ROLE OF EPIGENETIC REGULATION IN HUMAN SPERM AND EGG

ANALYSIS OF DNA METHYLATION AND POLYA REGULATION

by

Christian Pflüger

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Oncological Sciences

The University of Utah

August 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Christian Pflüger**

has been approved by the following supervisory committee members:

Bradley R. Cairns	, Chair	**5.1.2015** Date Approved
Katharine S. Ullman	, Member	**5.1.2015** Date Approved
Christopher T. Gregg	, Member	**5.1.2015** Date Approved
Katherine E. Varley	, Member	**5.1.2015** Date Approved
Jason Gertz	, Member	**5.1.2015** Date Approved

and by **Bradley R. Cairns**, Chair/Dean of the Department/College/School of **Oncological Sciences**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Our work focused on how germ cell DNA is packaged and if it is poised by distinctive chromatin to influence early embryo development. We studied human male and female germ cells and profiled their epigenetic repertoire. Further, we asked the question, is misregulation of that poising a common theme observed in infertility and aging in sperm? We pursued one part of this question in Chapter 2 where we investigated the changes in DNA methylation in sperm during the natural process of aging. We analyzed sperm from aged-matched donors and found that sperm on a global level gained DNA methylation over time, but notably, specific regions associated with genes linked to neuropsychiatric disorders significantly lost DNA methylation in sperm from aged donors. This raised the intriguing question if the increase in observed neuropsychiatric disorders in offspring from older fathers could be linked to the loss of DNA methylation in sperm of specific genes also implicated in neuropsychiatric disorders. Further, we asked if abnormalities in chromatin are seen in sperm from infertile patients. We tested seven imprinted regions in oligozoospermic and abnormal protamine patients and found a correlation of DNA methylation abnormalities at the promoters of these genes in infertile sperm. Our findings from this study are summarized in a publication detailed in Appendix A. Finally, we asked if changes in polyadenylation of transcripts in human oocytes could give us an insight into oocyte maturation and early human embryo development. We developed a novel bioinformatics tool called PANDA (Chapter 3) that enabled the analysis of relative polyA changes in transcripts between two different conditions. One of the advantages of this method is that it can utilize any

RNA-seq dataset as long as it has not been biased for PolyA selection. Hence, we applied PANDA to the existing RNA-seq dataset, generated by our lab from human oocytes and early embryos, and investigated PolyA changes (Chapter 4). Strikingly, we were able to identify transcripts in humans that were previously reported to gain polyA in oocytes of drosophila and xenopus. We also identified completely novel transcripts previously unknown to gain polyA during oocyte maturation. Hence, we are the first group to identify these transcripts in human oocytes and embryos. Together we hope that by studying the epigenetic setup of germ cells we have gained more insight into understanding the regulation of programs critical in early embryo development.

“The true sign of intelligence is not knowledge but imagination.” - Albert Einstein

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

Thinking back about my years in graduate school, I come to the realization that I experienced some of the best and toughest times in my life. Getting a PhD is a most humbling experience but also opened up my mind to truly outstanding science, phenomenal colleagues, loving friends and, most of all, a great mentor and human being, Brad Cairns. Brad has always been patient with me and allowed me to explore and further my PhD career. There were significant setbacks with projects failing or collaborations not working out. However, he always kept his faith in me and encouraged me to move on and find other great scientific projects to be part of. For all of that and so much more, I say thanks a million Brad.

Along with Brad comes a phenomenal lab in which to work and I am truly blessed to be part of it. I had the pleasure to work with some of the brightest people I know of and on top of it, to form bonds and friendships that I believe will be life lasting. Something that I cannot emphasize enough is the fact that the Cairns' lab is like a family. We work together and support each other, making it a sincerely collaborative environment I had the pleasure to be part of. Alisha is hands-down the best lab manager I have ever met. Her drive for efficiency and support for every one in the lab is unrivaled and very much appeals to my German soul. Maggie is one of the most skilled biochemists you can think of and you bet she can bake and make jams like no one else. Thanks for always sharing with your 'prosumer'. Cedric is one of the smartest and most insightful people I have met, even though he is from France. He also taught me that French people are more pessimistic than Germans and I had to fly all the way to Utah and do my PhD to actually learn that.

