UNDERSTANDING GERM CELL REGULATION USING BIOINFORMATIC APPROACHES

By

LEANNE WHITMORE

A dissertation submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Molecular Biosciences

DECEMBER 2015

© Copyright by LEANNE WHITMORE, 2015
All Rights Reserved
To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of LEANNE WHITMORE find it satisfactory and recommend that it be accepted.

____________________________________
Ping Ye, Ph.D., Co-Chair

____________________________________
Michael Griswold, Ph.D., Co-Chair

____________________________________
Jon Oatley, Ph.D.

____________________________________
Bala Krishnamoorthy, Ph.D.
ACKNOWLEDGEMENTS

I would like to take this opportunity to thank the people who have helped me complete my Doctorate degree. I would first like to thank Dr. Ping Ye for her unwavering guidance and support throughout my doctoral study. Additionally, I would like to thank previous laboratory member Dr. Debjit Ray for his helpful discussions and insights regarding bioinformatics.

Additionally, I would like to thank Dr. Wenfeng An for his direction and guidance specifically with the piRNA project presented in this dissertation. I sincerely thank other members of the An lab, Simon Newkirk and Fiorella Grandi for their help as well. I would also like to express my sincere gratitude to members of the Griswold lab, Dr. Cathryn Hogarth, Dr. Lisa Evans, Dr. Travis Kent, Kellie Agrimson, Estela Jauregui, Thanh Beedle, Debra Mitchell, and Traci Toppings, for their helpful suggestions, guidance and support. Also I would like to take the time to thank my committee members, Dr. Mike Griswold, Dr. Bala Krishnamoorthy and Dr. Jon Oatley for their valuable guidance and participation in my committee throughout my doctoral research.

Finally, I would like to thank my friends and family who have provided me with the emotional support needed for me to be successful at Washington State University.
UNDERSTANDING GERM CELL REGULATION USING BIOINFORMATIC APPROACHES

Abstract

by Leanne Whitmore, Ph.D.
Washington State University
December 2015

Co-Chairs: Ping Ye and Michael D. Griswold

Bioinformatics is a powerful tool which has been utilized to study diseases (i.e. cancer) and complex biological systems including the immune system, neural interactions in the brain and testis biology. Only recently has bioinformatics been used to study the intricacies of spermatogenesis, the differentiation and development of spermatogonia into spermatozoa. Successful spermatogenesis is essential for male reproduction and understanding regulatory mechanisms necessary for ensuring production of sperm is vital for the development of male infertility treatments or reversible contraceptive options. This dissertation focuses on two regulatory mechanisms necessary for spermatogenesis in mice: 1. metabolic activity of spermatogonia and 2. piRNA control of long interspersed element-1 (L1) retrotransposon evolution.
1. The bioinformatics approach flux balance analysis (FBA) is used to model metabolic activity of spermatogonia at 8 different time points during the initial differentiation of spermatogonia. Because it is known that central and vitamin A metabolic activities are crucial for successful spermatogenesis we analyze the activities of these metabolic pathways and uncover the roles they play in regulating spermatogenesis. Additionally, reaction and enzyme in-silico knockouts, within the vitamin A metabolic pathway are performed to elucidate their importance in regulating RA availability, which is crucial for differentiation of spermatogonia.

2. Bioinformatics analysis of piRNA sequencing and transcriptome datasets from mouse testes was performed to discover if piRNA repression of L1s is directing the single lineage evolution of L1s. PiRNAs suppress L1 activity at the transcriptional and post-transcriptional stages of retrotransposition and are essential for male fertility. PiRNA and L1 expression of L1 families were analyzed to determine if there was more activity (of piRNAs or L1 transcripts) to a particular age (young vs old) of L1 family. Additionally, expression was further analyzed to understand if piRNA or L1 activity was dependent on location of L1s relative to genes. Finally, where piRNAs target L1s for silencing (5’ or 3’ end) was examined to elucidate if piRNAs are exerting selective pressure on the 5’ end thereby causing L1s to acquire a new 5’ UTR as they evolve.

Overall, this thesis utilizes bioinformatics to understand spermatogenesis and advances our understanding of mammalian male reproduction.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>xii</td>
</tr>
</tbody>
</table>

## CHAPTER 1: Introduction

1. BIOINFORMATICS AND SPERMATOGENESIS ...........................................2
2. METABOLIC ACTIVITY OF GERM CELLS ..............................................7
   a. Role of Vitamin A Metabolism in Spermatogenesis .........................8
3. PIRNA CONTROL OF L1 EVOLUTION IN THE GERMLINE ..............................11
   a. Evolution of L1s .....................................................................11
   b. PIWI-Interacting RNAs (piRNAs) ...........................................12
4. CONCLUSIONS .................................................................................15
5. FIGURES ......................................................................................16
6. REFERENCES ..................................................................................20

## CHAPTER 2: Dissecting Germ Cell Metabolism through Network Modeling

1. TITLE PAGE ..................................................................................29
2. ABSTRACT ....................................................................................30
3. INTRODUCTION ...............................................................................31
4. RESULTS ......................................................................................34
CHAPTER 3: Role of Piwi – Interacting RNA in LINE-1 Evolutionary Dynamics

1. TITLE PAGE ................................................................. 80
2. ABSTRACT ................................................................. 81
3. INTRODUCTION .......................................................... 82
4. RESULTS ................................................................. 85
5. DISCUSSION ............................................................. 94
6. METHODS .............................................................. 98
7. REFERENCES .......................................................... 103
8. FIGURES ................................................................. 109
9. TABLES ................................................................. 125

CHAPTER 4: Conclusions and Future Directions

1. CONCLUSIONS AND FUTURE DIRECTIONS ..................... 128
   a. Chapter 2 ............................................................ 128
   b. Chapter 3 ............................................................ 132
   c. Final Conclusions .................................................. 135
2. REFERENCES ..................................................................................................................... 136
LIST OF TABLES

CHAPTER 2

1. Supplemental Table 1. Metabolites and genes known to be involved in vitamin A metabolism from the literature ............................................................... 75

CHAPTER 3

1. Supplemental Table 1: PiRNA Datasets ............................................................... 125

2. Supplemental Table 2: RNA Transcriptome Sequencing Datasets .............................. 126
LIST OF FIGURES

CHAPTER 1

1. Figure 1. Cross section of a mouse seminiferous tubule ........................................16
2. Figure 2. Vitamin A metabolism .................................................................17
3. Figure 3. Percentages of different types of transposons in the human genome.........18
4. Figure 4. Structure of a mouse L1 .................................................................19

CHAPTER 2

1. Figure 1. Vitamin A metabolic pathway .......................................................60
2. Figure 2. A mouse germ cell-specific metabolic network ..............................61
3. Figure 3. Flux of reactions that directly produce or degrade RA .................62
4. Figure 4. Identification of criteria to predict spermatogenic defects ..........63
5. Figure 5. Genes and reactions predicted to be required for spermatogenesis...64
6. Figure 6. Activities of other pathways in the germ cell-specific metabolic network ....65
7. Figure 7. The contribution of individual pathways to energy production and consumption ............................................................66
8. Figure 8. Flux of reactions that directly produce or degrade pyruvate ........67
9. Supplemental Figure 1. Types of germ cells enriched at different time points post-RA injection .................................................................68
10. Supplemental Figure 2. Flux of reactions that directly produce or degrade retinol and retinal .............................................................69
11. Supplemental Figure 3. Flux changes in the vitamin A pathway when Lipe is deleted in silico four hours after initiation of spermatogenesis via an RA injection ...........70
12. Supplemental Figure 4. Flux changes in the vitamin A pathway when the reaction retinol to CRBP-retinol is deleted in silico four hours after initiation of spermatogenesis via an RA injection..........................................................71

13. Supplemental Figure 5. Flux changes in the vitamin A pathway when the reaction CRBP-retinol to CRBP-retinal is deleted in silico four days after initiation of spermatogenesis via an RA injection..........................................................72

14. Supplemental Figure 6. Flux changes in the vitamin A pathway when the reaction CRBP-retinal to RA is deleted in silico four hours after initiation of spermatogenesis via an RA injection ...........................................................................73

15. Supplemental Figure 7. Flux changes in the vitamin A pathway when RA import into the cell is deleted in silico two days after initiation of spermatogenesis via an RA injection ...........................................................................74

CHAPTER 3

1. Figure 1: L1 Sub Classifications ..................................................................................109

2. Figure 2: PiRNA Expression Correlated with Age ......................................................110

3. Figure 3: PiRNA Expression to 5’ versus 3’ ends of L1s ...........................................112

4. Figure 4: Correlation between of piRNA and L1 Expression Values ..........................114

5. Figure 5: Genetic Locations of L1s .............................................................................115

6. Supplemental Figure 1: FL and Intact L1 Genomic Information ...............................117

7. Supplemental Figure 2: Length of Reads in piRNA Sequencing Datasets ................119

8. Supplemental Figure 3: Correlation between age and L1 copy .................................120

9. Supplemental Figure 4: Correlation between log (sense/antisense) value and age of L1 families .................................................................................................................121
10. Supplemental Figure 5: piRNA Mapping

11. Supplemental Figure 6: Location of Start Site for Genomic L1s in each of the L1 family
DEDICATION

This dissertation is dedicated to my family, Elizabeth, Paul, Nick and Laura Whitmore, who have supported me in everything I have done.
CHAPTER 1

INTRODUCTION
BIOINFORMATICS AND SPERMATOGENESIS

Bioinformatics is an essential and powerful method for studying intricate and complex biological systems. The field of bioinformatics involves the integration of mathematics, statistics, computer science, and large sets of biological data to allow for a comprehensive understanding of tissue/cellular subsystems such as metabolism, signal transduction and gene regulatory networks. Bioinformatics has been utilized with great success to study a variety of human diseases such as malaria, African sleeping sickness and cancer and provided researchers with potential drug targets (Bakker, et al., 1999; Bazzani, et al., 2012; Folger, et al., 2011; Haanstra and Bakker, 2015). Additionally, bioinformatics modeling and analysis of genome, transcriptome and small RNA sequencing data sets has been used to explore the intricacies of different mammalian biological processes including, but not limited to, the immune system, neural interactions in the brain and testis biology (de Rooij and van Beek, 2013; Dunster, 2015; Parikshak, et al., 2015; Ray, et al., 2014). While bioinformatics has been used extensively to study the immune system and brain (Dunster, 2015; Parikshak, et al., 2015), only a limited number of studies have utilized bioinformatics to study the development and differentiation of male germ cells.

Bioinformatics approaches have been utilized to study testis biology in rodents (de Rooij and van Beek, 2013; Ray, et al., 2014). De Rooij and van Beek, 2013 utilized bioinformatics to model the spermatogonial stem cell niche in rodent to further understand how stem cells give rise to spermatogonia. The adult mouse spermatogenic cycle was modeled by Ray, et. al. 2014 to elucidate the how spatial dynamics of germ cells regulate spermatogenesis, the process in which germ cells undergo differentiation and morphological changes to become sperm. Furthermore, bioinformatics analysis of transcriptome and small RNA sequencing datasets has led to the
discovery of important genes and small RNAs for sperm production (Endo, et al., 2015; Hogarth, et al., 2015; Li, et al., 2013; Modzelewski, et al., 2012; Soumillon, et al., 2013; Zamudio, et al., 2015). Because of the success of these studies which utilized bioinformatics to understand spermatogenesis, this dissertation utilizes bioinformatics to better understand regulation of male germ cell differentiation.

The testis is responsible for the production of billions of sperm over the course of a man’s reproductive life. Spermatogenesis occurs within the seminiferous tubules of the testis (Fig 1). The seminiferous tubule consists of three main cell types: 1) germ cells - which are cells differentiating into spermatozoa, 2) Sertoli cells - which provide support (i.e. structural and metabolic) to germ cells and 3) peritubular myoid cells - which are located outside of the basement membrane. Spermatogonial stem cells, a subset of spermatogonia, termed A single (As) spermatogonia, divide and differentiate into a progenitor population of spermatogonia, including As, A paired (pr) and A aligned (al) spermatogonia. Apr and Aal spermatogonia are progenitor spermatogonia joined together by an intracellular bridge. Apr and Aal spermatogonia, in response to retinoic acid (RA) a metabolic derivative of vitamin A (Mitranond, et al., 1979; Unni, et al., 1983), transition into A1 spermatogonia, in a process termed spermatogonia differentiation. A1 spermatogonia subsequently divide and differentiate into A2, A3, A4, Intermediate (In), and B spermatogonia. B spermatogonia then enter meiosis and become spermatocytes, where they replicate their genomic content and undergo two cell divisions, thereby transitioning from diploid to haploid cells. Finally, the post-meiotic germ cells go through spermiogenesis, which involves extensive morphological changes of round haploid spermatid cells into elongated spermatozoa. Spermatozoa are released into the lumen of the
seminiferous tubule (Fig 1) and transported through the rete testis to the rest of the male reproductive tract.

The initiation of spermatogenesis, or the differentiation of the first germ cells that will eventually become spermatozoa, is referred to as the first wave. This initial differentiation of germ cells is unique to subsequent waves. Precursor cells, prospermatogonia (gonocytes), simultaneously differentiate into A spermatogonia, including stem and progenitor cells, and A1 differentiating spermatogonia (Drumond, et al., 2011; Vergouwen, et al., 1991; Yang and Oatley, 2014). A1 spermatogonia then directly transition into A3 spermatogonia, skipping A2 division causing the first wave to be shorter than succeeding waves of spermatogenesis (Drumond, et al., 2011). Spermatogenesis initiates asynchronously along the length of seminiferous tubules ensuring that there is a continuous release of spermatozoa into the lumen (Snyder, et al., 2010). RA is necessary for the asynchronous onset of spermatogenesis (Snyder, et al., 2010). Snyder, et al., 2010 determined that RA signaling was present in patches along the seminiferous tubule, however, the molecular mechanisms controlling patchy RA availability is unknown. The success of the first wave is vital because it lays the foundation for ensuing rounds of spermatogenesis. The study of spermatogenesis, particularly with the aid of bioinformatics, is important for understanding not only the intricacies of this complex system, but also has the potential to redress serious male reproductive healthcare concerns, namely lack of infertility treatments and contraceptive options.

Approximately half of the 15% of couples that cannot conceive are due to male infertility (Anderson, et al., 2009). For 25% of male infertility cases the cause of infertility is unknown and, therefore, treatment is unavailable (Anderson, et al., 2009). In addition to the lack of infertility treatments, there is also a severe discrepancy in contraceptive options for men. While
women have a plethora of options, men have two: condoms and vasectomies. Currently, research is being conducted to develop reversible hormonal contraceptive options for men (Ilani, et al., 2012; Li, et al., 2015; Roth, et al., 2014) but more research is necessary to bring these options to market. Furthering our understanding of spermatogenesis is needed to address the gaping holes that exist in treatment for male infertility and development of contraceptives.

The continuous production of sperm is a complex process that involves numerous genes, signaling and metabolic pathways. Bioinformatics can help overcome the complexities of spermatogenesis in a way that cannot be executed in a traditional “wet laboratory” setting. This dissertation focuses on investigating the role of both metabolism and PIWI-interacting RNAs (piRNAs) during spermatogenesis using bioinformatics approaches.

The study of metabolism in a traditional wet laboratory can only focus on one pathway at a time. Radioactive tracers are used to understand the flow of metabolites through a pathway. Knockout studies can be compounded with radioactive tracers to further elucidate if a gene is important for a metabolic pathway. Additionally, kinetic studies on enzymes in a metabolic pathway can elucidate important rate limiting steps. These experiments are costly, time consuming and do not allow a comprehensive view of the entire metabolic activity for the organism, tissue or cell of interest. Using the computational approach flux balance analysis (FBA), reaction rates of every known reaction in an organism, tissue or cell can be investigated, which is impossible in a traditional laboratory environment. FBA has been utilized with great success to study the metabolism of many different organisms, tissues and cells including bacteria, yeast, plants, mice and human liver, macrophages, and neuronal cells (Adlakha, et al., 2015; Becker and Palsson, 2008; Bordbar, et al., 2011; Bordbar, et al., 2012; Chang, et al., 2010; Cheung, et al., 2015; Jerby, et al., 2010; Lewis, et al., 2010; Ray and Ye, 2013; Sigurdsson, et
In this dissertation (Chapter 2), FBA is used to investigate the importance of metabolism in germ cells during spermatogenesis.

