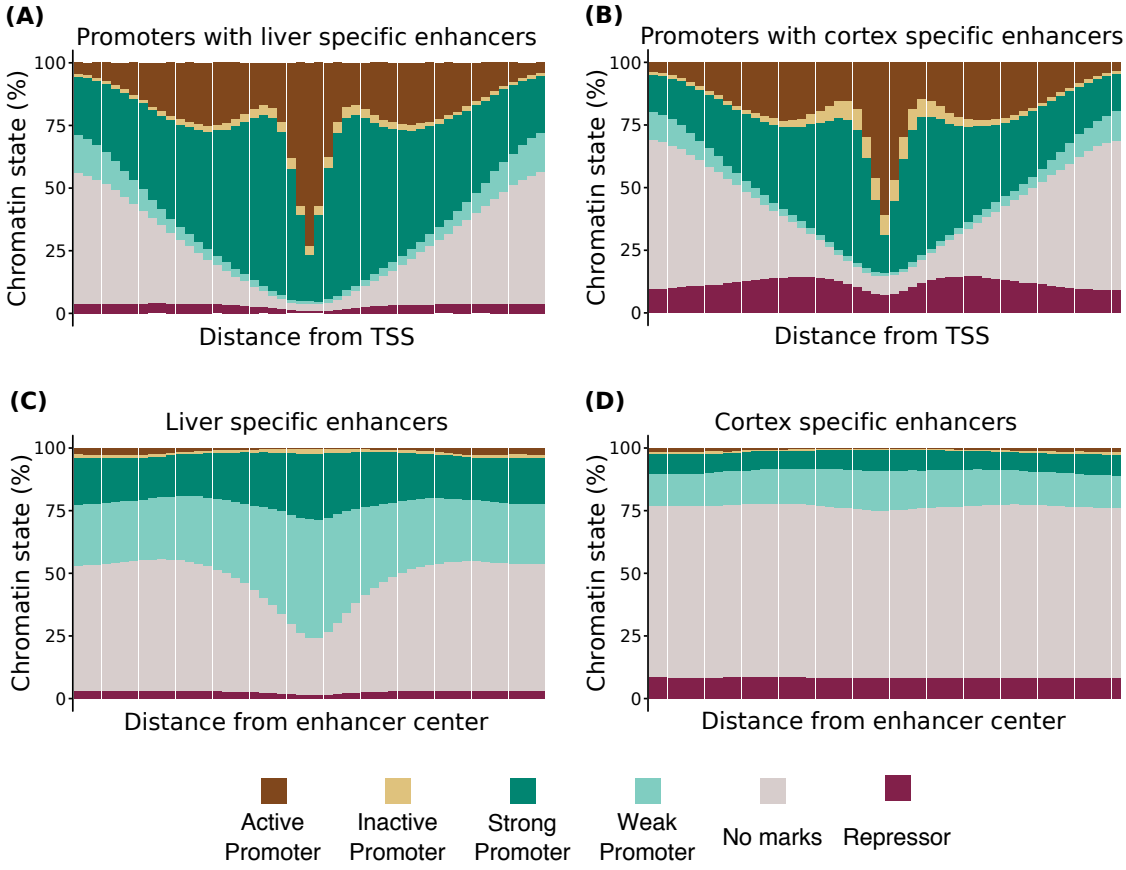

Figure 4.8. Chromatin state for promoters and cell-type-specific enhancers
Percentages of ChromHMM-based chromatin states around cell-type-specific enhancers, as well as the promoters targeted by those enhancers. (A) Chromatin states around the TSS of genes with liver-specific enhancers. (B) Chromatin states around the TSS of genes with cortex-specific enhancers. (C) Chromatin states around the liver-specific enhancers. (D) Chromatin states around the cortex-specific enhancers.

To test the applicability of these enhancers to our model of epigenetic regulation of gene expression, we looked to see if any gene's enhancers better corresponded to their expression than local chromatin state. We found that the local chromatin state of *Hddc2* did not correlate well with its expression (r = 0.4354). Its expression was highest in CAST and PWK hepatocytes, but its promoter did not clearly reflect why this would be the case (Figure 4.9A,B). Its annotated enhancer did show a CAST and PWK specific effect, in presence of strong enhancer marks (Figure 4.9C). Applying this analysis genome wide may improve the correlation between chromatin state and gene expression. These findings both validate that the liver-specific enhancers in this dataset are largely active in our hepatocytes, as well as that we can identify their activity with these ChromHMM-based chromatin states.

4.3.9. Additional epigenetic measurements add complexity to our model of gene expression in hepatocytes

To better classify the chromatin state in each of our strains, we incorporated an additional epigenetic modification, H3K27ac. This is a canonically activating mark found at regulatory elements that promote gene expression, such as enhancers (Shlyueva et al., 2014). Looking across this H3K27me3, H3K4me1, H3K27ac, and H3K4me3, we identified 15 non-redundant chromatin states with ChromHMM (Figure 4.10, left). These differentially correlated with gene expression (Figure 4.10, middle). We used these correlations, along with ENCODE annotations, to assign each state a functional interpretation (Figure 4.10, right). Because the addition of H3K27ac allowed us to segregate sub-states from the previous ChromHMM run (Figure 4.5A, left), we expected

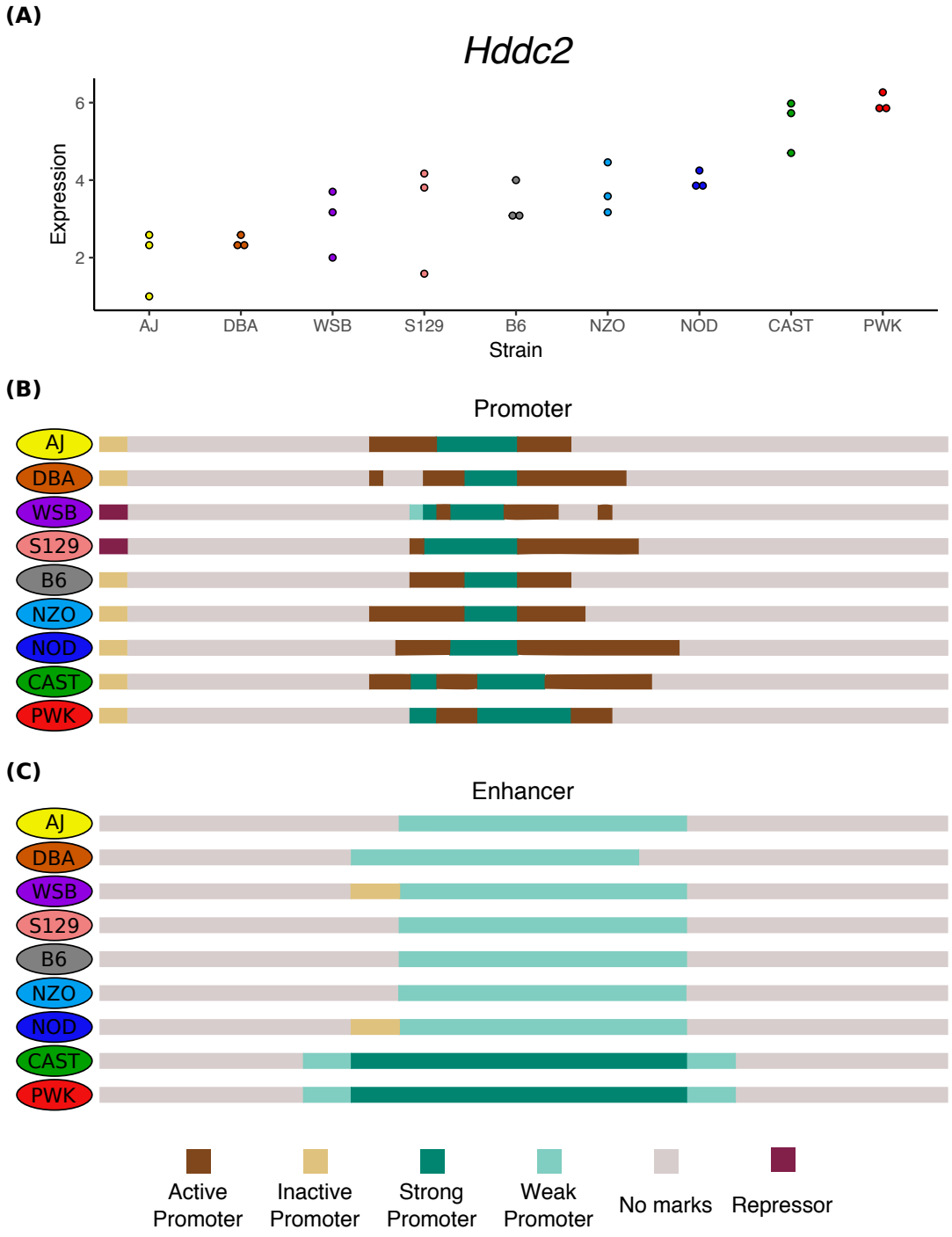

Figure 4.9. Chomatin state at enhancer elements correspond to gene expression
Local and distal regulatory elements for *Hddc2*. (A) Expression of *Hddc2* (Log_2(TPM+1)). (B) Chromatin state at the promoter of *Hddc2*. (C) Chromatin state at an enhancer of *Hddc2*.

to see greater variability in the correlation with gene expression. For example, as H3K27ac is annotated to activate enhancers, we thought that its incorporation might separate out highly active enhancers – which would correlate well with gene expression, from minimally active enhancers – which would not correlate with gene expression. Instead, we observe a limited range of average correlations with gene expression (Figure 4.10, middle), and had poor correlations with gene expression in aggregate (data not shown). The addition of these data adds valuable information to our model of chromatin state; however, our state assignments and/or current annotations of those states are presenting a challenge for the interpretation of that information.