Thanks for all the great talks! Some of my closest friends I have made are from the Cairns lab. Ravi is my partner in crime and we have spent countless hours furthering our bioinformatics skills, played badminton and got mocked for it (goose feather vs. duck feather), and we spent a lot of days exploring Utah. Thanks for all the unforgettable memories. Patrick and Pete are the newest additions to the Cairns lab and it has been nothing short of a blast to work, discuss, and party together with them. They are the reason that I finally learned how to ski and Patrick and his wife Kristin introduced me to the craft of home brewing. I am forever thankful for all the memories made. Tim Parnell is a tremendously skilled bioinformatician and a great friend. I have learned so much about Perl programming and computer administration from him; he should have charged me for it. I cannot thank him enough for all the things he taught me. Archana, you are the best bench mate someone can ask for, tolerating all my nonsense and being a barista's dream apprentice. I'm running out of space but I want to thank everyone else in the Cairns Lab for an amazing experience. Simon has made me laugh so much and often I might have pulled a muscle. Thanks for your amazing friendship and all the outstanding Oregon beers you kindly bootlegged for us. Jeff and Jaynie are not only the world's best landlords, but they are also the kindest and best friends you can hope for. We had tons of fun and great adventures, especially at our Moab wedding. Kian has introduced me to the fine art of sushi creation as well as mountain biking in beautiful Utah. This made a huge difference in getting through this dissertation. Also big thanks to Krista-Joon for always being there for me and lending us your adorable doggies, Wefos and Gracie, when I needed some doggie time. My longest and best friends from Germany, Nicole, Ralph and Artjom - I am eternally grateful to have you guys in my life and to show me what true friendship means, even if we are thousands of miles apart. I also cannot thank enough my in-laws, Ashok, Jyotika and Chhaya. You guys took me in and supported me as if I

am your own son or brother. Words cannot express how happy you guys make me. My late grandparents Omi Hannchen und Opi Claus and my paternal grandparents Omi Diddy and Opi Dad, I love you guys so much. You have always been there for me and supported your grandson going through all the education in the world (literally). My beautiful sister Ise and my adorable nephew and nieces gave me a lot of strength to get through my PhD. My step-dad Rene has visited me twice, drank all my home brewed beer and offered to be my gardener, provided I would pay for him. Thanks for always making me laugh and giving me all your support. I'm tremendously grateful for Sabine and her unshakable support during my journey in graduate school - thanks a million for always believing in me. My dad taught me to think critically and inspired me to apply my mind to tough problems. He has been a true inspiration to me and always encouraged me to get things right. I am enormously thankful for all the things he has done for me. My mom has given me so much emotional support and taught me how to deal with the toughest of situations. I do not think I would have made it through my thesis without her support. Thanks mom - I love you so much!

Finally, the love of my live, my best friend and the most beautiful and smart wife imaginable inside and out - words cannot describe how much strength, support and love you give me. I honestly would have no idea how to get everything done without your help and encouragement. Thanks for getting me through this dissertation in lab as a colleague as well as outside as a partner. Cheers to two Dr. Pflügers in the house!

CHAPTER 1

INTRODUCTION

1.1 Epigenetic Information: A Means to Give Instructions to the Genome

Genetic information is stored in the form of DNA in every cell in the nucleus. In order to store, organize and orchestrate the use of genetic material, DNA is packaged in a higher order structure called chromatin. Chromatin entails DNA that is wrapped around histone proteins[1] allowing for compaction of roughly 2 meters of DNA into a single 2 µm nucleus for each cell.

Epigenetics is defined as components in the cell that are regulating the usage of genetic material and are thought to be instructive as well as heritable. As such, any modification on the DNA itself as well as proteins binding to the DNA and their modification thereof can be considered epigenetic factors[2,3]. Chapter 2 of this dissertation details the epigenetic modification of cytosine methylation changes in context of aging and infertility in human sperm.