The second aspect of spermatogenic regulation investigated in this dissertation is control of long interspersed element 1 transposition activity (L1s) by piRNAs in the germline. L1s are shown to evolve in mammals in a single lineage, meaning that one family of L1s is active at a given time (Adey, et al., 1994; Cabot, et al., 1997; Khan, et al., 2006; Smit, et al., 1995; Sookdeo, et al., 2013). L1s are speculated to evolve in this manner to prevent depletion of host factors necessary for retrotransposition (Khan, et al., 2006). However, it is unknown if there are other factors (i.e. host defense mechanisms) exerting control over the evolution of L1s. PiRNAs are the most adaptive response germ cells possess for controlling L1 retrotransposition and knockout studies have determined that piRNAs are essential in suppressing L1s and consequentially male fertility in mice (Carmell, et al., 2007; Deng and Lin, 2002; Kuramochi-Miyagawa, et al., 2004). However, these studies alone cannot elucidate piRNA control of L1 evolution. Bioinformatics analysis of non-coding RNA sequencing data has shed light shown the importance of small RNAs are generated in regulating gene and L1 expression in different biological processes including development and differentiation (Iwakawa and Tomari, 2015; Suh and Blelloch, 2011). Therefore, Chapter 3 uses bioinformatics to analyze published piRNA sequencing data to uncover how piRNA target L1s of different ages throughout germ cell development to determine the role they may exert in controlling L1 evolution.

The remainder of this chapter discusses published literature and gaps in current understanding about germ cell metabolism and piRNA regulation of spermatogenesis, as well as brief overview of what I have done to redress these gaps in knowledge.
METABOLISM IN GERM CELLS

The metabolic activity of spermatogonia during the initial wave of spermatogenesis has not been thoroughly investigated. However, there has been substantial research in the metabolic activity of Sertoli cells and metabolic requirements for spermatocytes and spermatids. Sertoli cells have high anaerobic glycolysis activity thereby ensuring spermatocytes and spermatids obtain the necessary metabolites, pyruvate and lactate, for growth and development (Alves, et al., 2013; Bajpai, et al., 1998; Grootegoed, et al., 1984; Meinhardt, et al., 1999; Nakamura, et al., 1984; Robinson and Fritz, 1981). Thus, glycolytic activity in spermatocytes and spermatids is minimal because all required metabolites are received from the Sertoli cell. Whether spermatogonia also receive important energy metabolites from Sertoli cells or produce them on their own has not been determined.

In murine embryonic, human embryonic, pluripotent and induced pluripotent stem cells, anaerobic glycolytic activity is predominant (Folmes, et al., 2012; Kondoh, et al., 2007; Prigione, et al., 2010; Zhang, et al., 2011). As cells differentiate lactate production decreases and TCA activity increases (Chung, et al., 2007; Facucho-Oliveira, et al., 2007; Prigione, et al., 2010; Zhang, et al., 2011). A switch between anaerobic and aerobic glycolysis appears to play a role in determining cell fate. Whether or not central metabolism may also be determining the fate of spermatogonia is unknown. One of the goals of Chapter 2 is to use FBA to investigate the activities of central metabolic pathways such as glycolysis and the TCA cycle to elucidate the role they exert in spermatogonial differentiation (Chapter 2).

In addition to central metabolism, other metabolic pathways such as cholesterol metabolism, steroid biosynthesis, and vitamin A metabolism are crucial for successful spermatogenesis (Casado, et al., 2012; Hogarth, et al., 2015; Hogarth and Griswold, 2010;
Hogarth and Griswold, 2013; McCarthy and Cerecedo, 1952; Mitranond, et al., 1979; Raverdeau, et al., 2012; Szende, et al., 1990; Tong, et al., 2013; Unni, et al., 1983). While each of these pathways are important and perturbations to any of these pathways cause defects in spermatogenesis, our research focused on vitamin A metabolism because a number of studies have shown it regulates and controls multiple processes during spermatogenesis including spermatogonial differentiation, meiosis and spermiogenesis (Kent and Griswold, 2014).

Role of Vitamin A Metabolism in Spermatogenesis

In 1952 McCarthy and Cerecedo discovered that vitamin A deficiency (VAD) caused infertility in male and female mice. Subsequent studies determined that VAD caused a block in the transition from undifferentiated to differentiating spermatogonia (spermatogonial differentiation) (Mitranond, et al., 1979; Unni, et al., 1983). The active derivative of vitamin A, retinoic acid (RA) activates expression of genes required for spermatogonia differentiation through the retinoic acid receptors (RARs) and retinoic X receptors (RXRs), reviewed by Rhinn and Dolle, 2012. Because RA is required for numerous spermatogenic processes, in addition to spermatogonia differentiation, the metabolic control of RA concentration within the testis is of interest.

The basic metabolism of vitamin A (retinol) to RA involves three key metabolic steps: 1) retinol is reversibly converted to retinaldehyde, 2) the irreversible conversion of retinaldehyde to RA and 3) degradation of RA which produce 4-oxo-RA and 4-OH-RA (Fig 2). The three main reactions are catalyzed by retinol dehydrogenases (RDHs), retinaldehyde dehydrogenases (RALDHs) and the cytochrome P450 family 26 (CYP26s) enzymes respectively. Outside of these three main reactions there are many other enzymatic reactions which contribute to RA availability (Fig 1 of Chapter 2). For example, LRAT converts retinol to retinyl-ester while
other enzymes LIPE, CES1, and CES3 catalyze the reverse reaction. Additionally, binding proteins (CRBP and CRABP) affect the availability of metabolites in the vitamin A pathway (de Bruijn, et al., 1994; Ghyselinck, et al., 1999; Lampron, et al., 1995; Quadro, et al., 1999). All of these enzymes, as well as others, may be regulating RA availability in the testis. Hogarth, et al. 2015 observed a peak of RA approximately every 8.6 days corresponding to times during spermatogenesis where RA is crucial for success (Hogarth, et al., 2015), but the regulating mechanism of this pulse remains elusive. The metabolic control of the RA pulse is of extreme interest because identification of key enzymes and reactions may unveil sources of idiopathic male infertility issues or be a target for male contraception.

Numerous cell types, reactions and enzymes are currently under investigation for their contribution to regulating the RA pulse. Genes in the vitamin A metabolic pathway expressed in Sertoli cells are essential for the first wave of spermatogenesis (Raverdeau, et al., 2012; Tong, et al., 2013). Conditional knockout studies in Sertoli cells of Rdh10 and Raldhs caused a delay and block respectively in spermatogenesis (Raverdeau, et al., 2012; Tong, et al., 2013). Additionally, when Rdh10 is conditionally knocked out in both Sertoli and germ cells, the phenotype is more severe. Because RA is essential for this transition, these results indicate that the initial pulse of RA (pulse during the first wave of spermatogenesis) is hindered, therefore preventing spermatogenesis. Furthermore, the conditional knockout of Cyp26b1 in both Sertoli and germ cells exhibits spermatogenic phenotypes including vacuoles and missing germ cells resulting in infertility (Hogarth, et al., 2015). Even though most evidence suggests that Sertoli cells are providing a majority of RA during the first spermatogenic wave, the increased severity of the spermatogenic defect when Rdh10 and Cyp26b1 are knocked out in both Sertoli and germ cells
indicates germ cells may contribute to the regulation of the initial RA pulse. (Hogarth, et al., 2015; Tong, et al., 2013)

Recent gene expression analysis of the testis has provided new gene candidates that may control the RA pulse. *Lrat, Dhrs4* (a retinol dehydrogenase), and *Crabp1* decreased in expression in the testis at times when RA concentrations were high (Hogarth, et al., 2015). All of the previously listed genes encode for enzymes that reduce RA availability suggesting they may regulate the RA pulse. Whether the activities of *Lrat, Dhrs4*, and *Crabp1* in germ cells are essential for the generation of RA pulse is unknown. In order to elucidate the unknown aspects of germ cell metabolism, we use the computational approach FBA.

Utilizing FBA to investigate the enzymes and reactions regulating the RA pulse enables reaction and enzyme contributions to RA production and degradation to be quantified. Furthermore, enzyme and reaction knockouts can be easily modeled and metabolic changes in response to perturbations can be quickly and efficiently analyzed. Finally, the manner in which the vitamin A metabolic pathway interacts with other metabolic pathways (*i.e.* through energy metabolites (ATP or NADH) or retinyl-esters) can be quantified. In Chapter 2 of this dissertation, spermatogonial germ cell metabolism will be investigated through the first wave of spermatogenesis to understand global metabolism and, specifically, spermatogonial contribution to the generation of the first pulse of RA.
PIRNA CONTROL OF L1 EVOLUTION IN THE GERMLINE

Transposable elements (TEs) are segments of DNA which can move and integrate into different locations in the genome. They were first discovered in Zea mays (maize) by Barbra McClintock in 1953 (McClintock, 1953). Since this discovery, transposons have been classified into two groups, DNA transposons and retrotransposons. Due to the completion of the human genome sequence and subsequent bioinformatics analyses it is known that DNA transposons comprise only about 3% of the human genome while retrotransposons constitute approximately 44% (Fig 3) (Lander, et al., 2001). Retrotransposons include TEs with long terminal repeats (LTRs), long interspersed element (LINEs), and short interspersed elements (SINEs). SINEs lack the necessary protein factors to move autonomously and instead rely on L1s and LTRs for movement (Levin and Moran, 2011). LINEs are the largest group of retrotransposons and consist of three different families: L1, L2 and L3. Of all retrotransposons, L1s are the only one known to be capable of autonomous transposition in the human genome (Hancks and Kazazian, 2012; Lander, et al., 2001; Ostertag and Kazazian, 2001). L1 activity in the genome has provided novel genetic and epigenetic variations over the course of mammalian evolution (Furano, 2000; Zamudio and Bourc'his, 2010). The manner in which L1s have evolved in the genome and the factors contributing to this evolution are of interest because it will allow a more complete understanding of L1 biology and mammalian evolution.

Evolution of L1s

L1s have been evolving in mammalian genomes since before the diversification of placental mammals (Furano, 2000). Interestingly, L1s evolve in a single lineage evolutionary pattern, meaning generally only one family of L1s is active at a given time. This evolution is observed in many species including rats, mice and humans and is unique to L1 transposons.
(Adey, et al., 1994; Cabot, et al., 1997; Furano, 2000; Khan, et al., 2006; Smit, et al., 1995; Sookdeo, et al., 2013). Additionally, during the evolution of L1s, when a new family of L1s emerges a novel 5’UTR promoter (Fig 4) sequence is acquired. Coexistence of L1 families is possible if the two families have different 5’UTRs thereby using different host transcription factors (Khan, et al., 2006). For reasons that are not understood, a new and more dominant family will eventually emerge and older families will cease to replicate. In this dissertation it is postulated, for the first time, defense mechanisms - specifically piRNAs - are exerting selective pressure on replicating L1s and a new family will arise.

**PIWI-Interacting RNAs (piRNAs)**

While there are numerous defense mechanisms acquired by cells in order to control L1 activity, piRNAs are one of the most adaptive defenses germ cells possess. PiRNAs are small RNAs approximately 24 – 30 base pairs in length and form ribonucleoprotein complexes with PIWI proteins MILI, MIWI2, and MIWI to repress L1 activity. PiRNAs repress L1 activity at two stages of retrotransposition 1) transcriptional and 2) post-transcriptional. During genome reprogramming, at approximately embryonic day 14.5 in the mouse, piRNAs repress transcription of L1s by directing methylation to L1 sequences (Carmell, et al., 2007; Molaro, et al., 2014). Additionally, throughout germ cell development, piRNAs can facilitate degradation of L1 transcripts. PiRNAs target the ribonucleoprotein complex to L1 transcripts which are degraded by the slicer activity of the PIWI proteins. Degraded L1 transcripts can facilitate degradation of other L1s by becoming piRNAs themselves, a process termed the ping pong cycle (Aravin, et al., 2007; Brennecke, et al., 2007). The ping-pong cycle is the reason piRNAs are so adaptive and efficient at repressing L1s. The ability to adjust and target L1s that are more active
is thought to exert selective pressure on L1s thereby forcing them to evolve in order to evade
detection by piRNA.

PIWI proteins are expressed in both male and female germ cells to protect cells from
transposon integration within numerous species such as *Drosophila* (flies), mice, pigs and
humans (Gebert, et al., 2015; Ha, et al., 2014; Ishizu, et al., 2012). Global knockouts of PIWI
proteins in mice all result in a loss of L1 suppression and male infertility. MIWI2 and MILI
knockouts caused an infertility due to early meiotic arrest, while MIWI knockouts arrested at the
round spermatid stage (Carmell, et al., 2007; Deng and Lin, 2002; Kuramochi-Miyagawa, et al.,
2004). Knockout studies suggest that in the male germline piRNA repression of L1s is essential
for reproductive success.

Because piRNAs are essential for male fertility, likely due to their suppression of L1
activity, it is important to understand how they control L1s and the role this repression
mechanism exerts in shaping the evolution of L1s. Bioinformatics analysis of next generation
sequencing (NGS) datasets has proven to be a powerful tool in L1 biology. Specifically, there
are numerous sequencing datasets for piRNAs in drosophila and mice germ cells which have
analyzed using bioinformatics approaches to uncover the mechanism of piRNA control of L1s
(Aravin, et al., 2007; Brennecke, et al., 2007; Carmell, et al., 2007; Reuter, et al., 2011; Reuter,
computationally analyze piRNA sequencing data, have shown that piRNAs target younger L1s,
suggesting that piRNAs may be exerting selective pressure on active L1 families. However, a
more comprehensive bioinformatics analysis, is needed to determine if young L1s are always
preferentially targeted during germ cell development and how this directly affects L1 evolution.
Using new L1 annotations within the mouse genome, Chapter 3 of this dissertation examines
what ages of L1 families still contain fully retrotranspositionally active L1s and are more likely to be a threat to the integrity of the genome. Furthermore, bioinformatics analysis piRNAs sequencing and transcriptome sequencing datasets from mouse testes and specific germ cells (spermatogonia, spermatocytes, spermatids and spermatozoa) at different times (embryonic, neonatal, and adult) during germ cell development are implemented to determine the role of piRNA are exerting in L1 evolution.
CONCLUSIONS

Bioinformatics is a powerful, cost effective method that integrates biology, mathematics and statistics to understand the complexities in biological systems. This dissertation utilizes bioinformatics approaches to better understand the metabolic and piRNA controls required for successful spermatogenesis in mice. By uncovering the molecular intricacies of spermatogenesis, male infertility and contraceptive needs can be better addressed. Chapter 2 uses FBA to address the role of metabolism in regulating spermatogonia differentiation. The activities of important pathways (i.e. central metabolism and vitamin A) are examined to elucidate their contribution in regulating differentiation of spermatogonia. In Chapter 3, by utilizing improved L1 annotations in the mouse genome, piRNA and transcriptome sequencing data, the role of piRNAs in the evolution of L1s is addressed. Finally, chapter 4 will discuss the overall conclusions from both Chapters 2 and 3 and discuss the future directions of this research. Overall, the data presented in this dissertation, obtained using bioinformatics methods, has enhanced our understanding of metabolic regulation of spermatogonial differentiation and the role piRNA exerts in controlling L1 evolution thereby leading to improved comprehension of germ cell regulation and male reproduction.
**Fig 1:** Cross section of a mouse seminiferous tubule. Within the tubule are Sertoli cells and germ cells. Sertoli cell nuclei stained with SOX9. Picture of tubule provided by Estela Jauregui.
**Fig 2:** Diagram depicting the basic reactions in vitamin A metabolism and the corresponding enzymes families that catalyze them.
**Fig 3:** Percentages of different types of transposons in the human genome. Figure modified from (Lander, et al., 2001).
**Fig 4:** Structure of a mouse L1: 5’ UTR consists of 200bp monomeric repeats (yellow triangles) which at least one of which is required for transcription of the element. ORF (open reading frame) 1 and 2 have nucleic acid/RNA binding and reverse transcriptase/endonuclease activity respectively which is required for transposition.
REFERENCES


costs in flux balance analysis reveals alternative pathways and metabolite stores in an
illuminated Arabidopsis leaf, Plant physiology.