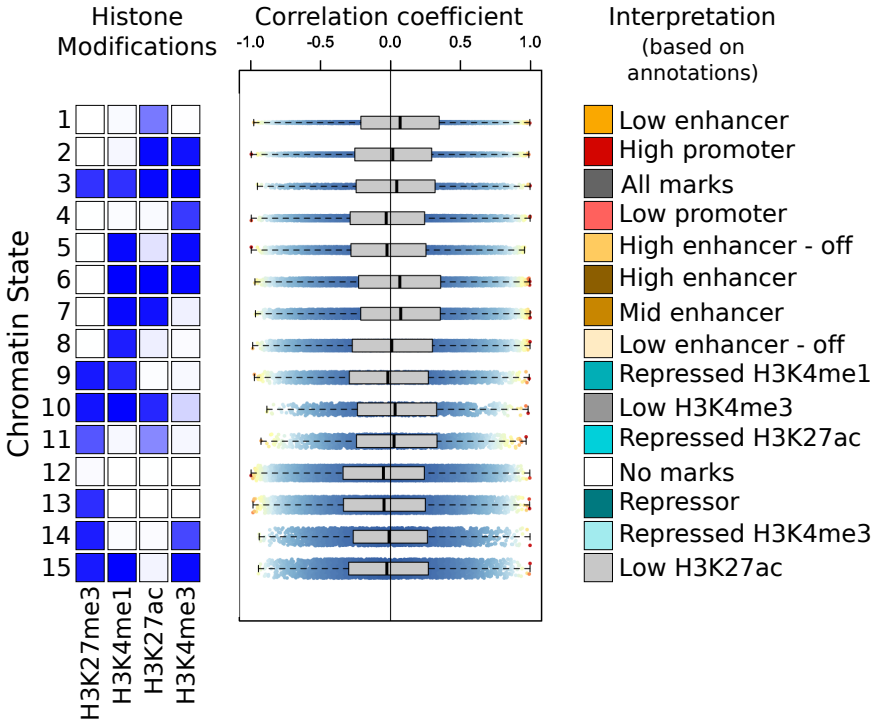

Figure 4.10. Chromatin state corresponds to known genetic features.
H3K27me3, H3K4me1, H3K27ac, and H3K4me3 divide the genome in 15 chromatin states. These epigenetic marks classify non-redundant states (left) that differentially correlate with gene expression (middle). They can be classified into known regulatory elements, based on these correlations, as well as previous annotations of similar chromatin states (right).

4.4.Discussion

By integrating multiple genome-wide measurements, we have been able to define features of the relationship between genetics, epigenetics, and gene expression. We used the genetic, epigenetic, and transcript abundance diversity across nine inbred strains of mice to assess how variation in genetics and chromatin relate to gene expression. We utilized the DO cis-eQTL from the liver as an avatar for local genetic effects of gene expression in hepatocytes. We observed many significant liver DO cis-eQTLs that correlated well with observed gene expression. Additionally, we measured the relationship between local chromatin state and gene expression, but those two measurements were more poorly correlated, on average. We were able to identify some underlying genetic signatures of epigenetic changes. Nonetheless, it will be necessary to continue and enhance these analyses before we have a more complete understanding of how genetic variation alters chromatin state. The work described here frames a pipeline for the continued integration of genetic, epigenetic, and transcriptomic data to build a local model of gene expression.

We demonstrated that local allelic variation frequently underlies transcriptomic differences. This could, in part, be due to coding mutations that affect the transcription or degradation rate of the gene product, as local regulatory variation was in linkage disequilibrium with coding variation. Another possibility is that this variation affected regulatory elements around the gene. Often, we observed coinciding epigenetic changes that correlated with gene expression. This could represent variants in promoters or enhancers that have consequences on their ability to be identified as regulatory elements and therefore to function normally in gene regulation. That said, many genes with local

genetic variation that correlated with transcript abundance did not exhibit measured chromatin changes. There could be several explanations for this, some technical and some biological. From a technical standpoint, our epigenetic data nowhere near completely represents the breadth of modifications that comprise the chromatin state around a gene. In addition to H3K27ac data, which has not yet been incorporated into this analysis, there are other histone modifications and chromatin alterations with known or potentially unknown functions in the regulation of gene expression. Additionally, ChromHMM puts ChIP-seq data, which is a quantitative measurement, into a Boolean measurement. More nuanced differences in chromatin state, more subtle than on-off changes of a histone modification, would go undetected by our analysis. Moreover, the windows in which we surveyed genetic variation – the DO cis-eQTL LOD peaks – were wider than our windows of chromatin state, only 1 kb out from the start and end of each gene. Therefore, distal enhancers that differentially regulate gene expression are likely included in our genetic analysis but excluded from our epigenetic pipeline. As demonstrated in our preliminary work on distal enhancers, these may account for some proportion of gene expression differences unexplained by our current model of chromatin state. Biologically, there could be gene regulation events affected by genetic variation, but not mediated by epigenetic changes. For example, transcription factors whose binding site near a gene varies from strain-to-strain might exert differential expression of that gene without a corresponding change in chromatin state. While genetic variation frequently correlated with gene expression differences, we are only partially able to decipher an epigenetic mechanism for those effects.

Occasionally, our analysis revealed gene expression variation that correlated with epigenetic changes but not allelic variation. Like in the case above, this could be due to technical and/or biological effects. Our interpretation of potentially causative genetic variation was based on the allele coefficients assigned in the liver DO cis-eQTLs. While these proved to be informative for many genes in our analysis, the data have limitations. The DO dataset was collected from livers and therefore could contain signals that do not correspond well to our dataset from purified hepatocytes. We attempted to reconcile this, at least in part, by determining if genetic variants within a ChIP-seq peak could explain the variance in a histone modification; however, this approach would have to be more thoroughly exploited to determine unequivocally when underlying genetic variation might explain differential chromatin state. The limitations of our model could also be due to trans-acting factors with their own genetic variation that modulate gene expression via epigenetic changes. Since these were inbred strains, unlike the DO mice that were used to generate the cis-eQTL data, trans-acting factors, such as histone methyltransferases, would intrinsically co-vary by strain. These could differentially deposit histone modifications to a region surrounding a gene by virtue of their own variation, regardless of any local genetic variation. So, while chromatin state was correlated with gene expression, the frequency with which it varies independent of underlying genetic variation must be investigated further.

We began a preliminary analysis based on the addition of a fourth histone modification to our dataset, H3K27ac (Figure 4.3, Figure 4.4, Figure 4.10). These data were re-processed, from alignment to peak calling. While the development of the methodologies mentioned here will be incorporated into future analyses, the results

presented here should be interpreted lightly. The poor epigenetic-to-expression correlations are suggestive of an alignment issue with these data, so while methods mentioned will continue to be included, the processing of these data will be revisited before that is completed.

We have demonstrated that a simple, local model of gene expression is powerful but incomplete. Local genetic variation correlates well with the expression of many of the genes in our dataset, but only a subset of those loci have coinciding epigenetic changes. These data show that significant eQTL data can be sufficient for predicting gene expression in parent lines in the same tissue. In the future, this could be validated by imputing the eQTL coefficients for DBA mice and calculating how well that correlates with observed gene expression. Further work in determining when genetic variation will alter gene expression through epigenetic mechanisms will be necessary to predict which variants are causative. This work builds upon the models of gene expression that the pairing of inbred and diverse mice has allowed us to create and should be incorporated when predicting the effects of genetic variation on gene expression in the future.

4.5. Contributions

This project was conceived and planned by myself with Drs. Tyler, Carter, and Petkov. Catrina Spruce generated the majority of the data collected. Anna Tyler and I shared the computational analyses. I generated the majority of the figures and wrote the manuscript.

CHAPTER 5: Discussion

5.1. Inclusion of variability to perturb a system reveals subtle biological insights

The basis of this overall project has been in a systems genetics approach: perturb a genetic system and take measurements genome wide. For each chapter, we have utilized one main disruption that allows for a systems analysis. In the project described in Chapter 2, we genetically knocked out a key meiotic gene, *Prdm9*, which led to a cascade of cellular and molecular phenotypes revolving around cell progression arrest resulting in infertility. In the second project (Chapter 3), genetic variation in PRDM9 binding sites genome-wide differentially allowed for its long ZF to bind. In the final project (Chapter 4), gene-to-gene differences in genetics and epigenetics defined variation across the transcriptome. The true power of these datasets lies in the diversity of each system. In Chapter 2, we describe a collection of samples at three timepoints, which increased the variation in our dataset to the point where we could identify substage-specific transcripts, leading to the key finding of the paper. In the project reported in Chapter 3, the genetic diversity between the B6 and CAST genomes provided built-in perturbations to a subset of Affinity-seq sites, defining a training and testing set for our computational model. Finally, in Chapter 4, we show how the diversity of genetics, epigenetics, and gene expression between nine genetically inbred strains enabled us to build a model that explored the relationship between each of these 'omic measurements and identify circumstances where that relationship is unpredictable.

When studying the relationship between cytological and transcriptional differentiation described in Chapter 2, we took multiple measurements across a developmental timeline. This approach revealed the complexity of the cellular phenotype

– a cytological delay preceding arrest. In addition, it increased the degree of variability across our samples, which enabled the use of PMCA to assign transcripts to specific meiotic substages. It was this substage assignment that revealed the most interesting finding – that the transcriptional and cytological programs in meiosis can become uncoupled when development is arrested. This finding not only has implications for meiotic research, but also will impact the way we study transcriptomics in other perturbed systems as well. The developmental diversity within our dataset was crucial for the proper analysis and interpretation of our data.

By incorporating data from two strains of mice in the work described in Chapter 3, we created built-in training and testing sets for our computational model of zinc-finger binding. We utilized Affinity-seq data from PRDM9^{Dom2} to both the B6 and CAST genomes. The naturally occurring genetic variation between these two strains of inbred mice provided a subset of perturbed PRDM9^{Dom2} binding sites that could be used to evaluate the power of our model. While there is genetic variation between binding sites within the B6 genome as well, the incorporation of CAST samples offered binding sites with a single SNP in otherwise identical sequences, generating a computational validation technique within our dataset.

Through the use of nine inbred strains of mice, as well as the utilization of data from a diverse, outbred population of mice, we built a model that correlates genetic and epigenetic variation with gene expression (Chapter 4). As in Chapter 3, these types of models can be built within a single genetic background; however, they are strengthened by the inclusion of diversity. For example, with pre-existing genomic annotations, one could compare all promoters within a single strain of mouse to examine genetic and

epigenetic commonalities between these regulatory regions. The context, regulation, and expression of each of the genes those promoters affect would influence the interpretation of those data. Thus, our approach utilized multiple strains, which allowed us to analyze each promoter, for example, at a gene-by-gene basis. We identified differences in the way that the expression levels of individual genes are regulated and perturbed by genetic and/or epigenetic differences, as well as gene expression that is resistant to these perturbations. The natural variation present in this dataset enabled a more precise examination of gene regulation, improving the computational models that summarize it.

5.2. Integration of cellular and molecular data improves biological accuracy of computational models

In the work reported in Chapter 2, we integrated gene expression and cellular phenotype data to model how transcriptomic and cytological differentiation programs interact when meiotic development is perturbed (Fine et al., 2019). It is far more typical to survey one or the other when studying developmental systems, particularly spermatogenesis. This can be due to cost restraints, scarce cell types, or other factors affecting study design. However, we demonstrate the value and importance to assessing cellular and molecular phenotypes in perturbed systems, as well as creating or using innovative data integration techniques to get the most out of the information collected.