In Chapters 3 and 4, polyadenylation changes in maturing human oocytes and early human embryos were investigated. Polyadenylation at RNA transcripts is not classically considered an epigenetic modification. However, during a time point when the genome is inactive, RNA loaded into cell needs to be regulated and instruct and time the genome reactivation. One proposed mechanism involved in regulating the reactivation of the genome is the maturation of transcripts and proteins necessary for this process. Hence, cytosolic polyadenylation of transcripts in the late oocyte development is of key interest in mediating early embryo development. In very broad terms, cytosolic polyadenylation can also be regarded as an extension of epigenetic regulation. This thesis describes a novel approach at investigating polyA changes in maturing human oocytes and early human embryos. The findings are detailed in Chapter 3 and Chapter 4 of this dissertation.

1.2 Histone and DNA interactions

Histones are key components in epigenetics, where the DNA wraps around these proteins providing tails that can be modified with small chemical groups such as a methyl group or an acetyl group. Each histone contains two tetramer subunits, comprised of Histone H2A, H2B, H3 and H4 forming a functional octamer complex[4,5]. Once histones are assembled and DNA is wrapped around it, the field names this functional unit a nucleosome. Regulation and hence instruction to DNA is mediated by chromatin remodeling, histone modifications and histone variant incorporation[6]. Nucleosomes can be moved by chromatin remodelers along the DNA, or alternatively, can be ejected to provide access to binding sites by transcription factors for example.

Transcription factors are DNA binding proteins that recognize sequence motifs on the DNA and provide platforms to enable transcription. Chromatin modifiers such as histone methyl transferases, for example SET1 H3K4 methyl-transferase, mediate this process. On the flip side of transcriptional activation, transcriptional repression is also a key element in regulating mRNA abundance. Hence, enzymes and proteins are needed to remove an activating mark such as H3K4me3. Two prominent mechanisms have been demonstrated in removing H3K4me3. One mechanism involves a H3K4 demethylase such as MLL1/2 that can remove the methyl groups at the H3K4 residue in a stepwise process, reducing tri-methyl to di-methyl, mono-methyl and eventually unmethylated state of H3K4. Another mechanism involves a histone chaperone or a nucleosome remodeler, removing or ejecting the histone with its modification from the DNA thus enabling an unmodified histone to take the place of the removed histone. Transcriptional repression is not only mediated by histone modifications but also by DNA modifications[7].

1.3 DNA Methylation

Besides modifications to histone tails, chemical modifications to the DNA are also well established in regulating and maintaining cellular identity. One of the most predominant modifications of DNA in vertebrate animals and plants is DNA methylation at the 5th carbon of the cytosine base (DNAme). The coordinated placement of DNAme in vertebrates is mediated by enzymes called DNA methyltransferases (DNMTs). DNMTs are distinguished into two important subclasses in the cell. DNMT1 plays the role of the maintenance DNA methyltransferase that recognizes hemi-methylated DNA after replication and ensures the methylation of the cytosine in the newly synthesized DNA strand. DNMT1 is thought to drive the majority of DNA methylation in vertebrate cells. The second subclass of DNMTs are considered to be *de novo* DNA methyltransferases, DNMT3a and DNMT3b. As the name of the subclass suggests, DNTM3s place cytosine methylation onto unmethylated CpG dinucleotides as well as in the sequence context of CHH or CHG[8,9].

It has been shown that DNAme is strongly correlated with inactive parts of the genome and that promoter methylation is incompatible with transcriptional activity[10-12]. A notable exception has been recently published from our lab in adult germ stem cells[13] where gametogenesis genes can be transcribed in the presence of DNA methylated promoters, but it is considered to be a notable exception to the rule. Another important aspect of DNAme is that it is inheritable and may instruct the accessibility to the genetic information for the transcription machinery[14-16]. The property of epigenetic information to regulate cellular functions such as proliferation, differentiation and inheritance illustrates its importance and explains the broad interest in studying its function. Inappropriate alterations and wrongful deposition, or a lack thereof, of epigenetic marks such as DNAme may lead to diseases such as cancer[17] and pose a risk for developmental disorders[18]. This

thesis work explores how DNAme is altered in aging human sperm in infertile patients and how sperm samples should be thoroughly and robustly handled to eliminate any somatic cell contamination from the DNAme analysis.