Chung, S., et al. (2007) Mitochondrial oxidative metabolism is required for the cardiac

acid binding protein-I (CRABPI) null mutant mice, Differentiation; research in
biological diversity, 58, 141-148.

de Rooij, D.G. and van Beek, M.E. (2013) Computer simulation of the rodent spermatogonial
stem cell niche, Biology of reproduction, 88, 131.

essential for spermatogenesis, Developmental cell, 2, 819-830.

essential for spermatogenesis, Dev Cell, 2, 819-830.

and kinetics during testis development in mice: a high-resolution light microscopy
approach, Reproduction, 142, 145-155.

Dunster, J.L. (2015) The macrophage and its role in inflammation and tissue repair:
mathematical and systems biology approaches, Wiley interdisciplinary reviews. Systems
biology and medicine.

competencies to regulate spermatogenesis, Proceedings of the National Academy of
Sciences of the United States of America, 112, E2347-2356.


Raverdeau, M., et al. (2012) Retinoic acid induces Sertoli cell paracrine signals for spermatogonia differentiation but cell autonomously drives spermatocyte meiosis,
Proceedings of the National Academy of Sciences of the United States of America, 109, 16582-16587.


CHAPTER 2

DISSECTING GERM CELL METABOLISM THROUGH NETWORK MODELING.

The following chapter is formatted in accordance with guidelines of the PLOS ONE, and is currently in press.
Dissecting Germ Cell Metabolism through Network Modeling

Leanne S. Whitmore\(^1\), Ping Ye\(^{1,2*}\)

\(^1\)School of Molecular Biosciences, Washington State University, PO Box 647520, Pullman, WA 99164, USA

\(^2\)Department of Molecular and Experimental Medicine, Avera Cancer Institute, 1000 E 23rd Street, Sioux Falls, South Dakota, 57105, United States of America

*Corresponding author

Email: Ping.ye@avera.org

Received: 23 December 2014/ Accepted: 20 August 2015/
Published: 21 September 2015
Abstract

Metabolic pathways are increasingly postulated to be vital in programming cell fate, including stemness, differentiation, proliferation, and apoptosis. The commitment to meiosis is a critical fate decision for mammalian germ cells, and requires a metabolic derivative of vitamin A, retinoic acid (RA). Recent evidence showed that a pulse of RA is generated in the testis of male mice thereby triggering meiotic commitment. However, enzymes and reactions that regulate this RA pulse have yet to be identified. We developed a mouse germ cell-specific metabolic network with a curated vitamin A pathway. Using this network, we implemented flux balance analysis throughout the initial wave of spermatogenesis to elucidate important reactions and enzymes for the generation and degradation of RA. Our results indicate that primary RA sources in the germ cell include RA import from the extracellular region, release of RA from binding proteins, and metabolism of retinal to RA. Further, in silico knockouts of genes and reactions in the vitamin A pathway predict that deletion of Lipe, hormone-sensitive lipase, disrupts the RA pulse thereby causing spermatogenic defects. Examination of other metabolic pathways reveals that the citric acid cycle is the most active pathway. In addition, we discover that fatty acid synthesis/oxidation are the primary energy sources in the germ cell. In summary, this study predicts enzymes, reactions, and pathways important for germ cell commitment to meiosis. These findings enhance our understanding of the metabolic control of germ cell differentiation and will help guide future experiments to improve reproductive health.
Introduction

Continuous production of sperm throughout reproductive life is a fundamental feature of mammalian spermatogenesis. Understanding the complex process of spermatogenesis has important implications for male infertility and contraception. Presently, 70 million (i.e., 15%) married couples worldwide are infertile and half of these cases are attributed to male factors [1]. In approximately 25% of male infertility cases, the cause is unknown and treatment options are limited. Additionally, men have fewer options for contraception compared to women. There is a clear need to further our understanding of male reproduction in order to address infertility and contraceptive issues.

The initial development of spermatogonia to spermatozoa is termed as the first wave of spermatogenesis. It commences prior to puberty and is followed by subsequent rounds of spermatogenesis. Retinoic acid (RA), a derivative of vitamin A, is required for several processes during spermatogenesis, including the transition from undifferentiated to differentiating spermatogonia [2-4]. This differentiation event commits germ cells to meiosis and subsequently the formation of mature spermatozoa. Recent evidence showed that the highest levels of RA within the testis were detected when spermatogonial differentiation took place, both in neonatal and adult mice [5]. The pulse of RA represents a sharp increase in RA concentration every 8.6 days in the testis. This observation suggests that pulses of RA are likely the stimulus for both the initial wave and subsequent rounds of spermatogenesis.

It remains unclear which enzymes and reactions in which cell types synthesize or degrade RA, thereby generating the RA pulse. Evidence suggests that RA from Sertoli cells—the somatic supporting cells—is critical for the first wave of spermatogenesis but not for subsequent waves [6, 7]. However, the role of germ cells in regulating RA pulse has not been thoroughly
investigated. There is a crucial need to better understand whether or how vitamin A metabolism in the germ cell contributes to the pulse of RA.

Other metabolic pathways may be crucial in regulating spermatogonial differentiation through connection with vitamin A metabolism. Lipid metabolism affects the availability of retinol, a precursor of RA, as retinol is primarily stored in the form of retinyl esters [8-11]. Vitamin A metabolism is further linked to central metabolism (e.g., citric acid cycle, glycolysis) through energy molecules such as NADPH and NADH. In addition, all these pathways are essential during differentiation of other cell types [12]. Lipid metabolism was shown to determine self-renewal or differentiation of hematopoietic stem cells [13]. Glycolysis is the primary energy source in many stem cell populations. As stem cells differentiate, the activity of anaerobic glycolysis decreases (i.e., fermentation of pyruvate to lactate) while the citric acid (TCA) cycle takes over the energy production [12]. However, it is unknown how these metabolic pathways function during spermatogonial differentiation.

These knowledge gaps are likely due to the enormous effort required to perform germ cell-specific gene knockouts (KOs) and measure metabolic activity within the germ cell. Individual gene KOs in the vitamin A pathway often yield no phenotype [8, 9, 14]. Double or triple KOs are difficult to obtain but would be needed to elucidate the role of vitamin A metabolism in generating the RA pulse. Investigation of metabolic pathways would require direct measurement of enzyme activities and metabolite levels.

Because germ cell metabolism is highly complex, a thorough understanding cannot be achieved by only studying individual genes in animal models. A more profound understanding requires computational models to integrate genome-scale and individual studies of enzymes, metabolites, and reactions into networks of interacting components. A metabolic network model
can reveal key players by performing *in silico* perturbations of enzymes and reactions and predict system-wide outcomes on a scale that would be impossible by *in vivo* approaches [15-18]. We developed a mouse germ cell-specific metabolic network. A manually curated vitamin A pathway was integrated into the network to ensure accurate representation. Flux balance analysis (FBA) was performed on the network constrained with germ cell-specific gene expression [19]. With this *in silico* platform, we predict enzymes and reactions that are imperative for controlling RA levels within the germ cell. Further, we identify active pathways and energy producing pathways in the germ cell. This study provides a comprehensive understanding of how global metabolism regulates RA availability and germ cell differentiation.
Results

Construction of a germ cell-specific metabolic network

Vitamin A metabolic pathway

We compiled a list of metabolites and genes known to be involved in vitamin A metabolism (S1 Table). Reactions containing these metabolites or genes were extracted from human and mouse metabolic databases including Recon1, Recon2, EHMN, Reactome, and BioCyc [20-24]. Human reactions were considered to exist in the mouse if the catalyzing human enzymes had mouse orthologs [25]. We manually curated the extracted reactions by standardizing metabolite and gene names, eliminating reactions specifically in the vision cycle, and removing redundant reactions. Isomers (i.e., 9-cis, 11-cis, and 13-cis) of retinol, retinal, and RA were excluded, as our goal was to discover enzymes and reactions in germ cells that may contribute to the pulse of all-trans RA detected experimentally [5]. Different forms of retinyl-esters (i.e., retinyl-palmitate, fatty acid retinols, and retinyl-esters) were condensed into one set of metabolites. Reactions were assigned to a sub-cellular compartment: cytoplasm, nucleus, or extracellular region. The resulting vitamin A pathway comprises 45 reactions (13 reversible, 26 irreversible, and 6 exchange/sink reactions), 79 genes, and 44 metabolites (Fig 1).

Germ cell-specific metabolic network

We constructed a genome-scale mouse metabolic network from human Recon2 [21] by following a similar procedure described previously [26]. Specifically, human genes were converted into mouse orthologs; reactions with mouse orthologs or previously without human gene association were included in the mouse network. The resulting network was tested
iteratively until it passed a set of 260 validation tests, which ensured necessary metabolites and biomass could be produced.

A germ cell-specific network was constructed by extracting pathways potentially active in germ cells from the mouse genome-scale network. We utilized time-course germ cell-specific gene expression during the first wave of spermatogenesis [19]. Specifically, mice were treated with the compound WIN 18,446 for nine days starting from 1-day post-partum (dpp). WIN 18,446 inhibits the conversion of retinal to RA by targeting aldehyde dehydrogenases [27]. At 9 dpp, mice were given an injection of RA to initiate the synchronized first wave of spermatogenesis. Testes were collected at 0, 4, 12 hours, 1, 2, 4, 6, 8 days after the RA injection for expression profiling. This protocol ensures the enrichment of specific germ cell types during the first wave. Undifferentiated spermatogonia were initially enriched as a result of WIN 18,446 treatment whereas A differentiating spermatogonia were abundant at 4 hours through 2 days post-RA injection. Intermediate spermatogonia were enriched at 4 days. Preleptotene and leptotene spermatocytes accumulated by 6 and 8 days post-RA injection, respectively. In addition, A differentiating spermatogonia from the second wave of spermatogenesis were also enriched at 6 and 8 days (S1 Fig) [19]. Expression data were obtained by measuring ribosome-associated mRNAs in the germ cell. By capturing the level of mRNAs potentially being translated into proteins, this method approximates enzyme activity better than monitoring total mRNAs [28]. In addition, this method captures germ cell-specific gene expression. Note that the WIN 18,446/RA protocol used to profile gene expression was similarly implemented to detect the RA pulse during spermatogonial differentiation [5]. Thus, it is appropriate to use this expression data to predict enzymes and reactions in germ cells that may contribute to the observed RA pulse.
To extract potentially active pathways in the germ cell, we first classified genes in the mouse genome-scale network as “highly expressed” or “lowly expressed” using the expression data described above. A gene is considered “highly expressed” in the germ cell if its expression value is above the median value of all genes across all time points. Conversely, a gene is considered “lowly expressed” if its expression value is below the median value. A metabolic pathway is considered active if it is significantly enriched for reactions that are associated with “highly expressed” genes (hypergeometric P-value < 0.05). Reactions without gene associations were excluded prior to the P-value calculation.

Utilizing this procedure, 10 out of 92 metabolic pathways (excluding transport and exchange reactions) were identified as active in the germ cell for at least one time point: cholesterol metabolism, TCA, fatty acid oxidation, fatty acid synthesis, folate metabolism, glycolysis and gluconeogenesis, keratan sulfate degradation, keratan sulfate synthesis, N-glycan synthesis, and nucleotide interconversion. Vitamin A metabolism was added as the 11th pathway because of its important role in spermatogonial differentiation [2-4]. Then extracellular transport reactions were incorporated if they contain at least one metabolite in the active pathways. Intracellular transport and exchange reactions were incorporated if all metabolites in the reactions are in the active pathways. These pathways and reactions were combined to generate a germ cell-specific compartmentalized mouse network. The network is capable of producing biomass without any modifications. We cast the network in the SBML format to facilitate model sharing (S1 File). In addition, we provide Cytoscape session files [29] to visualize the network (S2 File). A total of 2,916 reactions (1,597 reversible and 1,319 irreversible) in eight sub-cellular compartments (cytoplasm, nucleus, mitochondria, endoplasmic reticulum, Golgi apparatus,
peroxisome, lysosome, and extracellular region) are captured in the network, with 636 genes and 2,072 metabolites (Fig 2).

**Dynamic flux of the vitamin A metabolic pathway**

Using the germ cell-specific network, we simulated network flux (i.e., reaction rate) at eight time points during the first wave of spermatogenesis. As not all enzymes are expressed at a given time, we used the germ cell-specific gene expression [19] to define the maximum allowable flux for reactions (i.e., reaction constraints). This was performed by mapping the expression of genes to their catalyzed reactions in the network. Therefore, if an enzyme is lowly expressed, its associated reaction would have low activity. Conversely, a highly expressed enzyme would indicate that the associated reaction has the potential to be active. Then at each time point, reaction fluxes were computed by Markov chain Monte Carlo sampling implemented with the COBRA Toolbox [30]. The fluxes values are influenced by both network structure and expression-defined reaction constraints. Using the network model, we directly investigated intracellular RA availability.

Our result shows that total RA production exhibit two peaks: an initial peak from 4 to 24 hours and a second peak on day eight post-RA injection (Fig 3). The initial RA peak is derived primarily from import from the extracellular region, in addition to internal metabolism from CRBP-retinal and RA unbinding from CRABP. Because the first wave of spermatogenesis was initiated by an injection of RA in the mice [19], it is reasonable to observe from the model that extracellular import is the major RA source. The same three reactions are also responsible for the peak on day eight. However, RA release from the binding protein CRABP becomes the dominant source, suggesting the binding reaction is essential for regulating RA levels within the germ cell.
RA degradation also peaks on day one and eight post-RA injection (Fig 3). Our results indicate that cytochrome P450 enzymes contribute minimally to RA degradation in the germ cell throughout the first eight days of spermatogenesis. This observation is supported by experimental evidence that Cyp26 was detected in other cell types but not germ cells in the testis [31, 32]. RA binding to CRABP contributes more to the reduction of RA, specifically between 2 and 4 days post-RA injection. We proceeded to evaluate freely available RA in the germ cell with a sink reaction. Sink reactions are unbalanced network reactions that allow the accumulation of metabolites. The flux of this sink reaction measures RA being produced but not degraded. Available RA in the cytoplasm serves as an indicator of spermatogonial differentiation as RA initiates differentiation by being transported into the nucleus and binding to nuclear receptors [8, 9, 10, 33]. We observe high levels of available RA on day 1 and 8 post-RA injection, which is consistent with experimental observations that showed a pulse of RA at those time points [5].

We further examined reactions that control the levels of retinol and retinal, two RA precursors (S2 Fig). We find that cellular uptake of retinol via Stra6 is minimal throughout the first eight days of spermatogenesis. This is consistent with experimental evidence that Stra6 is not expressed in germ cells and global Stra6 KO shows no spermatogenic defect [34, 35]. In addition, our model shows low activities of retinol esterification catalyzed by Lrat, suggesting that retinol storage does not occur in germ cells. This is supported by experimental data that Lrat is not expressed in germ cells [31, 32]. Further, we discover that retinol and retinal levels are directly regulated by the reversible reaction between them and binding reactions with CRBP. Our results show that a group of reactions form a loop to regulate retinol and retinal levels. At 12 hours post-RA injection, this loop works by unbinding retinal from CRBP and converting it to
retinol. Next, retinol is bound to CRBP and converted to CRBP-retinal. The loop reverses the direction on day 1, and then flips back on day 4 post-RA injection. This loop maintains a supply of retinal that can be metabolized for RA production.