Studying cytological data alone in this system would have led to accurate, but incomplete findings. These data showed that there were no healthy *Prdm9*$^{-/-}$ germ cells past the zygotene/early-pachytene substages, reflecting their arrest in meiotic development at that stage. However, had less already been known about *Prdm9*$^{-/-}$ biology, the mechanism of this arrest would have been nearly impossible to determine. Improper

chromosomal synapsis, persistent DNA damage, and lack of formation of the sex body were all cytologically apparent in *Prdm9*$^{-/-}$ germ cells, and theoretically any of them could cause meiotic arrest. Transcriptionally, we uncovered Cell Death signaling co-occurring with the cytological phenotypes at 16 dpp, as well as G2/M DNA damage checkpoint activation at 12 dpp, which preceded any cytological phenotypes. These findings strongly suggest that the primary molecular trigger for cellular arrest in *Prdm9*$^{-/-}$ germ cells is specifically persistent DNA damage, which would not be apparent without transcriptomic profiling.

Conversely, without cytological analyses, our study of *Prdm9*$^{-/-}$ germ cells would have been thorough, but inaccurate. While the aforementioned molecular drivers of cellular arrest were revealed through RNA-seq, established cellular markers for late-stage meiosis were also expressed. Specifically, late-pachytene/diplotene transcripts were found in our RNA-seq samples from *Prdm9*$^{-/-}$ testes, despite there being no late-pachytene/diplotene cells in those samples. This demonstrated that transcriptional progression had continued on in a semi-normal fashion, despite cellular arrest. However, without cytological staging – or the knowledge of the well-established *Prdm9*$^{-/-}$ arrest around early-pachytene – these results could have indicated progression through the G2/M DNA damage checkpoint in this meiotic mutant. It was solely through the integration of transcriptomic profiling and deep cellular phenotyping that we were able to identify a likely mechanism for cellular arrest and avoid misinterpretation of cytological and transcriptional uncoupling.

Relevant to this type of data integration, we validated the applicability of PMCA, a computational tool that associates variation between two distinct measurements (Ball et

al., 2016). Specifically, we were able to identify which cell type a transcript belonged to from bulk RNA-seq and cellular phenotyping. This was crucial for the interpretation of both the cytological and transcriptomic data. PMCA could be used to relate cell abundances and gene expression in other systems as well, from assigning transcripts to cell types in the cancer microenvironment, to determining cell type specific gene expression differences in the aging brain. Between the current cost of scRNA-seq and wealth of previously collected bulk sequencing, this will continue to be a useful tool for the foreseeable future. Single cell RNA-seq data from this study, for example, would previously have been challenging to interpret. Late-pachytene/diplotene transcripts appearing in the same cells as late-leptotene/zygotene transcripts, as we hypothesize is the case in *Prdm9*$^{-/-}$ germ cells, could have meant a partial failure of transcriptional arrest or a partial delay in transcription. It was only through the integration of transcriptional and cytological data with PMCA that we were able to identify not just to co-expression of transcripts that didn't belong together, but what that functionally meant for the cell.

5.3. Integration of multi-omic data increases information attainable from each measurement

Here, we utilized two separate multi-omic datasets to demonstrate the power of applying multiple genomic measurements to a biological system. First, as reported in Chapter 3, we used genetic sequence data and protein binding data to build a computational model for the binding affinity of a long zinc-finger array based on the genetic variants in its binding site. Second, described in Chapter 4, we compared the annotated genetic variation across nine inbred strains of mice to quantitative measures of

histone modifications and RNA abundance. Both of these endeavors uncovered system complexity and interactions that would not have been apparent from single-omic studies.

In analyses described in Chapter 3, we used linear models to predict how genetic variation in the binding site of PRDM9^{Dom2} would affect its affinity. The most straightforward approach, which had already been tested, was to use the variance across the binding sites in a single genome to compose a binding motif for PRDM9 (Figure 3.1A). This model does accurately identify the most important positions within the PRDM9 binding site, the nearly-required nucleotides at its anchor positions. However, the systematic perturbations and measurements we made demonstrated the limitations of this model. First, by introducing PRDM9^{Dom2} to the comparative genetic diversity of CAST genome, it was apparent that bases within the binding site that were not well represented in the motif had an influence on binding affinity (Baker et al., 2015). Through the added dimension of affinity for each genetic sequence – measured by read counts per site, we were able to quantify the effects of each nucleotide across the entire binding site, as well as interactions between multiple positions. By fitting variation in the binding sites in the B6 genome to the measured affinity of PRDM9^{Dom2} for each site, we were able to train a model to identify the effect of each nucleotide on binding affinity, and subsequently test its predictive power with the natural genetic variation that exists in the PRDM9^{Dom2} binding sites across CAST genome. This multi-omic analysis generated a more predictive model of PRDM9^{Dom2} binding than would have been possible otherwise and enabled us to test the power of that model.

As described in Chapter 4, we modeled gene expression by integrating genetic, epigenetic, and transcriptomic data from mouse hepatocytes from nine different inbred

strains of mice. When studying any one of these genome-wide measurements in isolation, it is practically impossible to infer the others, thus limiting our ability to model the mechanistic complexity of a system. Including all three genome-wide measurements in a single study provided powerful computational capabilities, here to classify transcripts by their expression being correlated with local genetic and/or epigenetic variation. Genetic variation is the likely underlying cause for all downstream differences, changing the activity of regulatory elements and, in turn, gene expression. Histone modifications are the frequent mechanism by which these variants differentially regulate expression. Even among the histone modifications we measured, each mark provided a unique aspect of regulation, from distinguishing enhancers from promoters to determining the on/off state of a given regulatory element. Additionally, transcript abundance is the most straightforward way to measure and model the combined effects of genetic and epigenetic variation on gene expression. Without any one of these 'omic measurements in our model, we would lose our ability to infer the cause, mechanism, or magnitude of differences in gene expression.

5.4. Updating the framework for modeling biological systems

A linear model of gene expression provided researchers with a framework for modeling biological systems for decades, but, more recently, genome-wide measurements have demonstrated its limitations. Gene expression can no longer be considered an isolated, unidirectional process – there are system-wide interactions and multifaceted mechanisms of regulation that complicate the DNA-to-RNA-to-protein relationship. A more updated framework for modeling biological systems should be nonlinear and broader than just DNA, RNA and protein (Figure 1.1). The chromatin landscape, for one,

plays a crucial role in mediating the relationship between DNA and RNA and should be considered in models of gene expression. For example, we observed epigenetic variation that correlated with changes in gene expression, despite the absence of underlying genetic variation (Figure 4.7). Further, cellular phenotypes should also be considered when examining dynamic cell systems, as we have noted that molecular measurements can independent relationships with cellular phenotypes (Figure 2.10). Incorporating multi-omic measurements increases the power of each individual measurement, such as including a quantitative measure of differential protein binding when studying how genetic variation alters binding site selection (Figure 3.12). Across several examples, we have demonstrated the utility of including multiple genome-wide measurements to understand a genetic system.

Taking a multi-omics approach to studying biological questions can help address a persistent problem encountered when modeling biological systems: the determination of causation versus correlated "passenger" observations. For example, in many disease studies, causative mutations are hard to identify because of a wide preponderance of mutations, generally. Cancer genomes in particular accumulate genetic mutations frequently and rapidly, making the identification of passenger mutations a tenacious problem. Pairing genetic screens with downstream, genomic assays could help reveal causative mutations, those with effects, by deciphering which genes are misregulated, which genetic loci are mismethylated, etc. Another application of a multi-omics approach would be in disease systems that show drastic differences in cell abundance. It can be challenging to determine if gene expression changes reflect programmed differences within a single cell type or simply loss of a cell type. Similar to what we demonstrated in

Chapter 2, co-analysis of molecular and cellular assays can reveal which gene expression differences are a consequence of cell loss. This could be applied in fields such as neurodegeneration, where cell loss is an intrinsic part of the disease phenotype but likely a consequence of cell type-specific gene misregulation. As we continue to learn the complexities of genetics, health, and disease, we must use equivalently complex tools to interrogate those systems.

5.5. Future directions

Each of these projects has utilized multi-omic models to learn about complex genetic systems. Across three projects, we perturbed a biological system, measured the system at multiple genome-wide levels, and computationally modeled the complexity of the system to learn more about how it functions. However, by further studying these systems, either by increasing the degree of perturbations we make or broadening the scope by which we assess them, we can learn even more. Below I make proposals for extending this work.

5.5.1. Validate that transcripts uncoupled from cytological differentiation in meiotic arrest are expressed in alternative substages.

We demonstrated that cellular and molecular differentiation processes can become uncoupled in a perturbed developmental system (Fine et al., 2019). We showed that cytological and transcriptional progression through meiosis become abnormally uncoupled in *Prdm9*$^{-/-}$ mice. Despite there being no cells staged as late-pachytene/diplotene by cytological markers, transcripts annotated as specific to that substage of meiotic prophase were present in *Prdm9*$^{-/-}$ testes. These transcripts presumably are being expressed in cells cytologically staged as late-leptotene/zygotene.

This was inferred by the use of a computational tool, PMCA, which assigns transcripts to specific substages by correlating mRNA abundance to substage representation within the same sample. This work demonstrates the importance of pairing deep cellular phenotyping with modern computational analyses.

While the presence of late pachytene/diplotene transcripts in *Prdm9*$^{-/-}$ testes alone demonstrates that cellular and molecular processes are uncoupled, their expression in specific other meiotic substages is currently limited to conjecture. Moving forward, it will be important to verify this result in order to properly interpret these data and to define a mechanism by which others can interpret similar findings. The two most straightforward ways to accomplish this would be to visually show late-pachytene/diplotene transcripts in late-leptotene/zygotene cells or to isolate late-leptotene/zygotene cells and sequence them for late-pachytene/diplotene transcripts. The first of these options could be performed with RNA fluorescence in situ hybridization (FISH) (Mahadevaiah et al., 2009). By this method, RNA molecules can be visualized within their cell-of-origin; however, it is limited by the specific genes one chooses to interrogate and their signal strength, as well as methodological challenges that can disrupt accuracy. Due to these limitations and biases, it can be challenging to use this method for validation, as one must first choose a true-positive gene and then accurately perform the experiment. Therefore, cell selection and re-sequencing is likely a better route for validation. There are multiple ways to isolate subtypes of germ cells, including single-cell RNA sequencing (scRNA-seq), cell sorting, and spermatogenic synchrony (Davis et al., 2013; Gaysinskaya et al., 2014; Green et al., 2018; Hermann et al., 2018). Utilization of each of these methods has associated limitations. Performing scRNAseq is the best of these methods to ensure that

co-expressed transcripts are coming from the same cell. However, scRNA-seq is expensive and prone to drop-outs. Moreover, the cell type we are interested in is rare and thus minimally represented in the germ cell population as a whole. Cell populations collected via cell sorting can be sequenced in bulk, reducing the risk of drop-outs and false-negatives, making this method potentially more accurate than current scRNA-seq methods. Nonetheless, cell sorting remains challenging with respect to spermatocytes and this is a problem potentially exacerbated by presence of germ cells exhibiting meiotic arrest. Moreover, this method relies on previously published markers for substage identity, which, as found by this project, might not always be accurate in perturbed developmental systems. Exploiting spermatogenic synchrony provides a balance between these two methods: it uses bulk sequencing for accuracy, but cell populations are assessed by the testicular context and can be confirmed by classical cytological methods for precision. Synchronization of spermatogenesis should be a useful tool for validating the uncoupling of transcription and cytodifferentiation in our model of meiotic arrest (Davis et al., 2013).