1.4 DNA Methylation in Context of Germ Cells

DNA methylation plays an important role in compacting and organizing the genome. Somatic cells reportedly have bulk CpG DNA methylation levels of ~70% of all cytosines. In contrast, sperm is even more hypermethylated, averaging ~90% CpG methylation in the genome. The contrast of DNAme mesas versus canyons in the sperm genome is quite remarkable and has been extensively discussed before[19]. The comparatively few hypomethylated CpGs in the sperm genome can be found at poised promoters and imprinted loci of developmentally important genes[19]. It has been reported that DNAme aberrations at imprinted loci were detected in sperm from infertile patients[20-23]. Part of my thesis work was my involvement in writing software to identify aberrant loci of DNAme in infertile patients[23].

Another key interest in the Carrell and Cairns lab is the effect of aging on the DNAme landscape in germ cells. As mentioned before, the DNA of sperm is very hypermethylated. We were interested in the question, what happens to the distinct DNAme landscape in human sperm of aging donors? In collaboration with Tim Jenkins from the Carrell lab, we had the unique opportunity to investigate DNAme changes of sperm during the process of aging in healthy human donors. Importantly, we were able to age-match sperm from donors as well as follow sperm from individuals over an extended period of time (8-15 years). In summary, sperm kept the overall very defined and stable DNAme landscape with increased age; however, DNA hypomethylation was observed in aged sperm at interesting gene candidates that have been implicated in neuropsychiatric disorders[24,25]. In the light of recent publications detailing sperm's contribution to early embryo development

other than just delivering the appropriate amount of DNA, epigenetic modifications such as DNAme are gaining more and more interest as potential drivers for early embryo development[19,26]. In addition, recent reports suggest that hypomethylated regions with high CpG density also appear to drive nucleosome retention[26]. It is worth mentioning that human sperm exchanges the majority (>85%) of its histones with protamines. This exchange is thought to allow the sperm genome to be extremely tightly compacted[27,28]. Notably, the histones that are retained in the sperm genome exhibit histone modifications and are placed at regions in the genome that lack DNAme. This in turn raises the question of whether the changes observed in DNAme in aging sperm are associated with aberrant deposition or retention of histones at these regions. Taken together, our data strongly support the hypothesis that the sperm epigenome is not only well suited to facilitate mature sperm function, but that it may also contribute to events beyond fertilization[29]. My contributions to this publication were the analysis of DNAme changes during this study and the development of software to study sperm subpopulations. These findings were published in the journal PLOS Genetics[29] and are included in this dissertation as Chapter 2.

Finally, large retrospective epidemiological studies of fathers experiencing famine recently suggested that alterations to the sperm epigenome might have an effect on disease frequency with late onsets such as heart disease and diabetes[30,31]. Interestingly, these findings are supported by animal models suggesting a link of environmental influences onto the epigenome of sperm[32]. In summary, the sperm epigenome plays an important part in early embryo development as well as late onset diseases. It has the potential to confer properties for transgenerational inheritance and as such is of great importance and needs to be studied in great detail.