Model prediction of RA enzymes and reactions required for spermatogenesis

Identification of criteria to predict spermatogenic defects

Using the tractable computational model, we can perturb enzymes and reactions individually or collectively. These virtual experiments allow us to overcome genetic redundancy to predict germ cell-specific enzymes and reactions that are most important for controlling RA levels. Two statistical tests were used to assess the difference between wild type (WT) (previous simulations in the subsection of “Dynamic flux of the vitamin A metabolic pathway”) and perturbation simulations. The Spearman correlation coefficient measures differences in global network flux values between WT and perturbation simulations. The Kolmogorov-Smirnov (K-S) statistic measures differences in available RA within the germ cell between WT and perturbation simulations. Perturbations are considered to cause spermatogenic defects if both network flux and available RA are different from those of WT simulation for at least one time point.

To identify thresholds of the two statistical tests used to predict spermatogenic defects, we conducted in silico enzyme KOs and treatments for which the effect on spermatogenesis has been experimentally demonstrated. We collected a total of 17 experimental perturbations on retinoid metabolism, including global KOs, germ cell KOs, Sertoli cell KOs, and compound treatments [3, 6, 7, 35-43]. Among these experiments, six cause spermatogenic defects and 11 produce no defects (Fig 4). We simulated each of the 17 perturbations, taking into the account the origin of perturbation (e.g., global KO, cell-specific KO). All possible threshold
combinations of the two statistical tests were evaluated. Using a Spearman coefficient less than 0.95 and a K-S statistic greater than 0.9 to define defects, our model achieved the best performance, replicating 16 out of 17 experimental results (Fig 4). All of the KOs and treatments causing spermatogenic defects were predicted correctly. Of the 11 no defect KOs, all but one (Crabp1+Crabp2 global KO) were predicted correctly.

**Prediction of new genes and reactions required for spermatogenesis**

We simulated single gene KOs for all the genes in the vitamin A pathway that were not used to generate prediction thresholds. Our model predicts that deletion of only one gene, Lipe, would result in a spermatogenic defect at hour 4 post-RA injection (Fig 5). LIPE catalyzes two reactions in the model: conversion from retinyl-ester to retinol and conversion from CRBP-retinyl-ester to CRBP-retinol. Lipe KO causes minor flux changes in these two reactions. However, major increases in retinol being imported into the cell and metabolized into RA are observed. The end result is an increase in available RA within the germ cell (S3 Fig). We speculate that this increase is due to a compensation response from the cell to deal with the deletion of Lipe.

Similarly, we performed single reaction KOs for all reactions in the vitamin A pathway, and found six are predicted to be required for spermatogenesis using the thresholds determined previously (Fig 5). Two of these six reaction KOs cause an increase in available RA. When the retinol binding reaction is deleted, the network flux is altered at hour 4 post-RA injection. More retinol is transported into the cell and converted into retinal and RA (S4 Fig). This again causes an increase in available RA, indicating that the retinol-binding reaction is important for the homeostasis of the pathway. When the reaction of CRBP-retinol to CRBP-retinal is deleted, a spermatogenic defect is predicted on day 4. The alternative reaction from retinol to retinal is
activated, thereby generating more available RA than in the WT (S5 Fig). This is a compensation response likely due to the disruption in the pathway homeostasis.

In contrast to the above two reaction KOs, the remaining four reaction KOs are predicted to cause spermatogenic defects due to a decrease in available RA. When CRBP-retinal to RA is deleted, a spermatogenic defect is predicted at hour 4 post-RA injection. This reaction is the major source of RA at that time (Fig 3), thus the deletion causes the reduction of available RA. Although RA import into the cell increases to compensate for the loss, the majority binds CRABP, further reducing the available RA (S6 Fig). Three reaction KOs (RA(e) exchange, RA(e) to ALB-RA(e), ALB-RA(e) to RA) all prevent RA from being imported into the cell. Extracellular RA is the sole RA source on day 2 (Fig 3), therefore deletion of any of the three reactions reduces the available RA to zero. Although these reaction KOs cause changes in retinol and retinal conversions, they are not metabolized to produce RA (S7 Fig).

Our reaction KO predictions are consistent with important reactions indicated by the dynamic flux of the vitamin A pathway (Fig 3, S2 Fig). Predicted reactions are major sources or sinks of retinoids in the germ cell. Disruption in these reactions either increases or decreases available RA; both can result in spermatogenic defects as reported experimentally [3, 6, 7, 36, 43].

Activities of other pathways in the germ cell-specific metabolic network

Besides vitamin A metabolism, we examined the activity of the other ten pathways in the germ cell-specific metabolic network. Activity was quantified by averaging the median flux of reactions included in each pathway (Fig 6). The TCA cycle is the most active pathway, steadily increasing in activity until day 6 post-RA injection and then decreasing. Nucleotide
interconversion is the second most active pathway, consisting of conversion reactions of mono-, di-, and tri-phosphate nucleotides. Because undifferentiated and differentiating spermatogonia are actively growing and mitotically dividing, they require nucleotides for incorporation into DNA. Folate metabolism, important for DNA replication and epigenetic modifications such as DNA methylation [44], has stable activity throughout the first eight days of spermatogenesis. The activity of cholesterol metabolism peaks within a day and then declines. Glycolysis and gluconeogenesis, fatty acid synthesis and degradation all have low metabolic activity during the first wave in germ cells. Three pathways, keratan sulfate synthesis, keratan sulfate degradation, and N-glycan synthesis, are completely inactive.

Because energy is vital for successful differentiation of cells, we further examined which metabolic pathways are producing and consuming energy (Fig 7). The three inactive pathways were excluded from this analysis. Our results show that fatty acid synthesis is the primary source of energy production in germ cells across the first wave of spermatogenesis. Highest energy yields occur at hour 4 and day 8 after initiation of spermatogenesis via an RA injection, corresponding to the timing of spermatogonial differentiation [45]. Fatty acid oxidation is the second highest energy-producing pathway, followed by vitamin A pathway and TCA cycle. Energy consuming pathways mainly include nucleotide interconversion, glycolysis and gluconeogenesis, and cholesterol metabolism.

**Metabolites important for germ cell differentiation**

We investigated two metabolites known to be important for stem cell differentiation: pyruvate and lactate [12]. Pyruvate is mainly produced from glycolysis in stem cells and converted to lactate for anaerobic respiration. This prevents pyruvate from entering the TCA cycle, which induces differentiation [12]. As stem cells differentiate, pyruvate is shunted into the
TCA cycle, resulting in decreased lactate production. Although our model does not capture stem cell differentiation, it describes a differentiation process of germ cells. We examined whether the metabolic trend of spermatogonial differentiation could be similar to that of stem cells.

We find that pyruvate is primarily produced and consumed via transport in and out of the germ cell. Pyruvate production through glycolysis is extremely low (Fig 8). This is consistent with our earlier result of minimal glycolysis activity throughout the first eight days of spermatogenesis (Fig 6). Because it is known that glycolysis is highly active in Sertoli cells [46], and our results show that germ cell pyruvate primarily comes from the extracellular region, we speculate that pyruvate is mainly produced by Sertoli cells and transported into germ cells. Pyruvate is then converted into lactate. We observe a decrease in lactate production (Fig 8) and an increase in TCA activity during the first wave (Fig 6). This indicates that TCA activity and lactate production during spermatogonial differentiation follow the same trend as stem cell differentiation. The low glycolysis activity throughout the first wave, however, is unique to the germ cell.
Discussion

RA is essential for germ cell differentiation [2, 3]. RA acts in a pulse manner, with concentration peaking every 8.6 days in the mouse thereby initiating spermatogonial differentiation [5]. We constructed a mouse germ cell-specific metabolic network. Using flux balance analysis with germ cell-specific gene expression, we reproduced the pulse of RA during the synchronized first wave of spermatogenesis (Fig 3). Our results show that multiple reactions contribute to the RA pulse in the germ cell, including import from extracellular region, metabolism from retinal, and RA unbinding from CRABP (Fig 3). Because RA was injected into mice to initiate the first wave [19], it is reasonable to observe that extracellular transport is the primary source of RA at early time points. Experimental evidence shows that Sertoli cells are primarily responsible for producing RA during the first spermatogenic wave [6, 7], and likely the reason we still observe extracellular transport as a major RA source at later time points. Metabolism of retinal and unbinding from CRABP are also important RA sources, suggesting that internal RA production contributes to the total RA levels in the germ cell. Further, we observe that RA is minimally degraded by reactions within the germ cell indicating that other cell types may be responsible for RA degradation.

Using the FBA model, we predict genes and reactions that are critical for regulating RA availability (Fig 5). Overall, few gene and reaction KOs in the germ cell were predicted to cause spermatogenic defects, likely because of genetic redundancy in the vitamin A pathway or activity of neighboring somatic cells [6, 7]. *Lipe* was the only gene predicted to have an important role in regulating RA levels. Previous experimental studies show that global KO of *Lipe* in mice causes major defects in round spermatids and a decreased sperm count leading to infertility [47-49]. Spermatid abnormalities were evident as early as five weeks after birth [49]. However, the
mechanism of infertility was not thoroughly investigated. One study speculated that a metabolite downstream from *Lipe* reaction might be essential for membrane integrity in germ cells [49]. Here we show that, *in silico*, *Lipe* KO results in an increase in available RA, altering the regulation of spermatogonial differentiation in the first wave. Although phenotypic defects were not observed at this stage in *Lipe* KO mice, mis-regulation of vitamin A metabolism may affect downstream spermatid development and may only be phenotypically evident until later stages of spermatogenesis.

In addition, six reaction KOs are predicted to cause spermatogenic defects (Fig 5). Three of these reactions transport RA into the germ cell, further suggesting that somatic cells are responsible for generating the RA pulse. While investigation into the role of Sertoli cells has started [6, 7], other cells such as Leydig and peritubular myoid cells may also be important. The three remaining reaction KOs are involved in vitamin A metabolism in germ cells. Specifically, disruption in retinol to CRBP-retinol and in CRBP-retinol to CRBP-retinal increases RA levels while disruption of CRBP-retinal to RA decreases RA levels. Together, these results from our model suggest that although somatic cells are the major regulator of RA pulse, vitamin A metabolism in the germ cell also contributes to RA levels, and consequentially, spermatogonial differentiation.

Other pathways in our model have been examined to understand their roles in spermatogonial differentiation. The TCA cycle is the most metabolically active pathway, increasing in activity during the first wave of spermatogenesis (Fig 6). This occurs concurrently with decreased production of lactate (Fig 8), consistent with the metabolic trend observed in stem cell differentiation [12]. Pyruvate, the end product of glycolysis and the precursor to TCA and lactate, is primarily obtained from the extracellular region (Fig 8). This is unsurprising as
glycolysis activity in the germ cell is low throughout the first wave (Fig 6). Experimental evidence indicates that Sertoli cells have high rates of glycolysis and provide pyruvate and other essential metabolites to spermatocytes [46]. Therefore, our results suggest that Sertoli cells are likely the source of glycolysis intermediates for germ cells. In addition, our model predicts that fatty acid metabolism is the primary producer of energy throughout the first wave of spermatogenesis (Fig 7). Energy production peaks concurrently with the timing of the RA pulse and spermatogonial differentiation. This indicates that fatty acid metabolism may indirectly affect the RA pulse by controlling the availability of co-factors and energy molecules required for reactions. This interesting finding from our model will require experimental investigation to further elucidate connections between fatty acid metabolism and vitamin A metabolic pathway. In summary, our FBA model constitutes a powerful tool for manipulating genes and reactions to determine their contribution to vitamin A metabolism in the germ cell. These in silico perturbations prioritize cell-specific enzymes and reactions based on quantitative flux metrics, thus providing directions for future functional studies. Further, our FBA model traces metabolic activities of other pathways, allowing for the identification of pathways important for energy production and therefore spermatogonial differentiation. Future improvements of the computational approach will include integrating genome-scale metabolic networks of multiple cell types (e.g., germ, Sertoli, and Leydig cells) in the testis to elucidate how enzymes/reactions in other cells affect RA availability in germ cells.
Materials and Methods

Flux balance analysis

FBA calculates the flow of metabolites through a network under the assumption that each metabolite has equal production and consumption. All reactions in the network are represented in a stoichiometric matrix, $S_m \times r$, where $m$ is the number of metabolites and $n$ is the number of reactions. Matrix entries are metabolite coefficients in each reaction. Constraints are placed upon each reaction: $V_{\text{min}} \leq V \leq V_{\text{max}}$, defining the range of allowable flux values. Steady state analysis is carried out by setting mass balance equations equal to zero: $S_m \times rV = 0$.

FBA constraints

Maximum and minimum constraints were placed on each reaction in the network. Germ cell-specific gene expression was used to define the maximum flux. Briefly, 1 dpp mice were fed WIN 18,446 for 9 consecutive days. At 9 dpp, mice were injected with RA to initiate spermatogenesis. Testes were collected at 0, 4, 12 hours, 1, 2, 4, 6, 8 days post-RA injection for expression profiling. Ribosome associated mRNAs within germ cells were pulled down to evaluate global gene expression (GSE54408) [19]. Microarray data were normalized using the Robust Multi-array Average (RMA) method. The probe IDs were translated into Ensembl gene IDs based on the Mouse ST Gene 1.0 transcript annotation version 33.2. In the case of multiple probe IDs for one gene, the average expression value was used. Finally, signals from duplicate samples were averaged to yield the final gene expression value. Expression values for the 636 genes in the network were extracted. If a reaction is catalyzed by multiple isozymes, the maximum constraint was assigned with the value of the highest expressed isozyme. If a reaction is catalyzed by enzyme complexes, the maximum constraint was assigned with the value of the...
lowest expressed complex subunit. Once maximum constraints are determined for reactions associated with enzymes, they were scaled to have an average of 1,000 over the eight time points. For reactions without enzyme association or without expression data, the average of maximum constraints in the network was used.

Minimum constraints were determined based on reaction reversibility; irreversible reactions were given a value of zero while reversible reactions were assigned the negative value of the maximum constraint. For intracellular binding reactions of retinyl-ester, retinol, retinal, and RA, the minimum constraint was the maximum constraint at the time point subtracted by the highest maximum constraint across the eight time points.

**Sampling analysis of the network flux**

Monte Carlo Markov chain sampling was used to acquire the distribution for all possible flux states. This was performed using a hit-and-run algorithm—gpSampler in the Cobra Toolbox [30]. At each time point with expression-based constraints and network topology, the solution space was sampled with 6,000 points for 17 minutes. This results in 6,000 feasible solutions for all reactions acquired from the uniformly sampled points. From this distribution, the median flux value was obtained for each reaction.

**Enzyme and reaction deletion analysis**

To simulate germ cell-specific enzyme deletion, the expression of the enzyme was set to zero, and the maximum constraints were re-calculated for reactions catalyzed by this enzyme. To simulate germ cell-specific reaction deletion, the maximum and minimum constraints were both set to zero. Enzyme deletion in the Sertoli cell was simulated by altering the extracellular availability of metabolites through exchange reactions. Specifically, Sertoli KOs of *Rdh10* and *Aldh1a1,2,3* were mimicked by setting the minimum constraint of RA exchange reaction to zero,
thus eliminating RA from the extracellular region. Conversely, Sertoli KO of Cyp26b1 was mimicked by setting the maximum constraint of RA exchange reaction to -900, thus forcing RA to be present in the extracellular region. WIN 18,446 treatment was reproduced by setting maximum constraints to zero for all reactions catalyzed by aldehyde dehydrogenases, in addition to setting the minimum constraint of RA exchange reaction to zero. To replicate vitamin A deficiency, minimum constraints were set to zero for exchange reactions of any form of vitamin A (retinol, retinal, β-carotene, and RA). After the network was manipulated for each perturbation, candidate states were re-sampled and compared with normal candidate flux states by calculating the Spearman correlation coefficient and the K-S statistic.
Acknowledgments

The authors would like to thank Debjit Ray, Cathryn Hogarth, Travis Kent, and Lisa Evans for helpful discussions.
References


56


33. Mark M, Ghyselinck NB, Chambon P. Function of retinoid nuclear receptors: lessons from genetic and pharmacological dissections of the retinoic acid signaling pathway during...