To this end, we analyzed bulk RNA-seq data from retinoic acid-synchronized germ cells from *Prdm9*$^{+/+}$ and *Prdm9*$^{-/-}$ mice. We collected germ cells at two time points: 8- and 14-day-post-injection of retinoic acid (dpi), which by cytological staining, correspond to tubule stages X-XI and VII-VIII respectively (data not shown). In wildtype mice, seminiferous tubules are in stages X-XI at 8 dpi contain spermatogonia and leptotene/zygotene spermatocytes. As these stages occur even during *Prdm9*$^{-/-}$ meiotic arrest, the same should be true of mutant samples. At 14 dpi, seminiferous tubules are in stages VII-VIII, and wildtype mice have spermatogonia, leptotene spermatocytes, and

late-pachytene spermatocytes. In *Prdm9*$^{-/-}$ samples, we would expect to find healthy spermatogonia and leptotene spermatocytes, as well as zygotene spermatocytes that are initiating arrest in their meiotic progress. Therefore, if late-pachytene/diplotene transcripts are present in both 14 dpi conditions, but absent from both 8 dpi conditions, it can be inferred that those transcripts are bring specifically expressed in arrested zygotene cells in *Prdm9*$^{-/-}$ testes. However, when we performed this experiment, late-pachytene diplotene transcripts were not expressed in any *Prdm9*$^{-/-}$ samples, despite their high relative expression in *Prdm9*$^{+/+}$ spermatocytes (Figure 5.1A). This was also true of the subset of late-pachytene/diplotene transcripts that changed their specificity in bulk samples (Figure 5.1B). However, after examining the cytological data paired with these

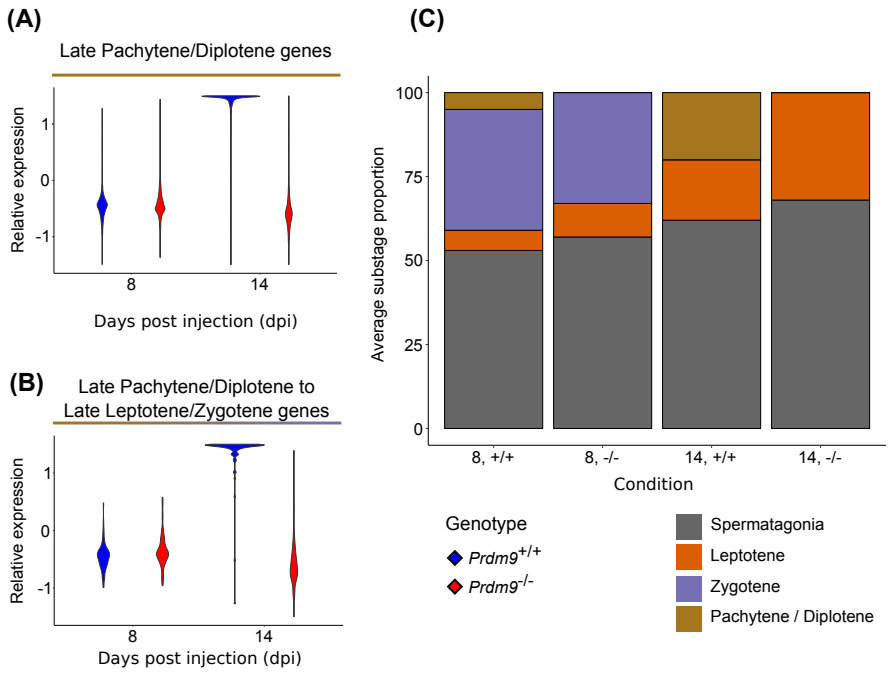

Figure 5.1. Exploitation of stage-synchonized testes fails to recapitulate findings from bulk germ cells.
(A) Relative expression of transcripts assigned to Late-Pachytene/Diplotene substage based on their wild-type expression patterns. (B) Relative expression of transcripts assigned to Late-Pachytene/Diplotene in wild-type samples but to Late-Leptotene/Zygotene in *Prdm9*$^{-/-}$ samples. (C) Cellular proportions of each substage in synchronized samples.

transcriptomes, it appeared that no post-leptotene cells were present in the *Prdm9*$^{-/-}$ samples at 14 dpi, despite their presence in the 8 dpi samples (Figure 5.1C). This suggests that arrested cells are cleared in this system before 14 dpi, making it impossible by this strategy to determine if uncoupled transcripts are expressed in late-leptotene/zygotene cells.

To identify the cell expressing the late-pachytene/diplotene transcripts detected in *Prdm9*$^{-/-}$ testes, more thorough sampling of synchronized germ cells or scRNA-seq could be performed. We surmise that at some time point between 8 and 14 dpi, arrested zygotene cells – those that may or may not be expressing late-pachytene/diplotene transcripts – must be eliminated from the seminiferous tubules. Sampling more frequently between these time points may be sufficient to catch the critical cell population. However, it is possible that, unlike unsynchronized testes, stage-synchronized testes eliminate arrested cells before ectopic uncoupled gene expression occurs. Therefore, a potentially more useful method would be to perform scRNA-seq on *Prdm9*$^{-/-}$ germ cells to detect uncoupled transcription. As mentioned earlier, this strategy would run the risk of imprecise transcript abundances due to loss of reads, but this problem might be circumvented by using a large cell population. This work would build upon the findings from chapter 2, strengthening the model that transcriptional progression can become uncoupled from cytodifferentiation in arrested meiotic cells.

5.5.2. Build a more generalizable model of long zinc-finger binding

We used a unique dataset to model how zinc-finger proteins select their binding site. The Affinity-seq dataset for PRDM9^{Dom2} in both the C57BL/6J and CAST/EiJ backgrounds provided us with the capability to measure how genetic variation in the

binding site of a long zinc-finger array alters the affinity of the zinc-finger to that binding site. Unlike *in vivo* assays for zinc-finger binding, our results were not confounded by protein cofactors or differentially accessible chromatin. Examining PRDM9^{Dom2} binding in the B6 and CAST genomes provided genetic variation in binding sites that coincided with affinity changes, further allowing us to study that relationship. Together, these enabled us to model the effect of SNPs across the PRDM9^{Dom2} binding site that alter binding affinity, as well as interactions between nucleotides at different positions that complicate these effects. Our findings suggest that a mechanism for these effects could lie in the DNA shape of a genetic sequence. These results could inform how models of long zinc-finger array binding should be constructed.

Modeling zinc-finger binding could be improved by adding greater perturbations to the system, at either the genetic or protein level. We could conduct PRDM9^{Dom2} Affinity-seq in the context of other genetic backgrounds to further perturb its binding sites. Observed PRDM9^{Dom2} binding sites did vary across the B6 and CAST genomes, but some positions – specifically anchor positions – had very minimal variation (Figure 3.1B). This could be a product of specific nucleotides being necessary (or nearly so) at those positions; however, since no position was completely non-variable, no single nucleotide in the binding site is completely necessary. Greater genetic variation may identify novel binding sites that lack nearly necessary nucleotides, informing us of what combinations of nucleotides across the binding site allow for these exceptions to be made. Moreover, we could introduce variability to the PRDM9 allele used for Affinity-seq. There are numerous alternative alleles of PRDM9 (Figure 5.2) that presumably have differing sets of binding sites. These zinc-finger arrays differ from PRDM9^{Dom2} to

varying extents. For example, PRDM9^{Dom3} has a single additional zinc-finger, whereas PRDM9^{Cst} only shares one finger with PRDM9^{Dom2}, apart from the first and last. Expanding this analysis to alternative alleles of PRDM9 could allow us to generalize our model by identifying relationships between zinc-fingers in the array and nucleotide importance and interactions in the binding site. These relationships could be applied to non-PRDM9 zinc-finger proteins, addressing the gap in our ability to predict the binding sites of long zinc-finger proteins genome wide.

Figure 5.2. Examples of alternative alleles of the zinc-finger array of PRDM9. Zinc-finger amino acid sequences for multiple alleles of PRDM9. Label designates the allele name and a genetic background on which it is found.

5.5.3. Further examine genetic influence on epigenetic state

We showed that local genetic variants rarely explain the variation in the intensity of H3K4me3 peaks at promoters (Figure 4.4). One limitation to this work was that it was performed on all H3K4me3 near promoters; however, not all of these demonstrated the same degree of variation. Unsurprisingly, when we examined r^2 values for each peak, ordered by the range of H3K4me3 activity, we observed that promoters with a wider range appeared to have greater degree of correlation with their underlying variants (Figure 5.3). But as even the high-variance promoters weren't all well correlated with genetic variation, it would be worth continuing to investigate this relationship, partially in the case of those H3K4me3 peaks that did have genetic variants predicteding peak

intensities. We could identify commonalities between variants that seem to affect H3K4me3 peak magnitude, such as their placement within the promoter or the frequency of specific polymorphisms. Also, broadening the genomic range of the variant search might improve our analysis, as not all causative variants are necessarily expected to fall within the H3K4me3 peak. A different method for causative SNP identification may also be beneficial. We could first query for common motifs in high-activity promoters to determine if low-activity promoters disrupt that motif. Further, this analysis could be performed at other types of regulatory elements. For example, previous work has found that H3K27ac peaks commonly have a genetic motif at the center of their peak (Quang et al., 2015). Perhaps variants in central motifs of regulatory elements would be better correlated with H3K27ac activity. While we can observe a relationship between genetics and epigenetics at some promoters, it would be worthwhile to continue this investigation more thoroughly and with a broader scope.