Fig1. **Vitamin A metabolic pathway.** Reactions involved in retinoid metabolism are depicted. Exchange/sink reactions and promiscuous metabolites (e.g., energy molecules, O$_2$, H$_2$O) are omitted from the diagram. The rounded rectangle denotes germ cell compartment.
Fig 2. A mouse germ cell-specific metabolic network. Reactions in the network are grouped into pathways. The number of reactions in each pathway is labeled.
Fig 3. Flux of reactions that directly produce or degrade RA. Median flux values from the model are plotted across eight days after initiation of spermatogenesis via an RA injection.
Fig 4. Identification of criteria to predict spermatogenic defects. A total of 17 experimental perturbations on male fertility are listed, including global KOs, cell-specific KOs, and compound treatments. Among these perturbations, six cause spermatogenic defects and 11 produce no defects. Using a Spearman coefficient less than 0.95 and a K-S statistic greater than 0.9 for at least one time point to predict defects, our model results match 16 out of 17 experimental perturbations.
Fig 5. **Genes and reactions predicted to be required for spermatogenesis.** One gene and six reactions are predicted to cause spermatogenic defects, defined by having a Spearman coefficient less than 0.95 and a K-S statistic greater than 0.9 for at least one time point.
Fig 6. Activities of other pathways in the germ cell-specific metabolic network. The average pathway flux is calculated by summing the median flux values of each reaction and dividing by the number of reactions in a pathway at each time point.
Fig 7. The contribution of individual pathways to energy production and consumption. All energy molecules are converted to the ATP unit: 1NADH=3ATP, 1NADPH=4ATP, 1FADH2=2ATP, 1GTP=1ATP, 1CTP=1ATP, 1TTP=1ATP. Net ATP production is calculated for each pathway at each time point by subtracting ATP consumption from production. A positive value means a pathway generates energy while a negative value indicates the pathway consumes energy.
Fig 8. Flux of reactions that directly produce or degrade pyruvate. Median flux values from the model are plotted across eight days after initiation of spermatogenesis via an RA injection.
S1 Fig. Types of germ cells enriched at different time points post-RA injection. Mice were treated with WIN 18,446 for nine days starting from 1 dpp. At 9 dpp, mice were given an injection of RA to initiate spermatogonial differentiation.
S2 Fig. Flux of reactions that directly produce or degrade retinol and retinal. Median flux values from the model are plotted across eight days after initiation of spermatogenesis via an RA injection.
S3 Fig. Flux changes in the vitamin A pathway when *Lipe* is deleted *in silico* four hours after initiation of spermatogenesis via an RA injection. A. Difference in median flux values between WT and *Lipe* KO for reactions in the vitamin A pathway. Reactions with difference less than 50 are not shown. B. Sampling flux distribution for the reaction of available RA.
S4Fig. Flux changes in the vitamin A pathway when the reaction retinol to CRBP-retinol is deleted *in silico* four hours after initiation of spermatogenesis via an RA injection. A. Difference in median flux values between WT and this reaction KO for reactions in the vitamin A pathway. Reactions with difference less than 50 are not shown. B. Sampling flux distribution for the reaction of available RA.
S5 Fig. Flux changes in the vitamin A pathway when the reaction CRBP-retinol to CRBP-retinal is deleted *in silico* four days after initiation of spermatogenesis via an RA injection.

A. Difference in median flux values between WT and this reaction KO for reactions in the vitamin A pathway. Reactions with difference less than 50 are not shown. B. Sampling flux distribution for the reaction of available RA.
S6 Fig. Flux changes in the vitamin A pathway when the reaction CRBP-retinal to RA is deleted in silico four hours after initiation of spermatogenesis via an RA injection. A. Difference in median flux values between WT and this reaction KO for reactions in the vitamin A pathway. Reactions with difference less than 50 are not shown. B. Sampling flux distribution for the reaction of available RA.
S7 Fig. Flux changes in the vitamin A pathway when RA import into the cell is deleted in silico two days after initiation of spermatogenesis via an RA injection. Three single reaction deletions yielded the same result: RA(e) exchange, RA(e) to ALB-RA(e), and ALB-RA(e) to RA. A. Difference in median flux values between WT and single reaction KO for reactions in the vitamin A pathway. Reactions with difference less than 50 are not shown. B. Sampling flux distribution for the reaction of available RA.
Table S1. Metabolites and genes known to be involved in vitamin A metabolism from the literature [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15].

<table>
<thead>
<tr>
<th>Metabolite</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2E,4E,6E,8E)-3,7-dimethyl-9-(2,6,6-trimethyl-1-cyclohexenyl)nona-2,4,6,8-tetraenyl hexadecanoate</td>
</tr>
<tr>
<td>(2E,4E,6E,8E)-3,7-dimethyl-9-(2,6,6-trimethylcyclohex-1-en-1-yl)nona-2,4,6,8-tetraen-1-yl acetate</td>
</tr>
<tr>
<td>(2E,4E,6E,8E)-3,7-dimethyl-9-(2,6,6-trimethylcyclohex-1-enyl)nona-2,4,6,8-tetraen-1-ol</td>
</tr>
<tr>
<td>(2E,4E,6E,8E)-3,7-dimethyl-9-(2,6,6-trimethylcyclohexen-1-yl)nona-2,4,6,8-tetraenal</td>
</tr>
<tr>
<td>(2E,4E,6E,8E)-3,7-dimethyl-9-(2,6,6-trimethylcyclohexen-1-yl)nona-2,4,6,8-tetraenoic acid</td>
</tr>
<tr>
<td>4-oh-RA</td>
</tr>
<tr>
<td>4-oxo-RA</td>
</tr>
<tr>
<td>acetylretinol</td>
</tr>
<tr>
<td>all-trans-retinal</td>
</tr>
<tr>
<td>all-trans-retinene</td>
</tr>
<tr>
<td>all-trans-retinol</td>
</tr>
<tr>
<td>all-trans-retinol acetate</td>
</tr>
<tr>
<td>all-trans-retinyl acetate</td>
</tr>
<tr>
<td>all-trans-Vitamin A acetate</td>
</tr>
<tr>
<td>all-trans-Vitamin A aldehyde</td>
</tr>
<tr>
<td>beta-carotene</td>
</tr>
<tr>
<td>Retinal</td>
</tr>
<tr>
<td>retinaldehyde</td>
</tr>
<tr>
<td>Retinoate</td>
</tr>
<tr>
<td>retinoic acid</td>
</tr>
<tr>
<td>Retinol</td>
</tr>
<tr>
<td>retinol acetate</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>retinol palmitate</td>
</tr>
<tr>
<td>Retinyl</td>
</tr>
<tr>
<td>retinyl-ester</td>
</tr>
<tr>
<td>retinyl acetate</td>
</tr>
<tr>
<td>retinyl palmitate</td>
</tr>
<tr>
<td>trans-retinol</td>
</tr>
<tr>
<td>Vitamin A</td>
</tr>
<tr>
<td>Vitamin A acetate</td>
</tr>
<tr>
<td>Vitamin A acid</td>
</tr>
<tr>
<td>Vitamin A palmitate</td>
</tr>
<tr>
<td>Vitamin A1</td>
</tr>
<tr>
<td>Vitamin A1 acetate</td>
</tr>
<tr>
<td>Vitamin A aldehyde</td>
</tr>
</tbody>
</table>

**Gene**

| Abca |
| Adh |
| Alb |
| Aldh |
| Arat |
| Bcma |
| Cel |
| Crabp |
| Cyp26 |
| Dgat1 |
| Hspg |
Ldlr
Lipe
Lpl
Lrat
Lrp
Plin2
Plrp2
Pnlip
Raldh
Rbp
Rdh
Scarb
Stra6
Vldlr
References


CHAPTER 3

ROLE OF PIWI–INTERACTING RNA IN LINE-1 EVOLUTIONARY DYNAMICS
Role of PIWI-Interacting RNA in LINE-1 Evolutionary Dynamics

Leanne S. Whitmore¹, Debjit Ray¹, Wenfeng An¹², and Ping Ye¹²,*

¹School of Molecular Biosciences, Washington State University, PO Box 647520, Pullman, WA 99164, USA
²Department of Pharmaceutical Sciences, South Dakota State University, Brookings, South Dakota 57007, USA
³Department of Molecular and Experimental Medicine, Avera Cancer Institute, 1000 E 23rd Street, Sioux Falls, South Dakota, 57105, USA.

* Corresponding author
Email: pye@wsu.edu
ABSTRACT

Transposons have been active in mammalian genomes for millions of years driving evolution by generating genetic and epigenetic changes in the genome. The autonomous retrotransposon long interspersed element-1 (L1) uses an RNA intermediate to move and constitutes approximately 17% of the human genome. L1 retrotransposition can be detrimental to the organism by causing disruptions in the coding sequences of genes and regulatory regions of genes. L1s evolve in a single lineage manner in numerous mammalian species acquiring a new 5’ untranslated region (UTR) when a novel family emerges. The biological mechanism driving this unique pattern of L1 evolution is not fully understood. PIWI-interacting RNAs (piRNAs), small non-coding RNAs, repress L1 transcription by targeting DNA methylation machinery to the promoter regions of L1s. PiRNAs can further prevent L1 transposition by degrading L1 transcripts. As the 5’UTR sequence has been shown to be over represented in prentatal piRNAs, we hypothesize that piRNA-mediated repression plays a major role in driving L1 lineage succession. To test this hypothesis we analyze piRNA abundance toward L1 families in mice. Our analysis shows that piRNA abundance is highest toward younger and transcriptionally more active L1 families suggesting that the piRNA system selectively acts against active L1 families, leaving room for new families to emerge.
INTRODUCTION

Long interspersed element-1 (L1) is one of the remaining active transposons in the mouse and the only active mobile element in the human genome [1]. L1s are a retrotransposon meaning they proliferate in genome via a RNA intermediate. Initial transcription of an L1 in the mouse is dependent on the presence of a functional monomeric promoter [2-4]. Mouse L1 monomeric promoters consist of 200 base pairs (bps); one full repeat is required for transcription [5]. Once the RNA intermediate is transported to the cytoplasm and ORF1, which contain nucleic acid and RNA binding domains, and ORF2, that has reverse transcriptase and endonuclease activities, are translated [1]. This ribonucleoprotein complex is transported back into the nucleus and is integrated into a new location in the genome by reverse transcription of the RNA intermediate. Approximately 17% and 20% of the human and mouse genomes respectively consist of L1s [6, 7]. The high percentage of L1s in both the human and mouse genomes demonstrates the ability of these elements to replicate within genomes.

L1s in a number of species, including mice and humans, evolve in a single lineage manner, where one family at a time is active and then is replaced with a novel L1 family [8-11]. As L1s evolve a novel 5’UTR promoter sequence is acquired [11-13]. The reason for this type of evolution remains unclear although it is hypothesized that limitation of the availability host factors necessary for retrotransposition make it possible for only one family to be active at a given time. Multiple families of L1s have shown to co-exist if they have different 5’UTR promoters as to not exhaust the supply of host transcription factors [11, 13]. Eventually a novel L1 family would emerge, become transcriptionally dominant and sequester other unknown essential host factors leading to the demise of older L1 families. However, no host factors that may limit L1 activity due to their inadequate availability have been identified [14-16]. There we
hypothesize that defense mechanisms in the host cells have may play a role in the single-lineage evolution of L1s.

There are many mechanisms developed by germ cells to defend the genome from integration of L1s [1]. A predominate defense against retrotransposition is PIWI-interacting RNAs (piRNAs) which are small RNAs that prevent L1 initial transcription and facilitate degradation of L1 transcripts. PiRNAs are 24-30bps in length and bind to agrounaute (PIWI) proteins MILI, MIWI2, and MIWI, which are expressed in the male germ cells of mice and humans [17, 18]. MIWI2 is expressed from embryonic day (E) 14.5 to postnatal day (P) 3, MILI from E12.5 to adulthood and MIWI from P14 to adulthood [19, 20]. PiRNAs which are initially transcribed from piRNA clusters, which consist of old degraded transposon sequences form a ribonucleoprotein complex with MIWI2, MILI and MIWI and target L1s for degradation[21, 22]. Fragments of degraded L1s can become piRNAs themselves, in a process termed the ping-pong loop which is active primarily in the prenatal germline and target other L1 transcripts for degradation. Furthermore MIWI2 and MILI bound piRNAs, facilitate the methylation L1 DNA sequences during male germline de-novo methylation [23, 24]. Because piRNAs are a dominant and adaptive defense mechanism against L1 activity in male germ cells we hypothesize that piRNAs play a significant role in the evolution of L1s.

PiRNAs were shown to preferentially target younger L1 families [7, 25]. These data suggest that the ability of piRNA to inhibit replication of active L1 families may force L1s to evolve in order to evade silencing. Therefore, expression of L1 families, specifically, younger L1s would be dependent on piRNA abundance. Higher piRNA abundance would suggest that there would be minimal L1 expression thereby limiting potential retrotransposition activity. Furthermore, previous work showed that piRNAs early in germ cell development, E16.5, target
the 5’UTR of L1s [19]. This may exert evolutionary pressure on the 5’UTRs of the L1, thereby forcing L1s to acquire new 5’UTRs in order to evade suppression by piRNAs.

Despite the numerous levels of control insertion of L1s does occur. The location of the L1 insertion is vitally important as integrating into a coding region or promoter region of a gene could be detrimental to the host cell/organism and subsequently the L1. It was shown that piRNAs preferentially target L1s that are located proximally to genes location suggesting for best replication efficiency L1s may preferentially insert distally or intragenically to a gene [7]. Therefore, understanding if there is a preference of location relative to genes for insertion of L1s will illuminate the complex evolution of L1s.

Utilizing new classifications for L1 families and several piRNA and transcriptome sequencing datasets we aim to understand how piRNA suppression of L1s contribute to the single lineage evolution observed in many species. Within each family we classify L1s based on different levels of retrotransposition potential to families of L1s piRNAs are likely to target. We utilize published piRNA-seq data and RNA-seq data from germ cells and whole testis to analyze the relationship between piRNA abundance, L1s expression and L1 age. Additionally, we examine whether piRNAs preferentially target the 5’UTR of young, middle aged, and old L1s. Together, our data from this study imply that the selective negative evolutionary pressure piRNAs put forth on young L1s are a significant factor in controlling single-lineage evolution of L1s.
RESULTS

L1 classifications

In order to examine L1 families that were likely to be active, the 29 families were further classified into 4 inclusive subcategories: all L1s, L1s with a 5’UTR, L1s with a 5’UTR and full-length (FL) (FL L1s) and L1s with a 5’UTR, FL and intact (intact L1s) (Fig 1).

Transcriptionally active L1s

First L1s that have the necessary L1 components to be potentially transcriptionally active (L1s with a 5’UTR) were classified. An L1 in the mouse genome was defined as potentially transcriptionally active if the L1 contained at least one complete monomer in the promoter sequence. Using the RepeatMasker results we were able to examine the alignment between the 5’end consensus and mouse genome, if the alignment contained at least one full monomer in the consensus (see Methods for how consensus sequences were annotated) the genomic L1 correspondingly had a complete monomer as well and was considered transcriptionally active.

Transcriptionally active and FL L1s

FL L1s were selected from the potentially transcriptionally active L1s. L1s that were found to have all 3 parts of an L1 consensus sequence (5’end, orf2, and 3’end), that were overlapping (meaning that there were no gaps between the 3 parts), on the same strand, and at-least 5500bps and less then 8000bps were considered full-length. The 8000bp upper bound was determined necessary in order to eliminate L1s that had multiple recombination events and were excessively long (Sup Fig 1a).