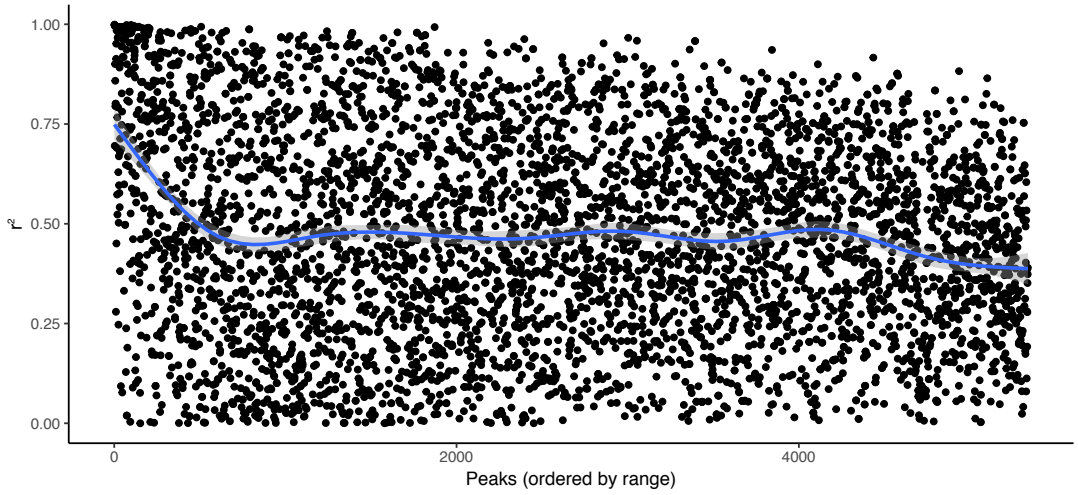

Figure 5.3. Peaks with a greater range of H3K4me3 activity show greater relationship with underlying variants.
The r^2 values for the linear models build to fit variation of H3K4me3 activity to underlying genetic variation. Peaks with a greater range of H3K4me3 activity seem to show overall higher predictive power.

5.5.4. Evaluate the predictability of our model of gene expression in hepatocytes

Our model studying the genetic basis of gene expression-correlated allele effects from the DO population with observed expression in the parental lines. We could assess our ability to predict gene expression in hepatocytes based on underlying variants by using the DBA mice as a testing set for our model of gene expression. One limitation of this analysis was its basis in allelic variation, rather than genetic variants. As DBA mice were not included in the generation of the DO mice, their alleles are not represented in that population and were therefore excluded from our analysis of the correlation between genetic and transcriptomic variation (Figure 4.2). However, this exclusion could be used to address the limitation of our analysis. After imputing DO coefficients for DBA alleles, based on their variant-level similarity to the other strains, those coefficients could be compared to observed gene expression. This would provide us with the ability to utilize the DBA mice for that analysis, as well as interpret the variant-level power of our correlation.

There are two apparent methods by which we could impute DBA allele effects. First, we could generate a similarity matrix for the genetic variants around a gene among all nine of our inbred strains. The DBA allele effect for the given gene could then be inferred to be equivalent to the effect of the most genetically similar strain. Alternatively, we could relate DBA to alternative alleles by using the alignment tool for CC lines that predicts the likelihood of a segment of DNA to belong to each DO founder strain. This reports a confidence score for each strain, which could be used as a weight for the DO coefficients of the other alleles. The sum of those weighted coefficients would act as the DBA coefficient for our analysis. Regardless of method, this imputed allele effect for

DBA would then be compared to the observed gene expression in DBA hepatocytes. We would expect that genes with a strong genetic-to-expression correlation will also have a good correlation between imputed genetic effect and observed gene expression in DBA. This analysis would suggest to us whether or not the relationship we observed between local genetics and gene expression is variant based and predictable.

5.5.5. Novel applications and innovations for genomic data integrations

Often the limiting factor for the analysis of a multi-omic dataset is the researcher's ability to interpret the amount of information that it provides. While a single lab may not have the resources to generate an extensive multi-omic dataset, the wealth of existing, published, available data is constantly growing. The greatest challenge then becomes the purposeful analysis to answer a precise biological question. Some existing tools could be applied more broadly to integrate multi-omic data, but on the horizon are new and anticipated data types for which interpretation and integration will require novel tools.

Many existing computational tools integrate multiple data types at the middle or late stage of analysis. Some have been used and discussed here, such as ChromHMM for multiple epigenetic marks, quantitative trait mapping for genetic and molecular data, and PMCA for transcriptomic and cytological data (Abiola et al., 2003; Ball et al., 2016; Ernst and Kellis, 2012, 2017). These are not the only methods to integrate multiple data types. GREAT compares ChIP-seq data with pre-existing functional annotations to assign a meaningful interpretation to the data (McLean et al., 2010). PATRI integrates genomic and transcriptomic data to identify biomarkers for diseases (Ukmar et al., 2018). As we learn more about biological complexity, it will be necessary to create new tools to

accommodate new insights. It is plausible to imagine a tool to integrate epigenetic data from histone modifications with chromatin data from three-dimensional DNA contacts in a way that would predict gene expression, or integrate transcriptomic data with epigenetic data, thereby predicting where gene regulation differs between multiple conditions. In addition to more middle- and late-stage integration tools, more early-stage integration tools are also needed. Of the aforementioned programs, ChromHMM is the most early-like tool but is limited in its utility, partially due to its binary mechanism for assigning chromatin state. An alternative method might maintain the quantitative nature of histone modifications while still allowing them to be integrated at an early stage, such as EpiCSeg or Segway (Hoffman et al., 2012; Mammana and Chung, 2015). Gene expression datasets could also be integrated at an early stage, allowing for the identification of differential splicing. Sometimes, due to data structure or the sequencing technology used, multiple gene expression datasets are integrated after gene quantification, meaning that isoform-level differences in gene regulation must be ignored. Integrating multiple types of -omic data at an early stage is common only with limited types of data. Gene expression data are usually processed within the context of known genetic variation but rarely in the context of epigenetic data. It is possible to conceptualize gene expression information that was processed within the context of data for transcriptional regulation, such that each gene would have an estimation of if it was being actively transcribed. While tools do exist to integrate various types of data, there are still paths for improving how we process multi-omic data.

Looking forward, there are new data collection methods that will require intricate and precise processing, particularly within the context of data integration. Perhaps the

most intriguing of these is single-cell genomics, which is becoming feasible for large numbers of cell types and diverse genome-wide measurements (Buenrostro et al., 2015; Chen et al., 2018; Hwang et al., 2018; Zheng et al., 2019). These data types provide a new and profound ability to assess the heterogeneity of a system but also present entirely new challenges in data integration. First, we should consider how to integrate single-cell data with bulk sequencing data. Currently, it remains unfeasible to perform single-cell analyses of every -ome, in every cell type, in every context – particularly given the time and resources that have already been put into measuring them with other methods. Thus, it is important to use relatively scarce single-cell data to enhance the interpretation of bulk datasets. Tools like PMCA, or other kinds of co-expression analyses, could potentially be utilized to measure cell type-specific gene expression by integrating bulk sequencing datasets across a perturbed system with scRNA-seq from a single condition. For example, scRNA-seq could identify cell type-specific transcripts, which could then be detected by bulk sequencing. Genes that co-vary with the cell type-specific transcripts could be assumed to be in the same cell type, informing us of how cell type frequencies are changing and perhaps how gene co-expression changes across conditions. Additionally, single-cell genomics must be integrated among multiple single-cell data types. Some approaches perform multiple assays in parallel (Angermueller et al., 2016; Stoeckius et al., 2017), but greater challenges will arise from integrating single-cell data that were collected separately. This will be an exciting, but challenging, problem. Single-cell genomics rely on the ability to assign a cellular context to the data collected. We are only beginning to assess how we can do that across multiple single-cell genomics methods and to then integrate those measurements for the same cell type (Duren et al.,

2018). If we can accurately integrate single-cell genomics, we will be able to observe the heterogeneity within cell types, identify subtypes of cells by the combination of their genome-wide states, and parse regulatory mechanisms in ways that have not previously been possible.

5.6. Conclusions

In these projects, I have demonstrated the current capabilities for surveying a biological system at a multi-omic level, the utility of current tools in the deliberate analysis of multiple data types, and the accuracy and precision achieved in the interpretation of data when studied within the context of the broader biological system. Across a spectrum of biological contexts, I have used a perturbation of the system – through genetic diversity, a targeted mutation, developmental time, or some combination of these – to uncover the molecular processes, such as gene regulation, that control those systems. This has been informative not only within each specific biological context but also within the schema of developing frameworks for when and how to integrate data types. The continued expansion of this kind of work will enable researchers to best integrate their datasets to create more predictive models of biology, through better analysis and interpretation of those data.

CHAPTER 6: Bibliography

Abiola, O., Angel, J.M., Avner, P., Bachmanov, A.A., Belknap, J.K., Bennett, B., Blankenhorn, E.P., Blizard, D.A., Bolivar, V., Brockmann, G.A*., et al.* (2003). The nature and identification of quantitative trait loci: a community's view. Nat Rev Genet *4*, 911-916.

Acevedo-Arozena, A., Wells, S., Potter, P., Kelly, M., Cox, R.D., and Brown, S.D. (2008). ENU mutagenesis, a way forward to understand gene function. Annu Rev Genomics Hum Genet *9*, 49-69.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881-888.

Andreux, P.A., Williams, E.G., Koutnikova, H., Houtkooper, R.H., Champy, M.F., Henry, H., Schoonjans, K., Williams, R.W., and Auwerx, J. (2012). Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. Cell *150*, 1287-1299.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T*., et al.* (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods *13*, 229-232.

Arnaud, L., Saison, C., Helias, V., Lucien, N., Steschenko, D., Giarratana, M.C., Prehu, C., Foliguet, B., Montout, L., de Brevern, A.G*., et al.* (2010). A dominant mutation in the gene encoding the erythroid transcription factor KLF1 causes a congenital dyserythropoietic anemia. Am J Hum Genet *87*, 721-727.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074-1077.

Baker, C.L., Kajita, S., Walker, M., Saxl, R.L., Raghupathy, N., Choi, K., Petkov, P.M., and Paigen, K. (2015). PRDM9 drives evolutionary erosion of hotspots in Mus musculus through haplotype-specific initiation of meiotic recombination. PLoS Genet *11*, e1004916.

Baker, C.L., Walker, M., Arat, S., Ananda, G., Petkova, P., Powers, N.R., Tian, H., Spruce, C., Ji, B., Rausch, D*., et al.* (2019). Tissue-Specific Trans Regulation of the Mouse Epigenome. Genetics *211*, 831-845.

Baker, C.L., Walker, M., Kajita, S., Petkov, P.M., and Paigen, K. (2014). PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Res *24*, 724-732.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature *533*, 452-454.

Baliga, N.S., Bjorkegren, J.L., Boeke, J.D., Boutros, M., Crawford, N.P., Dudley, A.M., Farber, C.R., Jones, A., Levey, A.I., Lusis, A.J*., et al.* (2017). The State of Systems Genetics in 2017. Cell Syst *4*, 7-15.

Ball, R.L., Fujiwara, Y., Sun, F., Hu, J., Hibbs, M.A., Handel, M.A., and Carter, G.W. (2016). Regulatory complexity revealed by integrated cytological and RNA-seq analyses of meiotic substages in mouse spermatocytes. BMC Genomics *17*, 628.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299-308.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res *21*, 381-395.
Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.
Basu, S., Kumbier, K., Brown, J.B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci U S A *115*, 1943-1948.
Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. Science *347*, 664-667.
Baudat, F., Imai, Y., and de Massy, B. (2013). Meiotic recombination in mammals: localization and regulation. Nat Rev Genet *14*, 794-806.
Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.
Berry, R.J., and Bronson, F.H. (1992). Life history and bioeconomy of the house mouse. Biol Rev Camb Philos Soc *67*, 519-550.
Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T., and Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol *2*, 152-163.
Birling, M.C., Herault, Y., and Pavlovic, G. (2017). Modeling human disease in rodents by CRISPR/Cas9 genome editing. Mamm Genome *28*, 291-301.
Blat, Y., and Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell *98*, 249-259.
Bogue, M.A., Churchill, G.A., and Chesler, E.J. (2015). Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. Mamm Genome *26*, 511-520.
Bolcun-Filas, E., and Handel, M.A. (2018). Meiosis: the chromosomal foundation of reproduction. Biol Reprod *99*, 112-126.
Breitling, R. (2010). What is systems biology? Front Physiol *1*, 9.
Brekke, T.D., Steele, K.A., and Mulley, J.F. (2018). Inbred or Outbred? Genetic Diversity in Laboratory Rodent Colonies. G3 (Bethesda) *8*, 679-686.
Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. Nature *436*, 701-703.
Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D., and Petukhova, G.V. (2012). Genetic recombination is directed away from functional genomic elements in mice. Nature *485*, 642-645.
Brodie, A., Azaria, J.R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? Nucleic Acids Res *44*, 6046-6054.
Brown, R.S., Sander, C., and Argos, P. (1985). The primary structure of transcription factor TFIIIA has 12 consecutive repeats. FEBS Lett *186*, 271-274.
Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of

open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213-1218.
Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486-490.
Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327-339.
Capell, B.C., and Berger, S.L. (2013). Genome-wide epigenetics. J Invest Dermatol *133*, e9.
Carballo, E., Lai, W.S., and Blackshear, P.J. (1998). Feedback inhibition of macrophage tumor necrosis factor-alpha production by tristetraprolin. Science *281*, 1001-1005.
Casellas, J. (2011). Inbred mouse strains and genetic stability: a review. Animal *5*, 1-7.
Caspary, T. (2010). Phenotype-driven mouse ENU mutagenesis screens. Methods Enzymol *477*, 313-327.
Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., and Raschella, G. (2017). Zinc-finger proteins in health and disease. Cell Death Discov *3*, 17071.
Catarino, R.R., and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev *32*, 202-223.
Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell *157*, 77-94.
Challen, G.A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J.S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y.*, et al.* (2011). Dnmt3a is essential for hematopoietic stem cell differentiation. Nat Genet *44*, 23-31.
Chen, T., and Dent, S.Y. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat Rev Genet *15*, 93-106.
Chen, X., Miragaia, R.J., Natarajan, K.N., and Teichmann, S.A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. Nat Commun *9*, 5345.
Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature *534*, 500-505.
Choo, Y., Sanchez-Garcia, I., and Klug, A. (1994). In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. Nature *372*, 642-645.
Chuang, H.Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. Annu Rev Cell Dev Biol *26*, 721-744.
Churchill, G.A., Airey, D.C., Allayee, H., Angel, J.M., Attie, A.D., Beatty, J., Beavis, W.D., Belknap, J.K., Bennett, B., Berrettini, W.*, et al.* (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet *36*, 1133-1137.
Churchill, G.A., Gatti, D.M., Munger, S.C., and Svenson, K.L. (2012). The Diversity Outbred mouse population. Mamm Genome *23*, 713-718.

Cleary, M.D., Meiering, C.D., Jan, E., Guymon, R., and Boothroyd, J.C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. Nat Biotechnol *23*, 232-237.

Coate, J.E., and Doyle, J.J. (2010). Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol Evol *2*, 534-546.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A.*, et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nat Rev Genet *10*, 184-194.

da Cruz, I., Rodriguez-Casuriaga, R., Santinaque, F.F., Farias, J., Curti, G., Capoano, C.A., Folle, G.A., Benavente, R., Sotelo-Silveira, J.R., and Geisinger, A. (2016). Transcriptome analysis of highly purified mouse spermatogenic cell populations: gene expression signatures switch from meiotic-to postmeiotic-related processes at pachytene stage. BMC Genomics *17*, 294.

Dao, L.T.M., Galindo-Albarran, A.O., Castro-Mondragon, J.A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T.*, et al.* (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Genet *49*, 1073-1081.

Davis, J.C., Snyder, E.M., Hogarth, C.A., Small, C., and Griswold, M.D. (2013). Induction of spermatogenic synchrony by retinoic acid in neonatal mice. Spermatogenesis *3*, e23180.

Daxinger, L., Harten, S.K., Oey, H., Epp, T., Isbel, L., Huang, E., Whitelaw, N., Apedaile, A., Sorolla, A., Yong, J.*, et al.* (2013). An ENU mutagenesis screen identifies novel and known genes involved in epigenetic processes in the mouse. Genome Biol *14*, R96.

de Castro-Catala, M., Mora-Solano, A., Kwapil, T.R., Cristobal-Narvaez, P., Sheinbaum, T., Racioppi, A., Barrantes-Vidal, N., and Rosa, A. (2017). The genome-wide associated candidate gene ZNF804A and psychosis-proneness: Evidence of sex-modulated association. PLoS One *12*, e0185072.

de Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499-506.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E.*, et al.* (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390-394.

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat Methods *3*, 17-21.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell *2*, 819-830.

Dikstein, R. (2011). The unexpected traits associated with core promoter elements. Transcription *2*, 201-206.
Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Mol Cell *62*, 668-680.
Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.
Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A.T., Chang, H.Y., Wang, Y., and Wong, W.H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci U S A *115*, 7723-7728.
Eddy, E.M. (1998). Regulation of gene expression during spermatogenesis. Semin Cell Dev Biol *9*, 451-457.
Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nat Rev Genet *2*, 919-929.
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.
Elton, L., Carpentier, I., Verhelst, K., Staal, J., and Beyaert, R. (2015). The multifaceted role of the E3 ubiquitin ligase HOIL-1: beyond linear ubiquitination. Immunol Rev *266*, 208-221.
Endrullat, C., Glokler, J., Franke, P., and Frohme, M. (2016). Standardization and quality management in next-generation sequencing. Appl Transl Genom *10*, 2-9.
Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods *9*, 215-216.
Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc *12*, 2478-2492.
Fallahi, M., Getun, I.V., Wu, Z.K., and Bois, P.R. (2010). A Global Expression Switch Marks Pachytene Initiation during Mouse Male Meiosis. Genes (Basel) *1*, 469-483.
Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci U S A *115*, 2628-2631.
Fedotova, A.A., Bonchuk, A.N., Mogila, V.A., and Georgiev, P.G. (2017). C2H2 Zinc Finger Proteins: The Largest but Poorly Explored Family of Higher Eukaryotic Transcription Factors. Acta Naturae *9*, 47-58.
Fine, A.D., Ball, R.L., Fujiwara, Y., Handel, M.A., and Carter, G.W. (2019). Uncoupling of transcriptomic and cytological differentiation in mouse spermatocytes with impaired meiosis. Mol Biol Cell *30*, 717-728.
Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2014). Ensembl 2014. Nucleic Acids Res *42*, D749-755.
Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A *89*, 1827-1831.

Fu, M., and Blackshear, P.J. (2017). RNA-binding proteins in immune regulation: a focus on CCCH zinc finger proteins. Nat Rev Immunol *17*, 130-143.
Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A.*, et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res *39*, D876-882.
Gaffney, D.J. (2013). Global properties and functional complexity of human gene regulatory variation. PLoS Genet *9*, e1003501.
Gan, H., Cai, T., Lin, X., Wu, Y., Wang, X., Yang, F., and Han, C. (2013). Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. Mol Cell Proteomics *12*, 1144-1157.
Gaysinskaya, V., Soh, I.Y., van der Heijden, G.W., and Bortvin, A. (2014). Optimized flow cytometry isolation of murine spermatocytes. Cytometry A *85*, 556-565.
Gibson, T.J., Postma, J.P., Brown, R.S., and Argos, P. (1988). A model for the tertiary structure of the 28 residue DNA-binding motif ('zinc finger') common to many eukaryotic transcriptional regulatory proteins. Protein Eng *2*, 209-218.
Goncalves, E., Fragoulis, A., Garcia-Alonso, L., Cramer, T., Saez-Rodriguez, J., and Beltrao, P. (2017). Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. Cell Syst *5*, 386-398 e384.
Green, C.D., Ma, Q., Manske, G.L., Shami, A.N., Zheng, X., Marini, S., Moritz, L., Sultan, C., Gurczynski, S.J., Moore, B.B.*, et al.* (2018). A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. Dev Cell *46*, 651-667 e610.
Grewal, S.I., and Jia, S. (2007). Heterochromatin revisited. Nat Rev Genet *8*, 35-46.
Gupta, A., Christensen, R.G., Bell, H.A., Goodwin, M., Patel, R.Y., Pandey, M., Enuameh, M.S., Rayla, A.L., Zhu, C., Thibodeau-Beganny, S.*, et al.* (2014). An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins. Nucleic Acids Res *42*, 4800-4812.
Gustafsson Sheppard, N., Heldring, N., and Dahlman-Wright, K. (2012). Estrogen receptor-alpha, RBCK1, and protein kinase C beta 1 cooperate to regulate estrogen receptor-alpha gene expression. J Mol Endocrinol *49*, 277-287.
Hamid, J.S., Hu, P., Roslin, N.M., Ling, V., Greenwood, C.M., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics *2009*.
Hammoud, S.S., Low, D.H., Yi, C., Carrell, D.T., Guccione, E., and Cairns, B.R. (2014). Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell *15*, 239-253.
Handel, M.A., and Schimenti, J.C. (2010). Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. Nat Rev Genet *11*, 124-136.
Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C.*, et al.* (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res *32*, D258-261.
Harrison, P.W., Wright, A.E., and Mank, J.E. (2012). The evolution of gene expression and the transcriptome-phenotype relationship. Semin Cell Dev Biol *23*, 222-229.
Hawkins, R.D., Hon, G.C., and Ren, B. (2010). Next-generation genomics: an integrative approach. Nat Rev Genet *11*, 476-486.