Transcriptionally, FL & intact L1s

Finally, L1s were considered fully capable of transposition if they had 5’UTR, FL and contained an intact ORF1 & ORF2. To identify L1s with intact ORF1s and ORF2 protein
sequences were first extracted from L1s if the ORFs contained >300 amino acids (aa) from each FL L1 identified. ORFs from L1s that contained at least two ORFs >300 aa’s were then aligned using blastp to the corresponding consensus ORF1 and ORF2 protein sequences [26]. Note that if an L1 was recombined the genomic ORFs were aligned to the consensus ORFs for each L1 family the genomic L1 was aligned to by RepeatMasker. After testing numerous different thresholds for classifying intact L1s (Sup Fig 1b) we selected that the 2 ORFs must have 99% similarity and were within 99% length to ORF1 and ORF2 protein consensus sequences respectively. Additionally, we tried using percent identity instead of similarity as a cutoff threshold, however, based on results of mutation studies for ORF1 and ORF2, similarity was determined more accurate. This is because when substituting an aa with another similarly charged and structured AA L1s retain ability for transposition [27-29]. In examining the number of L1s that had intact ORF1 versus intact ORF2 sequences, we observed that more L1s in each family have intact ORF1s than intact ORF2s, independent of what cutoff threshold was used (Sup Fig 1c). This suggests that ORF2 undergoes more mutations (i.e. RNAi editing or recombination with other L1s). Overall we determined that 12 families have intact L1s.

These classifications (Fig 1) show that older families have more L1s that are 5’ truncated and are unable to transpose. Conversely, younger families have many more full-length and intact L1s suggesting that these families have more potential to be detrimental to the genome.

**PiRNA abundance toward young and old L1 families**

In order to examine the relationship between piRNA abundance and age we utilized 8 RNA sequencing datasets on small RNAs from MIWI2, MILI and MIWI collected at different times during germ cell development (Sup Table 1) [30-32]. For each data set we examined the length of processed reads to ensure piRNAs, which are 24 – 30 bps in length, were present in the
datasets (Sup Fig 2). To capture how piRNA abundance and L1 age are related, piRNA sense and antisense abundance toward L1s in each family and subcategory were summed up and then normalized for the genomic copy number of L1s. Normalization for L1 copy number was necessary because older L1 families have more copies and therefore without normalization we may only be observing copy number dependence of piRNA abundance rather than age (Sup Fig 3). Sense and antisense piRNA abundance were split in order to understand how piRNA abundance is related to autonomous L1 expression.

For sense piRNA we observed a strong negative correlations (meaning there is higher piRNA abundance toward younger families) for the subcategories all L1s, L1s with a 5’UTR and FL L1 from E18 until adulthood (except MIWI adult piRNA) (Fig 2a). Younger L1s throughout germ cell development are targeted more by piRNAs and therefore are contributing more to the piRNA pool thereby exerting selective pressure on these active L1s. Intact L1s which all are in young families generally do not exhibit strong negative or positive correlations because all 12 families have high abundance of sense piRNA (data not shown). Antisense piRNA abundance exhibit strong negative correlations for MIWI2 and MILI bound piRNA at E18 and P10 respectively for all L1s, L1s with a 5’ UTR and FL L1s. MIWI2 directs methylation machinery to younger L1 sequences; therefore, it is unsurprising that these piRNAs are more abundant toward younger L1s. Interestingly, positive correlations (more piRNA abundance for older L1 families) for MILI E18, MILI P0, MILI adult, MIWI P14, P20 and adult for all L1s. This indicates that piRNA toward older families is more antisense orientated and therefore not being generated from L1 transcripts like younger families. This observation between sense and antisense correlations are further supported by examining the log ratio of sense versus antisense piRNA stranded for each L1 family and correlated to age (Sup Fig 4). There are significant
negative correlations (meaning there is more sense oriented piRNA for younger families compared to old) for all piRNA data sets for all subcategories except intact L1s. Intact L1s do not show this correlation because all 12 families are relatively young and all have a greater abundance of sense versus antisense piRNA (data not shown). This observation of greater sense piRNA abundance toward young L1s suggests that many piRNAs are being generated from active L1 transcripts thereby building up a defense towards active younger L1s. To further support our conclusion we examined the ratio of primary to secondary piRNAs as it relates to age of L1 families.

To examine if the ping-pong cycle is more active toward younger families, which converts L1 transcripts to piRNA, we examined the abundance primary and secondary piRNA in sense and antisense orientated piRNA for the different L1 families. PiRNA was considered primary if it started with a U and secondary if it had an A in the 10th position [19]. If a read had both a U and an A at the 1st and 10th position the read was not included in the analysis. Log ratios of primary/secondary normalized for copy number piRNA expression for sense and antisense were calculated for each L1 family and then correlated with ages (Fig 2b). A negative correlation indicates that younger families have more primary piRNAs and older families have more secondary piRNAs while a positive correlation means that older families have more primary piRNAs and younger families have more secondary piRNAs. Previous studies have demonstrated that there is a higher fold of primary piRNAs relative in sense oriented piRNA and secondary in antisense piRNAs [19]. Because we observe negative correlation (E18 to P0 for PIWI proteins) of sense piRNA in the subcategory FL L1s and in subcategories “upstream” (L1s with 5’UTR and all L1s) these piRNAs are likely a result of the ping pong cycle. Interestingly, the positive correlations at E18 and P0 for antisense piRNAs observed for FL L1s and upstream
subcategories suggest that for younger families more secondary piRNAs exist thereby providing more support that the ping pong mechanism is actively suppressing these L1s. Antisense primary piRNAs are hypothesized to be from the primary processing piRNA pathway and not entering the ping pong cycle, likely due to the minimal expression of these middle aged to older L1s. At MILI P10 the ping pong mechanism is no longer active [33] therefore piRNAs targeting younger L1 families are less prevalent, however, piRNAs are still being generated via the primary piRNA pathway which includes transcribing older L1s which are part of the piRNA cluster thereby explaining the positive correlations we observe for both sense and antisense piRNAs after MILI P0. Again, the differences in correlations results for the intact L1 subcategory is because only 12 relatively young families are included in this correlation and all ages of L1s appear to be targeted equally by the ping pong mechanism at early time points. Antisense piRNA from MILI P10 and later appear to target FL and intact L1s even the mechanism of production has changed. Because the piRNA are generated from older L1 families these piRNA maybe preferentially older intact and FL L1s. These results suggest that piRNAs preferentially target active L1s early in germ cell development that are a threat to the integrity of the organism. Targeting of young L1s by piRNAs are exerting selective pressure on the L1s to evolve in order to evade detection.

As L1s new L1 families emerge they acquire a novel 5’UTR sequence [9, 11-13]. Previous studies have shown that piRNAs preferentially target the 5’ end of the L1 sequence at embryonic day16.5 during germ cell development and therefore put selective pressure on the 5’UTR of L1s to evolve [19]. We wanted to examine for all L1 families if this pattern of 5’ end preferentiality is consistent for genomically mapped piRNAs. First piRNA were mapped directly to the 29 consensus sequences to examine if piRNA similarly targets different age of L1s in the same way as was reported by [19] when piRNAs were directly mapped to a general L1
consensus sequence. PiRNAs mapped to the 5’end and 3’ end were summed up and the log2 ratio of 5’end piRNA divided by 3’ end was calculated (Fig 3a, Sup Fig 5a). MIWI2 E18 piRNA for young families preferentially target the 5’end while for older families the distribution of 5’end and 3’end piRNAs is about equal for both sense and antisense piRNA. MILI E18 and P0 piRNA strongly target the sense strand 5’end of younger and older L1s although there is a slight decrease in the sense 5’end preference for older L1 families. Antisense also preferentially target the 5’end of younger families but show a slight preference toward the 3’end of older families. MILI P10, adult, MIWI P14, P20 and adult sense and antisense piRNA target the 3’end which becomes more dominant as L1s age.

We next determined whether this trend is similar for genomically mapped piRNAs to L1s. Interestingly, different results were observed for piRNAs mapped genomically versus mapped directly to consensus sequences (Fig 3b, Sup Fig 5b). For sense and antisense orientated MIWI2 E18 piRNAs for younger L1s there is approximately an equal distribution of 5’end and 3’end sense and antisense piRNAs. Middle aged and older family sense and antisense piRNAs preferentially target the 3’ end. A similar trend is observed for MILI E18 and P0 piRNAs with exception of how piRNAs are targeting older L1s. At P10 and later piRNAs from any PIWI protein preferentially target the 3’ end of any aged L1s.

The results between the consensus and genomic mapping are different at E18 and P0 for either PIWI protein. To understand why a bimodal peak observed in young families (Sup Fig 5) we examined where L1 genomic copies in each family the TSS was in relation to the entire consensus sequence. When examining the starting position of genomic L1s we see that a majority start either at the 5’end or 3’end of the sequence with very few starting in the middle explaining the lack of piRNA to the middle of the sequence (Sup Fig 6).
**L1 expression correlated piRNA expression**

In order to further examine the relationship between piRNA and L1 expression we correlated piRNA and L1 expression. L1 expression was determined by using RNA-Sequencing data from whole testes and specific types of germ cells (Sup Table 2) [34-38]. There is a significant negative correlation (meaning that L1 families with high expression have low piRNA expression and for L1 families with low expression there is high piRNA expression) between sense piRNA and L1 expression for MIWI P14 piRNA and L1 whole testis expression at P14 for subcategories all L1s, L1s with a 5’UTR and FL L1s (Fig 4). At P20 there are significant positive correlations (meaning for L1 families with low expression the piRNA expression is also low and for L1 families with high expression there is correspondingly high piRNA expression) between MIWI P20 sense piRNAs and whole testis L1 expressions for all subgroups of L1s. MIWI and MILI adult sense piRNAs significantly positively correlate with adult whole testis L1 expression. In examining cell specific L1 expression from different germ cell types all from adult testes we observe significant positive correlations between all piRNA and L1 expression datasets for all L1s. Additionally, positive correlations were observed between spermatocytes L1 expression and MIWI adult sense piRNA for L1s with a 5’UTR and FL L1s, spermatid L1 expression with MIWI and MILI adult sense piRNA for L1s with a 5’ UTR and spermatid and spermatozoa L1s expression with MILI adult sense piRNAs for intact L1s.

The lack of significant correlation between piRNA and L1 expression of families at early time points and higher abundance of piRNAs compared to L1 transcripts (can be observed in Fig 5) suggests that piRNAs with potential help from other defense mechanisms are significantly preventing L1 expression. At P14 and P20 meiosis is occurring and chromosomes are undergoing massive restructuring to allow recombination. At P14 piRNA clusters, which contain
truncated older L1 sequences, that are not as active earlier in development are being expressed and therefore maybe clusters are just starting to become transcriptionally active but have yet to produce piRNA. Conversely, the positive correlations observed at P20 for all of the subcategories suggests that as time progresses during meiosis more piRNAs are being produced from piRNA clusters (older expressed L1s). Similarly, the positive correlation between L1 and piRNA expression in the subcategory all L1s is observed in specific germ cells from adult is likely due to the same reason positive correlations are observed at P20 and adult testis. The strong correlations observed in germ cell specific samples may be due to the enrichment of expression data specifically from germ cells.

**piRNA and L1 expression based relative location to genes**

Where L1s insert in the genome is important to understand as it can have detrimental effects on the host. Prevalence of L1s located intragenically (genic) (within a gene), proximally (10kb up or downstream of a gene) or distally (greater than 10 kb up or downstream of a gene) was determined for each of the 29 families (Fig 5a). L1s predominately insert genically or distally in relation to genes for all 29 families. To understand if piRNAs target L1s preferentially based on their location to genes we examined piRNA expression to genic, proximal and distal L1s for each subgroup of L1s (Fig 5b).

MIWI2 E18 sense and antisense piRNAs show a highest expression toward intact L1s independent of gene location. Similarly, sense MILI E18, P0 & P10 piRNA show the same trends, intact L1s being targeted more regardless of location to gene. MILI E18 and P0 piRNA antisense expression for L1s for is much less and exhibits limited preference toward any subcategories. MILI P10 antisense piRNAs for distal and genic L1s are more abundant to intact L1s while proximal piRNA target L1s with a 5’UTR. After P10, piRNA expression toward L1s
decreases which is similar to observations made in other studies [33]. MILI adult, MIWI P14 and P20 sense piRNA expression is slightly greater toward intact L1s while antisense piRNAs are more abundant to all L1s and L1s with a 5’UTR, with an exception of MIWI P14 antisense distal and genic piRNA expression is slightly more toward intact L1s. Finally, MIWI adult sense and antisense piRNA exhibit slightly more expression toward all L1s. The L1 expression of these L1s was then analyzed (Fig 5c).

At E16.5 of distal and genic FL and intact L1s are more transcriptionally active while L1s with a 5’UTR are the most active L1s of proximally located L1s. L1 expression is significantly less at P3, however, we observe for distal L1s that FL L1s are most active while genic all L1s and proximal L1s with a 5’UTR are more transcriptionally active. Interestingly, at P10 we observe the same L1 activity trends as E16.5. At P14 L1 expression is very low for all locations and subcategories. At P20 and adult testes L1 expression exhibits similar trends to L1s at P10. For all cell-specific L1 expression intact L1s have higher expression for all locations except for spermatogonia distal L1s which FL L1s have a slightly higher expression. Antisense expression of for all sub categories is extremely low and therefore data is not shown here.
DISSCUSSION

For the first time we investigate the role of piRNA in the single lineage evolution of L1s which is observed in many different species. The current theory for L1 evolution that there is a limitation in resources provided by the host cell necessary for retrotransposition, although to date there are no identified host factors required for L1 transposition which would contribute to this type of L1 evolution [5, 15, 16]. PiRNAs have previously [7, 25] and in this study been shown to preferentially target young L1s for degradation and transcriptional silencing at the most vulnerable times during germ cell development. This exerts selective pressure on those L1s that are actively replicating (Fig 1, 2). Specifically, sense stranded piRNAs which are primarily generated from L1 transcripts indicates piRNAs are generated from autonomous L1 expression, generating an immune defense to active L1s. Additionally, we observed that the ping pong cycle is predominately active toward young L1s in subcategories all L1s, L1s with 5’UTR and FL L1s (Fig 2b). Later during germ cell development, the primary processing pathway is the dominant mechanism of piRNA generation and piRNA production is no longer dependent on transcriptionally active young L1s.

Previous studies have shown that when piRNAs target an L1 for degradation or methylation they target the 5’end of the L1 (Fig 3a) [19, 23]. While we were able to replicate this finding by mapping piRNA directly to consensus sequences, when genomically mapped piRNAs were examined it was discovered that for a majority of L1 families piRNAs preferentially targeted the 3’end (Fig 3b). However, young L1 families from E18 to P0 showed approximately equal 5’ and 3’ end piRNA expression. To explain the bimodal peaks observed in young families and the 3’ preference to older families we examined the starting position of genomic L1s copies in each of the 29 families (Sup Fig 6). It was observed that genomic L1
copies primarily start at the 5’ end or the 3’ and very few start in the middle the L1, similar results were found by [39]. For older families even fewer L1s were found to start at the 5’end. These results suggest for young families we are observing peaks at both ends due to FL L1s which are primarily being targeted at the 5’end and truncated L1s which are being target at 3’end. Older families which have less copies of FL L1s (Fig 1, Sup Fig 6) are target more at the 3’end. Because very few L1s start in the middle of the L1 we observe very little piRNA targeting this region (Sup Fig 5, 6). Additionally, we observe 3’ targeting for all 29 families for all piRNAs collected after P10 (Fig 3b). This is because the ping pong cycle is no longer active and the primary processing pathway takes over piRNA production from piRNA clusters which are usually generated from old truncated L1s. These results suggest that piRNAs are targeting the 5’ UTRs of young active L1 families specifically, at early time points during germ cell development, thereby exerting selective pressure on these L1 5’UTRs.

We further evaluated the correlation between piRNA and L1 expression toward of the 29 families. No correlations were observed between piRNA and L1 expression in whole testes from E16.5 to P10 likely because the high expression of piRNAs is extremely limiting L1 expression of any L1 family. AT P14 both piRNA and L1 expression is very low (Fig 5 b, c), however, there is a negative correlation observed. Older families, likely within piRNA clusters, are starting to be expressed more while younger families are not as transcriptionally active (data not shown). We hypothesize that these older L1s are being transcribed as a part of piRNA clusters to allow for future piRNA production via the primary processing pathway. At P20 the positive correlation between piRNA and L1 expression suggests piRNAs have been generated from these L1 transcripts which are still being transcribed to produce more piRNA. Positive correlations are observed in specific germ cells but not in whole testes expression data. This suggests piRNAs
are still being generated in these cells from piRNA clusters/primary processing pathway and may not have significant correlations because expression is being diminished from other cells. Additional work, specifically examining if there is overlap between location of piRNA clusters and old L1 families, would be required to prove that these older L1 transcripts may be a part of piRNA clusters.

Finally, we examined the location of L1 insertions relative to genes. For each of the 29 families L1s were classified as proximally, distally or genically located to a gene. The prevalence of genically located L1s compared to proximal is likely because L1s are inserting into introns which would cause no detrimental effect while regions immediately upstream and downstream of the gene are important for enhancer and repressor binding and L1 insertions could drastically damage regulation of a gene (Fig 5a). We further examined piRNA abundance to each subcategory for each location independent of age. Similar trends for sense piRNA expression regardless of location were observed. Intact L1s are their primary targets showing that piRNAs preferentially target more active (younger) L1s (Fig 5b). MIWI2 E18 antisense piRNA target L1s similar to sense piRNA regardless of location. MILI E18 and MILI P0 antisense piRNA expression significantly diminished compared to sense expression, however, is higher toward distal all L1s and proximal and genic L1s with a 5’UTR. This supports the finding in Fig 2a that shows sense piRNAs are targeting active younger L1s and antisense are more abundant toward older L1s. At MILI P10 proximal antisense piRNA expression towards L1s is higher for L1s with a 5’ UTR which is also observed for L1 expression in whole testes data at P10. This suggests that intact L1s proximally located are not as active as L1s with a 5’UTR, but not necessary FL. Expression is upregulated potentially because genes they are located nearby are also upregulated or there are more of these elements located proximally. From P14 to
adulthood piRNA and L1 expression is extremely limited however more expression for both is apparent for intact L1s.

We have shown that piRNAs preferentially target young L1s via the ping pong mechanism early in germ cell development, which thereby exerts selective pressure on these elements to avoid detection. Furthermore, the 5’UTR of FL L1s are targeted by piRNAs which explains why piRNAs change promoters when evolving. Finally we have shown that piRNAs target active L1s independent of location to a gene. While it is probable that other defense mechanisms play a role in the evolution of L1s we have shown that piRNAs are a major player in the single lineage evolution of L1s.
METHODS

L1 Genome Annotation

L1s were recently categorized into 29 different families [11]. L1s in each of the 29 families were identified in the mouse genome using RepeatMasker (http://www.repeatmasker.org/species/mm.html), RepeatLibrary v4.0.5 made available Jan 31, 2014.

Consensus Sequence Annotation

Consensus sequences used to annotate the L1s in the mouse genome were obtained from the alignment file provided by RepeatMasker (mm10.fa.align). Each of the 29 consensus sequences had been split by RepeatMasker into 3 segments (5’end, orf2, and 3’end). Each piRNA library was mapped to each of the 29 sequences. The 3 split parts of the 29 consensus sequences was combined together to generate the full consensus sequence. There was 141bp overlap between the 5’end and orf2 consensus sequences and a 296 bp overlap between orf2 and 3’ end consensus sequences for each L1 family. When we examined these overlaps there were occasional SNPs in the overlap regions. Therefore, when we combined the overlap sequences to generate the full-length consensus sequences we accounted for these SNPs by adding ambiguous bases at these positions.

For each of the 29 full consensus sequences, put together from the split consensus sequences in the mm10.align file from Repeatmasker, the location and protein sequence of ORF1 and ORF2 for each consensus was extracted using MacVector. The ~200bp monomeric promoter sequences location in each consensus sequence was discovered using the program Tandem Repeat Finder [40]. Tandem Repeat finder was able to find monomer sequences for all but four L1 family consensus sequences (L1MdFanc_I, L1MdFanc_II, L1MdF_V, and L1Lx_II).
For these cases monomeric promoter and tether sequences from related consensus sequence families were aligned to these consensus sequences using MacVector to deduce where the monomers were located. L1Lx_II consensus sequence was aligned to L1Lx_I, L1Lx_III consensus sequences to deduce where the 5’ monomeric sequences were located. Finally, consensus sequences L1MdFanc_I, L1MdFanc_II and L1MdF_V were aligned to L1MdF family consensus sequences (L1MdF_I, L1MdF_II, L1MdF_III, and L1MdF_IV) to find the monomeric sequences.

Upon examination of the RepeatMasker output genomic L1s were named based on their 3’ end annotation. Because the 5’end is responsible for the transcription of the L1 and we were interested in identifying how the 5’end evolves based on the evolutionary pressure of piRNAs we renamed L1s based on their 5’end L1 annotation. L1s that had multiple L1 5’end annotations with at least one monomeric promoter sequence was named by the L1 5’end annotation that highest Smith-waterman score. Only genomic L1s from the 29 families [11], were included in this study.

**L1 Location Relative to Genes**

We further classified L1s in each family based on their proximity to coding sequences of genes, using the gene annotation file (genes.gtf) provided with the mm10 genome by UCSC. Three categories were used: distal, proximal and intragenic (genic). L1s were first classified as genic if the L1s’ midpoints were located within the coding region of a gene, then considered proximally located if their midpoint was located within 10Kb up or downstream of a gene [7] and finally categorized as distally located if the midpoint of the L1 was greater than 10Kb up or down stream of a gene. Using this classification there could be no overlapping definitions.

**PiRNA Sequencing Datasets and Normalization**
To analyze mouse piRNA expression toward L1s we obtained multiple piRNA sequencing datasets from GEO (Sup Table 1)[30-32]. Illumina was used for sequencing for all the GEO datasets. Fastqc [41] was used evaluate quality of reads and cutadapt [42] was used to remove barcodes from the 5’ and adapters from the 3’ ends of reads. Reads with an average quality score of 20 were kept for further analysis (Table 1). Because piRNAs are longer than most small RNAs we evaluated the size of reads in the sequencing datasets to ensure piRNAs were being sequenced. Reads between 15 and 36 bases in length were counted and then the percentage of reads of each length were calculated (Sup Fig 2). Reads from each dataset were aligned to the mm10 genome using bowtie [43] allowing 0 mismatches and multiple alignments. Duplicates of reads were removed before mapping and then taken into account when calculating expression of piRNA to L1 transcripts.

PiRNAs mapping to L1s were counted using HTSeq [44]. Parameters used for HTSeq include the mode intersection_strict which indicates that a read must be completely enclosed in an L1 annotation to be counted and “no” specifies a read should be counted if it is mapping to the sense or antisense strand relative to the L1 or gene. Counts were normalized using the method similar to [7], each read was divided by the number of times it mapped to the genome and the number of reads mapping in the genome. The sum count of sense and antisense piRNAs to a family or particular subgroup of L1s was divided by the number of L1s in the family and/or subgroup.

**Consensus piRNA Mapping**

Each piRNA dataset was mapped to each consensus sequence using the alignment program Mosaik Align [45] allowing 0 mismatches. PiRNA expression toward each family was
normalized for the number of times a read mapped to the consensus sequence and the total number of reads that mapped to a consensus sequence.

*Genomic piRNA Mapping back to Consensus*

To identify if genomically mapped piRNAs target L1 copies at the 5’end as is indicated by the results of others and our piRNA consensus mapping results we examined where piRNAs are targeting the genomic L1s copies. We first binned the consensus sequences for each of the 29 families into 100bps bins. For each non-recombined with no gaps (see Results for meaning of no gaps) genomic L1 copy we determined how much of the consensus sequence was covered using RepeatMasker results. Then the genomic L1 was also binned to represent the bins in the corresponding consensus sequence. Normalized piRNA expression mapping in each bin were then summed for each family.

*Gene Expression Analysis*

To analyze expression of L1 transcripts we obtained transcriptome RNA sequencing datasets from whole testis and specific groups of germ cells at similar time points to the piRNA sequencing (Sup Table 2)[34-38]. Quality of reads were analyzed using fastqc and reads were processed by removing adapters and reads less than 25bps and with a quality score 20 were excluded from analysis. Reads for each dataset were mapped to the mm10 genome using bowtie1 [43]. Parameters for mapping included only allowing one alignment per read and number of mismatches depended on maximum length of reads for each dataset (Sup Table 2).

HTSeq was used to count reads mapping to L1s (sense or antisense) with the same parameters as utilized for counting piRNA. Counts were normalized by dividing the number of reads by the number of reads mapped to the genome. For time points that had replicates the IDR package was used to analyze reproducibility (data not shown). Because reproducibility was
extremely high we utilized the mean expression value to represent gene expression for a given gene or L1.
REFERENCES


7. Mourier T: Retrotransposon-centered analysis of piRNA targeting shows a shift from active to passive retrotransposon transcription in developing mouse testes. *BMC genomics* 2011, **12**:440.


**Fig 1: L1 Sub Classifications** Depicts number of L1s in each of the subcategories (all L1s, L1s with a 5’UTR, FL L1s and intact L1s) within the 29 families.
Fig 2: PiRNA Expression Correlated with Age (a) Heat map depicts Pearson correlation values for sense and antisense piRNA abundance correlated with age of L1 families. Negative correlation indicates younger families have more piRNA expression while positive correlation indicates older families have more piRNA expression. (b) Heat map shows Pearson correlation values for log ratio primary/secondary piRNA expression correlated with age of L1 families.
Negative correlation indicates young families have more primary piRNA while positive correlation shows that older families have more primary piRNA. * indicates pvalue < .05.
Fig 3: PiRNA Expression to 5’ versus 3’ ends of L1s (a) Shows log ratio of 5’end piRNA expression/3’ end piRNA expression for piRNA mapped directly to consensus sequences for sense (blue) and antisense (red) strands. (b) Log ratio of 5’end piRNA expression/3’ end piRNA expression for piRNA mapped to the genome back to consensus sequence for sense (blue) and antisense (red) strands.
Fig 4: Correlation between of piRNA and L1 Expression Values

Heat Map shows Pearson correlation coefficients for piRNA expression for each L1 family correlated with L1 expression for each L1 family. * indicates pvalue < .05.
Fig 5: Genetic Locations of L1s (a) Number of L1s located intragenically (genic), proximally (within 10 kb up or downstream of a gene) or distally (located greater than 10 kb from a gene) in each of the 29 families. (b) PiRNA expression to L1s in each subcategory located genic, proximal or distally to a gene. PiRNA expression for each subcategory was normalized for number of L1s in subcategory that were located genically, proximally or distally. (c) L1 expression to L1s in each subcategory located genic, proximal or distally to a gene. L1.
Sup Fig 1: FL and Intact L1 Genomic Information (a) Depicts number of L1s of a given length in each subcategory. (b) Shows number of L1s considered intact using a number of different length and similarity cutoffs. (c) Number of L1s with intact ORF1, intact ORF2, and intact ORF1 and ORF2 using a 99% length and similarity cutoff.
Sup Fig 2: Length of Reads in piRNA Sequencing Datasets Percent of reads of a given length in each dataset.
Pearson correlation coefficient = 0.63
p-value = 2e-04

Sup Fig 3: Correlation between age and L1 copy
Sup Fig 4: Correlation between log (sense/antisense) value and age of L1 families Heat map showing Pearson correlation coefficients for the log ratio of sense/antisense piRNA expression for each L1 correlated with age of L1 families. Negative correlation indicates younger families have a higher abundance of sense piRNA while a positive correlation indicates older families have a higher abundance of sense piRNA.
**Sup Fig 5: piRNA Mapping** (a) Raw graphing results for piRNA mapped to 29 consensus sequences and (b), genomically mapped piRNA mapped back to consensus sequences, one family selected from young middle and old L1s.
Sup Fig 6: Location of Start Site for Genomic L1s in each of the L1 family
## Sup Table 1: PiRNA Datasets

Table showing library size and mapping statistics for each of the 8 piRNA libraries used.

<table>
<thead>
<tr>
<th>Experiment/Time</th>
<th>GSE ID</th>
<th>Library Size</th>
<th>Mapped Reads</th>
<th>% Mapped Reads</th>
<th>Uniquely Mapped Reads</th>
<th>Uniquely Mapped Reads</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIWI2 E18</td>
<td>GSE39203</td>
<td>8,143,020</td>
<td>6,466,046</td>
<td>79.41</td>
<td>3,081,341</td>
<td>37.84</td>
<td>Xioli J. et al 2012</td>
</tr>
<tr>
<td>MILI E18</td>
<td>GSE39203</td>
<td>6,732,090</td>
<td>5,404,433</td>
<td>80.28</td>
<td>4,249,271</td>
<td>63.12</td>
<td>Xioli J. et al 2012</td>
</tr>
<tr>
<td>MILI P0</td>
<td>GSE39203</td>
<td>4,311,060</td>
<td>3,414,313</td>
<td>79.2</td>
<td>249,271</td>
<td>59.5</td>
<td>Xioli J. et al 2012</td>
</tr>
<tr>
<td>MILI P10</td>
<td>GSE39203</td>
<td>5,629,389</td>
<td>4,311,768</td>
<td>76.59</td>
<td>2,771,526</td>
<td>49.23</td>
<td>Xioli J. et al 2012</td>
</tr>
<tr>
<td>MIWI P14</td>
<td>GSE32183</td>
<td>11,885,456</td>
<td>8,308,004</td>
<td>69.9</td>
<td>7,430,012</td>
<td>62.51</td>
<td>Reuter et. al 2011</td>
</tr>
<tr>
<td>MIWI P20</td>
<td>GSE32183</td>
<td>11,899,571</td>
<td>8,308,882</td>
<td>69.83</td>
<td>7,698,356</td>
<td>64.69</td>
<td>Reuter et. al 2011</td>
</tr>
<tr>
<td>MIWI Adult</td>
<td>GSE32183</td>
<td>13,163,017</td>
<td>9,498,936</td>
<td>72.16</td>
<td>8,895,228</td>
<td>67.58</td>
<td>Reuter et. al 2011</td>
</tr>
<tr>
<td>Total Small RNA P0</td>
<td>GSE39203</td>
<td>7,134,113</td>
<td>52,581,14</td>
<td>73.7</td>
<td>2,982,453</td>
<td>41.81</td>
<td>Xioli J. et al 2012</td>
</tr>
<tr>
<td>Experiment/Time</td>
<td>GSE ID</td>
<td>Library Size</td>
<td>Mapped Reads</td>
<td>% Mapped Reads</td>
<td>Mismatches allowed</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
<td>---------------</td>
<td>---------------</td>
<td>----------------</td>
<td>-------------------</td>
<td>-------------------------</td>
<td></td>
</tr>
<tr>
<td>WT E16.5</td>
<td>GSE7747</td>
<td>24,406,691</td>
<td>20,572,798</td>
<td>84.29</td>
<td>1</td>
<td>Zamudio et. al 2015 Genes Dev</td>
<td></td>
</tr>
<tr>
<td>WT P3</td>
<td>GSE9652</td>
<td>38,937,518</td>
<td>32,724,565</td>
<td>84.04</td>
<td>1</td>
<td>Modzelewski et. al 2012 Dev Cell</td>
<td></td>
</tr>
<tr>
<td>WT P10</td>
<td>GSE7747</td>
<td>24,298,722</td>
<td>20,619,120</td>
<td>84.86</td>
<td>1</td>
<td>Zamudio et. al 2015 Genes Dev</td>
<td></td>
</tr>
<tr>
<td>WT P14.5</td>
<td>GSE44654</td>
<td>24,776,019</td>
<td>15,700,572</td>
<td>63.37</td>
<td>2</td>
<td>Li et. al 2013 Mol Cell</td>
<td></td>
</tr>
<tr>
<td>WT P20</td>
<td>GSE7747</td>
<td>90,964,128</td>
<td>69,025,241</td>
<td>75.88</td>
<td>3</td>
<td>Zamudio et. al 2015 Genes Dev</td>
<td></td>
</tr>
<tr>
<td>WT Adult R1</td>
<td>GSE67169</td>
<td>15,587,995</td>
<td>11,369,899</td>
<td>72.94</td>
<td>0</td>
<td>Endo et. al 2015 Proc Nat Acad Sci</td>
<td></td>
</tr>
<tr>
<td>WT Adult R2</td>
<td>GSE67169</td>
<td>18,209,692</td>
<td>13,338,765</td>
<td>73.25</td>
<td>0</td>
<td>Endo et. al 2015 Proc Nat Acad Sci</td>
<td></td>
</tr>
<tr>
<td>WT Adult R3</td>
<td>GSE67169</td>
<td>13,317,023</td>
<td>11,093,938</td>
<td>83.31</td>
<td>0</td>
<td>Endo et. al 2015 Proc Nat Acad Sci</td>
<td></td>
</tr>
<tr>
<td>CS Spermatogonia 6dp R2</td>
<td>GSE43717</td>
<td>32,736,287</td>
<td>19,634,561</td>
<td>59.98</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatogonia 6dp R3</td>
<td>GSE43717</td>
<td>38,579,378</td>
<td>23,893,109</td>
<td>61.93</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatocytes Adult R1</td>
<td>GSE43717</td>
<td>23,945,985</td>
<td>13,187,009</td>
<td>55.07</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatocytes Adult R2</td>
<td>GSE43717</td>
<td>33,240,606</td>
<td>18,327,174</td>
<td>55.13</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatocytes R3</td>
<td>GSE43717</td>
<td>41,475,288</td>
<td>23,356,876</td>
<td>56.32</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatids Adult R1</td>
<td>GSE43717</td>
<td>20,110,262</td>
<td>11,372,885</td>
<td>56.55</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatids Adult R2</td>
<td>GSE43717</td>
<td>35,347,416</td>
<td>19,030,786</td>
<td>53.84</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatids Adult R3</td>
<td>GSE43717</td>
<td>37,616,883</td>
<td>21,398,329</td>
<td>56.88</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatocytes R1</td>
<td>GSE43717</td>
<td>30,683,403</td>
<td>21,156,114</td>
<td>68.95</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatocytes R2</td>
<td>GSE43717</td>
<td>32,517,036</td>
<td>21,558,231</td>
<td>66.30</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
<tr>
<td>CS Spermatocytes R3</td>
<td>GSE43717</td>
<td>38,460,282</td>
<td>26,015,594</td>
<td>67.64</td>
<td>2</td>
<td>Soumillon et. al 2013 Cell Rep</td>
<td></td>
</tr>
</tbody>
</table>
Sup Table 2: RNA Transcriptome Sequencing Datasets Table showing library size and bowtie mapping parameter (mismatch) for each of the gene expression sets used.
CHAPTER 4