Hayashi, K., and Matsui, Y. (2006). Meisetz, a novel histone tri-methyltransferase, regulates meiosis-specific epigenesis. Cell Cycle *5*, 615-620.
Hayashi, K., Yoshida, K., and Matsui, Y. (2005). A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature *438*, 374-378.
Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1-8.
Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A*., et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311-318.
Henrichsen, C.N., Vinckenbosch, N., Zollner, S., Chaignat, E., Pradervand, S., Schutz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. Nat Genet *41*, 424-429.
Hermann, B.P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K.N., Chen, I.C., Gildersleeve, H., Lehle, J.D., Mayo, M., Westernstroer, B*., et al.* (2018). The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. Cell Rep *25*, 1650-1667 e1658.
Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J*., et al.* (2007). Genome-wide in situ exon capture for selective resequencing. Nat Genet *39*, 1522-1527.
Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods *9*, 473-476.
Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., and Fedoroff, N.V. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci U S A *97*, 8409-8414.
Hrabe de Angelis, M., and Balling, R. (1998). Large scale ENU screens in the mouse: genetics meets genomics. Mutat Res *400*, 25-32.
Huisinga, K.L., Brower-Toland, B., and Elgin, S.C. (2006). The contradictory definitions of heterochromatin: transcription and silencing. Chromosoma *115*, 110-122.
Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med *50*, 96.
Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet *2*, 343-372.
Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218-223.
Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. Proc Natl Acad Sci U S A *94*, 5617-5621.
Isalan, M., Klug, A., and Choo, Y. (1998). Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. Biochemistry *37*, 12026-12033.
Jamieson, A.C., Miller, J.C., and Pabo, C.O. (2003). Drug discovery with engineered zinc-finger proteins. Nat Rev Drug Discov *2*, 361-368.
Janky, R., Verfaillie, A., Imrichova, H., Van de Sande, B., Standaert, L., Christiaens, V., Hulselmans, G., Herten, K., Naval Sanchez, M., Potier, D*., et al.* (2014).

iRegulon: from a gene list to a gene regulatory network using large motif and track collections. PLoS Comput Biol *10*, e1003731.
Jansen, M., de Moor, C.H., Sussenbach, J.S., and van den Brande, J.L. (1995). Translational control of gene expression. Pediatr Res *37*, 681-686.
Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007a). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.
Johnson, J.A., Lu, Y.Y., Van Deventer, J.A., and Tirrell, D.A. (2010). Residue-specific incorporation of non-canonical amino acids into proteins: recent developments and applications. Curr Opin Chem Biol *14*, 774-780.
Johnson, W.E., Li, C., and Rabinovic, A. (2007b). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.
Kaplan, T., Friedman, N., and Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol *1*, e1.
Karczewski, K.J., Dudley, J.T., Kukurba, K.R., Chen, R., Butte, A.J., Montgomery, S.B., and Snyder, M. (2013). Systematic functional regulatory assessment of disease-associated variants. Proc Natl Acad Sci U S A *110*, 9607-9612.
Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A *107*, 2926-2931.
Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V.*, et al.* (2013). Extensive variation in chromatin states across humans. Science *342*, 750-752.
Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J.*, et al.* (2014). Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A *111*, 6131-6138.
Kim, J.B., Sebastiano, V., Wu, G., Arauzo-Bravo, M.J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrich, M., van den Boom, D.*, et al.* (2009). Oct4-induced pluripotency in adult neural stem cells. Cell *136*, 411-419.
Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S.*, et al.* (2014). A draft map of the human proteome. Nature *509*, 575-581.
Kim, Y.S., Lewandoski, M., Perantoni, A.O., Kurebayashi, S., Nakanishi, G., and Jetten, A.M. (2002). Identification of Glis1, a novel Gli-related, Kruppel-like zinc finger protein containing transactivation and repressor functions. J Biol Chem *277*, 30901-30913.
Kirschner, M.W. (2005). The meaning of systems biology. Cell *121*, 503-504.
Kitano, H. (2002a). Computational systems biology. Nature *420*, 206-210.
Kitano, H. (2002b). Systems biology: a brief overview. Science *295*, 1662-1664.
Klug, A. (2010). The discovery of zinc fingers and their applications in gene regulation and genome manipulation. Annu Rev Biochem *79*, 213-231.
Korkuc, P., Schippers, J.H., and Walther, D. (2014). Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. Plant Physiol *164*, 181-200.
Krämer A, G.J., Pollard J Jr, Tugendreich S (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics *30(4)*, 523-530.

Kuhn, M., Weston, c.f.S., Wing, J., Forester, J., and Thaler, T. (2016). contrast: A Collection of Contrast Methods. R package version 021.
Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet *48*, 206-213.
La Salle, S., Sun, F., and Handel, M.A. (2009). Isolation and short-term culture of mouse spermatocytes for analysis of meiosis. Methods Mol Biol *558*, 279-297.
Ladomery, M., and Dellaire, G. (2002). Multifunctional zinc finger proteins in development and disease. Ann Hum Genet *66*, 331-342.
LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res *37*, 4181-4193.
Laity, J.H., Lee, B.M., and Wright, P.E. (2001). Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol *11*, 39-46.
Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *175*, 598-599.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.
Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435-1439.
Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237-1251.
Leek, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., and Storey, J.D. (2017). sva: Surrogate Variable Analysis. R package version 3260.
Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882-883.
Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
Li, X.Z., Roy, C.K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B.W., Xu, J., Moore, M.J., Schimenti, J.C., Weng, Z.*, et al.* (2013). An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell *50*, 67-81.
Lifton, R.P., Goldberg, M.L., Karp, R.W., and Hogness, D.S. (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harb Symp Quant Biol *42 Pt 2*, 1047-1051.
Likic, V.A., McConville, M.J., Lithgow, T., and Bacic, A. (2010). Systems biology: the next frontier for bioinformatics. Adv Bioinformatics, 268925.
Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M.*, et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.
Lithgow, G.J., Driscoll, M., and Phillips, P. (2017). A long journey to reproducible results. Nature *548*, 387-388.

Liu, E.T., Bolcun-Filas, E., Grass, D.S., Lutz, C., Murray, S., Shultz, L., and Rosenthal, N. (2017). Of mice and CRISPR: The post-CRISPR future of the mouse as a model system for the human condition. EMBO Rep *18*, 187-193.

Liu, G., Gramling, S., Munoz, D., Cheng, D., Azad, A.K., Mirshams, M., Chen, Z., Xu, W., Roberts, H., Shepherd, F.A.*, et al.* (2011). Two novel BRM insertion promoter sequence variants are associated with loss of BRM expression and lung cancer risk. Oncogene *30*, 3295-3304.

Lomniczi, A., Wright, H., Castellano, J.M., Matagne, V., Toro, C.A., Ramaswamy, S., Plant, T.M., and Ojeda, S.R. (2015). Epigenetic regulation of puberty via Zinc finger protein-mediated transcriptional repression. Nat Commun *6*, 10195.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. Genome Res *15*, 945-953.

Lupo, A., Cesaro, E., Montano, G., Zurlo, D., Izzo, P., and Costanzo, P. (2013). KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. Curr Genomics *14*, 268-278.

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.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). cluster: Cluster Analysis Basics and Extensions. R Package version 208.

Mahadevaiah, S.K., Costa, Y., and Turner, J.M. (2009). Using RNA FISH to study gene expression during mammalian meiosis. Methods Mol Biol *558*, 433-444.

Mahadevaiah, S.K., Turner, J.M., Baudat, F., Rogakou, E.P., de Boer, P., Blanco-Rodriguez, J., Jasin, M., Keeney, S., Bonner, W.M., and Burgoyne, P.S. (2001). Recombinational DNA double-strand breaks in mice precede synapsis. Nat Genet *27*, 271-276.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823-826.

Mammana, A., and Chung, H.R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. Genome Biol *16*, 151.

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med *363*, 166-176.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet *9*, 387-402.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29-59.

Mastrangelo, T., Modena, P., Tornielli, S., Bullrich, F., Testi, M.A., Mezzelani, A., Radice, P., Azzarelli, A., Pilotti, S., Croce, C.M.*, et al.* (2000). A novel zinc finger gene is fused to EWS in small round cell tumor. Oncogene *19*, 3799-3804.

Matthews, B.J., and Waxman, D.J. (2018). Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. Elife *7*.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.

Mazo, A., Hodgson, J.W., Petruk, S., Sedkov, Y., and Brock, H.W. (2007). Transcriptional interference: an unexpected layer of complexity in gene regulation. J Cell Sci *120*, 2755-2761.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Meehan, T.F., Conte, N., West, D.B., Jacobsen, J.O., Mason, J., Warren, J., Chen, C.K., Tudose, I., Relac, M., Matthews, P.*, et al.* (2017). Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. Nat Genet *49*, 1231-1238.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Miller, J., McLachlan, A.D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J *4*, 1609-1614.

Mills, M.C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. Commun Biol *2*, 9.

Montagutelli, X. (2000). Effect of the genetic background on the phenotype of mouse mutations. J Am Soc Nephrol *11 Suppl 16*, S101-105.

Moore, M., Klug, A., and Choo, Y. (2001). Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. Proc Natl Acad Sci U S A *98*, 1437-1441.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Mouse Genome Sequencing, C., Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344-1349.

Nakamura, M., Runko, A.P., and Sagerstrom, C.G. (2004). A novel subfamily of zinc finger genes involved in embryonic development. J Cell Biochem *93*, 887-895.

Namekawa, S.H., Park, P.J., Zhang, L.F., Shima, J.E., McCarrey, J.R., Griswold, M.D., and Lee, J.T. (2006). Postmeiotic sex chromatin in the male germline of mice. Curr Biol *16*, 660-667.

Neufeld, D.S. (1997). Isolation of rat liver hepatocytes. Methods Mol Biol *75*, 145-151.

Ngo, V., Chen, Z., Zhang, K., Whitaker, J.W., Wang, M., and Wang, W. (2019). Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. Proc Natl Acad Sci U S A *116*, 3668-3677.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci *368*, 20120362.
Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet *6*, e1000888.
Nishioka, Y. (1995). The origin of common laboratory mice. Genome *38*, 1-7.
Nishizaki, S.S., and Boyle, A.P. (2017). Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends Genet *33*, 34-45.
Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J.*, et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.
Ovcharenko, I., Stubbs, L., and Loots, G.G. (2004). Interpreting mammalian evolution using Fugu genome comparisons. Genomics *84*, 890-895.
Pan, G.J., Chang, Z.Y., Scholer, H.R., and Pei, D. (2002). Stem cell pluripotency and transcription factor Oct4. Cell Res *12*, 321-329.
Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669-680.
Parker, S.C., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Program, N.C.S.*, et al.* (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A *110*, 17921-17926.
Parvanov, E.D., Petkov, P.M., and Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. Science *327*, 835.
Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.
Peirce, J.L., Lu, L., Gu, J., Silver, L.M., and Williams, R.W. (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet *5*, 7.
Perez, E.E., Wang, J., Miller, J.C., Jouvenot, Y., Kim, K.A., Liu, O., Wang, N., Lee, G., Bartsevich, V.V., Lee, Y.L.*, et al.* (2008). Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. Nat Biotechnol *26*, 808-816.
Perlman, R.L. (2016). Mouse models of human disease: An evolutionary perspective. Evol Med Public Health *2016*, 170-176.
Persikov, A.V., Osada, R., and Singh, M. (2009). Predicting DNA recognition by Cys2His2 zinc finger proteins. Bioinformatics *25*, 22-29.
Persikov, A.V., and Singh, M. (2014). De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res *42*, 97-108.
Petersen, A., Alvarez, C., DeClaire, S., and Tintle, N.L. (2013). Assessing methods for assigning SNPs to genes in gene-based tests of association using common variants. PLoS One *8*, e62161.
Plessy, C., Dickmeis, T., Chalmel, F., and Strahle, U. (2005). Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. Trends Genet *21*, 207-210.