CONCLUSIONS AND FUTURE DIRECTION
CONCLUSIONS AND FUTURE DIRECTIONS

Bioinformatics is the integration of many different fields (biology, mathematics, statistics and computer science) to investigate a biological system. This dissertation used bioinformatics to examine the metabolic regulation of germ cell differentiation during the first wave of spermatogenesis and the role piRNA suppression of L1s in germ cells is exerting on L1 evolution. Both piRNA suppression of L1s and metabolism are essential for differentiation of spermatogonia into sperm, in process termed spermatogenesis. Additionally, their complexities and intricacies make the use bioinformatics essential in order to gain a comprehensive understanding of these processes.

Chapter 2

Chapter 2 of this dissertation investigated the role spermatogonia metabolic activity may exert in controlling differentiation. Metabolic activity was analyzed using flux balance analysis (FBA). FBA is a bioinformatics approach which can be used to analyze the entire metabolism of an organism, tissue or cell of interest (Orth, et al., 2010). Metabolism in spermatogonia was modeled at 8 time points over the first 8 days of the first wave of spermatogenesis. Spermatogonia cell-specific gene expression obtained utilizing the WIN18,446/RA protocol, corresponding to 8 time points modeled, was first utilized to construct a spermatogonia cell specific metabolic network (Evans, et al., 2014). This network consisted of metabolic pathways which were determined to be potentially active, if there was an enrichment of enzymes in the pathway considered to be highly expressed (expression value above the median), at any of the 8 time points modeled. Additionally, curated vitamin A metabolic network was generated by consulting numerous metabolic databases, such as Recon1, Recon2, EHMN, Reactome, and
BioCyc, and integrated into the spermatogonia cell specific network (Caspi, et al., 2012; Croft, et al., 2011; Duarte, et al., 2007; Hao, et al., 2010; Thiele, et al., 2013).

Using this network and integrating gene expression data, metabolic activity was simulated over the first wave spermatogenesis at 8 distinct times over the first wave of spermatogenesis at 0, 4, 12 hours (hrs), 1, 2, 4, 6, 8 day. Reactions contributing to retinoic acid (RA) production and degradation were analyzed to their contribution to the RA pulse (Hogarth, et al., 2015). We observed that three reactions provide a majority of RA: 1) extracellular RA import into the cell, 2) conversion of CRBP (cellular retinol binding protein)-retinal to RA and 3) release of RA from binding protein CRABP (cellular retinoic acid binding protein). RA degradation was minimal within the cell, CYP26 catalyzed reactions and CRABP binding of RA contributing minimally to the degradation of RA. Freely available RA was evaluated in the cell and found to exhibit a peak at 4 and 12hrs and 8 days during the first wave. The peak observed at 4 and 12 hrs is likely due to the injection of RA utilized in the WIN 18,446/RA synchrony protocol (Evans, et al., 2014). The observed peak of freely available RA at day 8 suggests that our model is able to replicate the RA pulse (Hogarth, et al., 2015). Additionally, these results suggest that apart from internal metabolism of RA and release of RA from CRABP, another source other than the spermatogonia cells are responsible for RA production. This source is hypothesized to be the Sertoli cell since a number of knockout studies have shown that impairing vitamin A metabolism in these cells hinders spermatogenesis (Raverdeau, et al., 2012; Tong, et al., 2013). In-silico knockouts of reactions and enzymes were performed in the vitamin A pathway to further investigate the role reactions and enzymes in spermatogonia may play in the regulation of RA.
Knockout studies predicted that one gene and six reactions would cause significant disruption to the entire flux of the metabolic pathway and freely available RA. *Lipe* at 4 hrs catalyzes 2 reactions in our network, retinyl-ester to retinol CRBP-retinyl-ester to CRBP-retinol. Reducing efficiency of these reactions causes a disruption in other reactions. Previous studies had shown that global ablation of *Lipe* caused spermatogenic defects resulting in male infertility (Casado, et al., 2012; Chung, et al., 2001; Hermo, et al., 2008). In these studies it was hypothesized that LIPE activity was required for membrane integrity, however, here we show that LIPE activity may also be essential for regulating RA availability. Reaction knockouts of retinol to CRBP-retinol and CRBP-retinol to CRBP-retinal at 4 hours and 4 days respectively also caused an increase in freely available RA. These reactions although they do not directly contribute to RA production or degradation are important for overall vitamin A metabolic homeostasis. Knockouts of CRBP-retinal to RA at 4 hrs and RA transport reaction at 2 days caused a disruption in the network flux and a decrease in RA availability. Since these reactions are major contributors to RA at these time points it is unsurprising there disruption would be predicted to cause a spermatogenic defect.

The activities of other metabolic pathways in our network were examined. The tricarboxylic acid (TCA) cycle was found to be the most active pathway followed by nucleotide interconversion. Folate, cholesterol, fatty acid synthesis, degradation, glycolysis and gluconeogenesis all have steady activity while keratan sulfate synthesis, degradation and N-glycan synthesis all have no activity across the first wave. Of the pathways which exhibited metabolic activity fatty acid synthesis and degradation produced the most energy, which is essential for growth and development of these cells. Interestingly, coincided with times fatty acid synthesis and degradation had highest energy production levels when spermatogonia
differentiation occurs suggesting these pathways may play an indirect role in regulating differentiation (Hogarth, et al., 2015).

Finally, the activity of anaerobic glycolysis activity, conversion of pyruvate to lactate, was evaluated. Because glycolysis activity was determined to be low and TCA activity was high in differentiating spermatogonia it was examined if pyruvate was being provided by an extracellular source. It was observed that pyruvate was being transported into the cell suggesting that pyruvate is being generated by glycolytic activity from another cell, which was hypothesized to be the Sertoli cells due to previous studies which showed pyruvate was provided to later germ cells, spermatocytes and spermatids, via Sertoli cells (Alves, et al., 2013). Pyruvate was converted into lactate spermatogonia cell, however this activity decreased as cells differentiated, suggesting that differentiating spermatogonia follow the same metabolic trends as other differentiating cells (Folmes, et al., 2012; Shyh-Chang, et al., 2011; Zhang, et al., 2011).

Overall, the research in chapter 2 showed that metabolic activity in spermatogonia is essential for successful differentiation and development of these cells. Vitamin A metabolism, while not responsible for a majority of RA production or degradation, is essential for differentiation of spermatogonia, suggested by our in-silico reactions and enzymes knockouts within spermatogonia contribute to the observed RA pulse (Hogarth, et al., 2015; Tong, et al., 2013). Additionally, activities of central metabolic pathways were found to be essential for differentiation. However, because many essential metabolites were indicated to be provided by the extracellular environment, other cells within the testis should be modeled to elucidate their role in regulating the RA pulse and providing other essential metabolites to spermatogonia.

Metabolism of other cells within the testis likely providing crucial metabolic support to spermatogonia including Sertoli cells and peritubular myoid cells should be modeled. Sertoli
cells are reported to provide essential metabolites RA, pyruvate and lactate to spermatocytes and spermatids (Bajpai, et al., 1998; Grootegoed, et al., 1984; Meinhardt, et al., 1999; Nakamura, et al., 1984; Raverdeau, et al., 2012; Robinson and Fritz, 1981; Tong, et al., 2013). Furthermore, the vitamin A metabolic activity in peritubular myoid cells is thought to create a metabolic barrier which prevents circulating RA from entering the seminiferous epithelium (Vernet, et al., 2006). This is predicted to allow cells within the seminiferous tubule to have complete control of RA levels thereby ensuring that differentiation happens successfully. Finally, the Lipe knockout which was predicted to cause defects in spermatogenesis due to alteration in RA availability should be validated. To validate this finding RA levels should be measured in Lipe knockout mice (Chung, et al., 2001) to see if infertility may also be caused by miss-regulation of the RA pulse. Although work in chapter 2 has increased our understanding of the role metabolism exerts in spermatogonia differentiation, work is still necessary to fully comprehend the scope metabolism plays in regulating spermatogenesis and ensuring male fertility.

Chapter 3

Chapter 3 focused on identifying the role piRNAs may exert in controlling the evolution of long interspersed element-1 (L1s). L1s have evolved in mammals in a single lineage manner for reasons which are still unclear. Using publically available L1 annotations in the mouse genome L1s in the reclassified 29 L1 families (Sookdeo, et al., 2013) were subcategorized into 4 inclusive categories, all L1s, L1s with a functional 5’ promoter (L1s with a 5’UTR), L1s with a 5’UTR and full-length (FL) (referred to as FL L1s) and L1s with a 5’UTR, FL and intact (intact L1s) meaning they have a functional ORF1 and ORF2. From these classifications we observed that younger families have more FL and intact L1s making them a greater threat to genome integrity. Older L1s although they have more copies are truncated at the 3’ or 5’ ends making
them inactive. Because younger L1s have more FL and intact L1s it is expected that piRNAs selectively act on these families. We next examined the expression of piRNA toward L1s in each family and subcategory to understand if piRNAs preferentially target young (potentially active) L1s.

We observed that sense stranded piRNA exhibited more expression toward young L1s. Sense stranded piRNAs indicate that the piRNA was generated from the L1 transcripts therefore suggesting piRNA abundance is dependent on L1 expression. To further support this idea we examined whether the ping pong cycle was more active toward younger families. The ping pong cycle which uses degraded L1 transcripts to generate more piRNA is active early during germ cell development. From E18 to P0 we observed that the ping pong mechanism of piRNA generation was active dominantly toward younger L1s. This activity suggests that an “immune” like defense to active younger L1s thereby exerting selective pressure on these L1s to evolve in order to evade detection.

Additionally, it was shown that piRNAs preferentially target the 5’UTR of L1s (Aravin, et al., 2007). We wanted to investigate if this preferentiality was consistent across all ages of L1s and with genomically mapped piRNA. Using genomically mapped piRNA is a more powerful method because we obtain a more comprehensive and detailed picture of how piRNAs are targeting genomic L1s. In mapping piRNAs mapped to the genome back to the consensus sequence, we observed in young families 2 peaks of piRNA expression at the 5’ end and 3’ end different than what was seen by directly mapping piRNA to consensus sequences. In this case, a significantly larger peak of piRNA expression is observed at the 5’end. To explain the observations of biomodal peaks of genomically mapped piRNAs back to the consensus sequence we examined the location of transcription start sites (TSS) of genomic L1s for each family and
TSS start at the very 5’ end or 3’ end of the L1. This suggests that the 5’ UTR peak of piRNA expression is from L1s that are FL while the 3’ is generated from truncated L1s. Because middle aged and old L1 families have fewer FL L1s there is less piRNA expression at the 5’ end. Additionally, because very few L1s start in the middle of the consensus sequence we also observe less expression toward the middle of the consensus sequence. This data supports the idea that piRNAs target the 5’ UTR of active L1s thereby forcing them to acquire new L1s to evade suppression.

Finally, piRNA and L1 expression were examined to elucidate if there was a difference in expression based on their location relative to genes. L1s were predominantly distally and genically located to genes. Only a small number of L1s for each family were located proximally. Because the ratio of distal, genic and proximal L1s was about the same for each of the 29 families piRNA and L1 expression was analyzed of the 4 subcategories for each of the 3 locations independent of age. We observed that throughout germ cell development piRNA and L1 expression was greater for intact L1s independent of location. The exception is antisense piRNA and L1 expression toward proximally located L1s which is greater toward L1s with a 5’ UTR. Because this is not observed for FL or intact L1s this suggest L1s that are located proximally are 3’ truncated meaning that the promoter region of the L1s are located up or downstream of a gene. This supports the idea that L1s insertions may act as promoters for genes and some insertions may have benefited genome evolution (Zamudio and Bourc’his, 2010).

Our work has provided evidence to suggest piRNAs are a contributing factor in the single lineage evolution of L1s. Additionally, piRNAs exert selective pressure on the 5’ end of L1s thereby forcing L1s to change promoter sequences to evade detection. In order to definitively prove that L1s target the 5’ UTR thereby suppressing L1 activity an L1 transgene with a novel 5’
UTR sequence should be integrated into the genome of mice to investigate how it is targeted by piRNAs. If our theory is correct, due to the novel 5’ UTR and low copy number, piRNAs would not recognize the promoter therefore it would escape methylation repression and insertions from the transgene would be observed.

**Final Conclusions**

The work in this dissertation has investigated two molecular processes required for spermatogenesis. Using bioinformatics the metabolic activity in spermatogonia and its role in the initial differentiation was understood. Furthermore, piRNAs were determined to exert controlling evolution of L1s. The knowledge obtained from the studies in this dissertation has increased our understanding of mammalian male reproduction and can help address the need for male fertility treatment and contraceptive development.
REFERENCES


Raverdeau, M., et al. (2012) Retinoic acid induces Sertoli cell paracrine signals for spermatogonia differentiation but cell autonomously drives spermatocyte meiosis,