Pollex, T., and Furlong, E.E.M. (2017). Correlation Does Not Imply Causation: Histone Methyltransferases, but Not Histone Methylation, SET the Stage for Enhancer Activation. Mol Cell *66*, 439-441.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M.*, et al.* (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet *41*, 882-884.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629-641.

Quang, D.X., Erdos, M.R., Parker, S.C.J., and Collins, F.S. (2015). Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. Epigenetics Chromatin *8*, 23.

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N.*, et al.* (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol *29*, 436-442.

Raghupathy, N., Choi, K., Vincent, M.J., Beane, G.L., Sheppard, K.S., Munger, S.C., Korstanje, R., Pardo-Manual de Villena, F., and Churchill, G.A. (2018). Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. Bioinformatics *34*, 2177-2184.

Ramskold, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol *5*, e1000598.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E.*, et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Reynolds, L., Ullman, C., Moore, M., Isalan, M., West, M.J., Clapham, P., Klug, A., and Choo, Y. (2003). Repression of the HIV-1 5' LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors. Proc Natl Acad Sci U S A *100*, 1615-1620.

Rinchik, E.M. (1987). Molecular analysis of heritable mouse mutations. Environ Health Perspect *74*, 41-48.

Rodgers, K., and McVey, M. (2016). Error-Prone Repair of DNA Double-Strand Breaks. J Cell Physiol *231*, 15-24.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R.*, et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389-393.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797-803.

Salinger, A.P., and Justice, M.J. (2008). Mouse Mutagenesis Using N-Ethyl-N-Nitrosourea (ENU). CSH Protoc *2008*, pdb prot4985.

Sandoval, J., Rodriguez, J.L., Tur, G., Serviddio, G., Pereda, J., Boukaba, A., Sastre, J., Torres, L., Franco, L., and Lopez-Rodas, G. (2004). RNAPol-ChIP: a novel application of chromatin immunoprecipitation to the analysis of real-time gene transcription. Nucleic Acids Res *32*, e88.

Sanz, E., Yang, L., Su, T., Morris, D.R., McKnight, G.S., and Amieux, P.S. (2009). Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. Proc Natl Acad Sci U S A *106*, 13939-13944.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schultz, N., Hamra, F.K., and Garbers, D.L. (2003). A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proc Natl Acad Sci U S A *100*, 12201-12206.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337-342.

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.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V.*, et al.* (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345-353.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol *26*, 1135-1145.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet *15*, 272-286.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G.*, et al.* (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res *43*, W589-598.

Smith, A.D., Sumazin, P., Xuan, Z., and Zhang, M.Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. Proc Natl Acad Sci U S A *103*, 6275-6280.

Solomon, M.J., Larsen, P.L., and Varshavsky, A. (1988). Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell *53*, 937-947.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthes, P., Kokkinaki, M., Nef, S., Gnirke, A.*, et al.* (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep *3*, 2179-2190.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613-626.

Srivastava, A., Morgan, A.P., Najarian, M.L., Sarsani, V.K., Sigmon, J.S., Shorter, J.R., Kashfeen, A., McMullan, R.C., Williams, L.H., Giusti-Rodriguez, P.*, et al.* (2017). Genomes of the Mouse Collaborative Cross. Genetics *206*, 537-556.

Stadhouders, R., van den Heuvel, A., Kolovos, P., Jorna, R., Leslie, K., Grosveld, F., and Soler, E. (2012). Transcription regulation by distal enhancers: who's in the loop? Transcription *3*, 181-186.

Stark, R., and Brown, G.D. (2011). DiffBind: differential binding analysis of ChIP-seq peak data. Bioconductor.

Steen, H., and Pandey, A. (2002). Proteomics goes quantitative: measuring protein abundance. Trends Biotechnol *20*, 361-364.

Stitzel, M.L., Sethupathy, P., Pearson, D.S., Chines, P.S., Song, L., Erdos, M.R., Welch, R., Parker, S.C., Boyle, A.P., Scott, L.J.*, et al.* (2010). Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. Cell Metab *12*, 443-455.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat Methods *14*, 865-868.

Strambio-De-Castillia, C., Niepel, M., and Rout, M.P. (2010). The nuclear pore complex: bridging nuclear transport and gene regulation. Nat Rev Mol Cell Biol *11*, 490-501.

Sul, J.Y., Wu, C.W., Zeng, F., Jochems, J., Lee, M.T., Kim, T.K., Peritz, T., Buckley, P., Cappelleri, D.J., Maronski, M.*, et al.* (2009). Transcriptome transfer produces a predictable cellular phenotype. Proc Natl Acad Sci U S A *106*, 7624-7629.

Sun, F., Fujiwara, Y., Reinholdt, L.G., Hu, J., Saxl, R.L., Baker, C.L., Petkov, P.M., Paigen, K., and Handel, M.A. (2015). Nuclear localization of PRDM9 and its role in meiotic chromatin modifications and homologous synapsis. Chromosoma *124*, 397-415.

Svenson, K.L., Gatti, D.M., Valdar, W., Welsh, C.E., Cheng, R., Chesler, E.J., Palmer, A.A., McMillan, L., and Churchill, G.A. (2012). High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics *190*, 437-447.

Takada, T., Ebata, T., Noguchi, H., Keane, T.M., Adams, D.J., Narita, T., Shin, I.T., Fujisawa, H., Toyoda, A., Abe, K.*, et al.* (2013). The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. Genome Res *23*, 1329-1338.

Taylor, B.A., Wnek, C., Kotlus, B.S., Roemer, N., MacTaggart, T., and Phillips, S.J. (1999). Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. Mamm Genome *10*, 335-348.

Thibault-Sennett, S., Yu, Q., Smagulova, F., Cloutier, J., Brick, K., Camerini-Otero, R.D., and Petukhova, G.V. (2018). Interrogating the Functions of PRDM9 Domains in Meiosis. Genetics *209*, 475-487.

Thivierge, C., Kurbegovic, A., Couillard, M., Guillaume, R., Cote, O., and Trudel, M. (2006). Overexpression of PKD1 causes polycystic kidney disease. Mol Cell Biol *26*, 1538-1548.

Threadgill, D.W., and Churchill, G.A. (2012). Ten years of the Collaborative Cross. Genetics *190*, 291-294.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B.*, et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Toth, L.A., Trammell, R.A., and Williams, R.W. (2014). Mapping complex traits using families of recombinant inbred strains: an overview and example of mapping susceptibility to Candida albicans induced illness phenotypes. Pathog Dis *71*, 234-248.

Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet *13*, 36-46.

Tubbs, A., and Nussenzweig, A. (2017). Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell *168*, 644-656.

Ukmar, G., Melloni, G.E.M., Raddrizzani, L., Rossi, P., Di Bella, S., Pirchio, M.R., Vescovi, M., Leone, A., Callari, M., Cesarini, M.*, et al.* (2018). PATRI, a Genomics Data Integration Tool for Biomarker Discovery. Biomed Res Int *2018*, 2012078.

Urrutia, R. (2003). KRAB-containing zinc-finger repressor proteins. Genome Biol *4*, 231.

van Haaften-Visser, D.Y., Harakalova, M., Mocholi, E., van Montfrans, J.M., Elkadri, A., Rieter, E., Fiedler, K., van Hasselt, P.M., Triffaux, E.M.M., van Haelst, M.M.*, et al.* (2017). Ankyrin repeat and zinc-finger domain-containing 1 mutations are associated with infantile-onset inflammatory bowel disease. J Biol Chem *292*, 7904-7920.

van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nat Biotechnol *28*, 1089-1095.

Vanhooren, V., and Libert, C. (2013). The mouse as a model organism in aging research: usefulness, pitfalls and possibilities. Ageing Res Rev *12*, 8-21.

Veitia, R.A., and Birchler, J.A. (2015). Models of buffering of dosage imbalances in protein complexes. Biol Direct *10*, 42.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet *101*, 5-22.

Vrana, K.E., Churchill, M.E., Tullius, T.D., and Brown, D.D. (1988). Mapping functional regions of transcription factor TFIIIA. Mol Cell Biol *8*, 1684-1696.

Vu, V., Verster, A.J., Schertzberg, M., Chuluunbaatar, T., Spensley, M., Pajkic, D., Hart, G.T., Moffat, J., and Fraser, A.G. (2015). Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. Cell *162*, 391-402.

Walker, M., Billings, T., Baker, C.L., Powers, N., Tian, H., Saxl, R.L., Choi, K., Hibbs, M.A., Carter, G.W., Handel, M.A.*, et al.* (2015). Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage. Epigenetics Chromatin *8*, 31.

Wamstad, J.A., Wang, X., Demuren, O.O., and Boyer, L.A. (2014). Distal enhancers: new insights into heart development and disease. Trends Cell Biol *24*, 294-302.

Whitaker, J.W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. Nat Methods *12*, 265-272, 267 p following 272.

Witte, J.S. (2010). Genome-wide association studies and beyond. Annu Rev Public Health *31*, 9-20 24 p following 20.

Wu, Y., Williams, E.G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S.M., Argmann, C.A., Faridi, P., Wolski, W., Kutalik, Z.*, et al.* (2014). Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell *158*, 1415-1430.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res *42*, D98-103.

Xu, S., Grullon, S., Ge, K., and Peng, W. (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. Methods Mol Biol *1150*, 97-111.

Yoshiki, A., and Moriwaki, K. (2006). Mouse phenome research: implications of genetic background. ILAR J *47*, 94-102.

Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. EMBO Rep *16*, 1467-1481.

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.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhang, Y., Malone, J.H., Powell, S.K., Periwal, V., Spana, E., Macalpine, D.M., and Oliver, B. (2010). Expression in aneuploid Drosophila S2 cells. PLoS Biol *8*, e1000320.

Zheng, M., Tian, S.Z., Capurso, D., Kim, M., Maurya, R., Lee, B., Piecuch, E., Gong, L., Zhu, J.J., Li, Z.*, et al.* (2019). Multiplex chromatin interactions with single-molecule precision. Nature *566*, 558-562.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res *41*, W56-62.