1.5 DNA Demethylation

As mentioned previously, mature sperm cells are hypermethylated in comparison to most somatic cells. In contrast, oocytes are hypomethylated in comparison to both sperm cells and somatic cells with an average CpG methylation of ~40%. Upon fertilization, both the maternal and the paternal pro-nucleus lose global DNA methylation levels (Figure 1.1). A notable difference is the active DNA demethylation in the paternal pro-nucleus in mammals versus passive DNA demethylation in the maternal pro-nucleus. The active DNA demethylation is mediated through ten-eleven enzymes (Tet enzymes) that oxidize the 5mC to 5-hydroxymethylcytosine (5hmC, Figure 1.2). The modified cytosine, 5hmC, can then be further oxidized to 5-formyl-cytosine (5fC) and 5-carboxyl-cytosine (5caC). The latter two cytosine modifications are known targets of TDG (Thymine-DNA glycosylase), which can recognize both 5fC and 5caC as a mismatch with guanosine on the opposite strand. TDG will cleave the base, leaving an abasic site for the base excision repair machinery (BER) to fix. BER is shown to repair the abasic site with a cytosine, completing the complicated cycle of 5mC removal. It is important to point out that active induced deaminase (AICDA/AID), an enzyme most studied in context of B-cell maturation, was also shown to remove 5mC on a global level in zebrafish. While AICDA is not essential and most likely not the preferred for the cell to remove 5mC from the genome, it is worth mentioning that AICDA is overexpressed in a lot of cancers that also exhibit a DNA hypomethylated phenotype. AICDA functions as a cytosine deaminase, removing the amine group at the fourth carbon in cytosine, effectively converting 5mC to thymine. This in turn will also create a T:G mismatch which can be fixed by either TDG or MBD4 (Methyl-binding-domain containing protein 4) by excising the thymine base creating an abasic site as a result. As mentioned before, this abasic site is recognized by the BER machinery, which targets the abasic site and replaces it with a cytosine

(Figure 1.2).

1.6 Protection of the Maternal Pro-nucleus from Active DNA Demethylation

A key task for the zygote is to ensure that active DNA demethylation primarily targets the paternal genome. The maternal genome contains imprinted loci, regions of regulatory DNA elements that are hypermethylated in the maternal genome and hypomethylated in the paternal genome (Figure 1.3). A protein called DPPA3/Stella/PGC7 has been shown to bind these maternally imprinted loci by binding to histones marked by H3K9me2 histone modification. DPPA3binding prevents Tet3, the active DNA demethylase, from binding and demethylating. It has been postulated that the maternal genome is only targeted by Stella since the paternal genome by and large lacks histones due to its packaging by protamines. Remarkably, we found DPPA3 to be in the top 25 transcripts that receive the most polyadenylation during human oocyte maturation. This finding is described in more detail in Chapter 4 of this dissertation.

1.7 DNA Methylation and Transgenerational Inheritance

The previous section already touched upon the notion that there is a strong interplay of histones with DNA and DNA methylation. In fact, non-coding RNAs also play and important role in deposition and maintenance of DNAme. As seen in Figure 1.1, DNA demethylation is significantly reduced upon fertilization in both the maternal and paternal genome. It is, however, reestablished after the embryo implants into the uterus wall of its mother in the epiblast stage (~E6.5). A second wave of DNA demethylation was shown to occur during the maturation of primordial germ cells (PGCs) to mature germ cells such as oocytes and sperm cells. Consequently, any changes in DNAme that persist in the offspring germ

cells need to survive two rounds of DNA demethylation. A possible mechanism by which “memory” of DNA methylation status is thought to occur comprises a combination of histone modifications and non-coding RNA. For example, histone H3K9me3 modifications are strongly correlated with DNAme. Proteins such as HP1 (Heterochromatin Protein 1) can both bind DNAme with the interaction of MeCP2 and recruit SUV39H1, a histone H3K9 methyltransferase. Non-coding RNAs (ncRNAs) have been shown to either directly recruit DNA methyltransferases to their site of expression[33] or to recruit repressive histone modifications to sites of repression. For example, it is well recognized in the field that ncRNAs from loci such as XIST or Hox can recruit the PRC2 machinery[34,35]. The PRC2 complex is responsible for depositing H3K27me3, a histone mark also associated with repressive chromatin. Targeting H3K9me3 and, by association, DNAme to imprinted loci, ncRNAs from the Igf2r and Kcnq1 loci have been shown to recruit EHMT2, a H3K9me3 methyltransferase[36-38], to the site of imprinting. In summary, regulation of DNA methylation is part of complex network of regulation, involving not only enzymes directly responsible for DNAme deposition but also ncRNA and histone modification pathways. Chapter 2 and the Appendix B in this thesis touch upon the DNAme in sperm and their regulation. Notably, we also set out to test the robustness of DNAme in germ cells by purposefully altering their DNAme levels with a drug called 5-aza-cytidine (5azaC). 5azaC blocks DNMT function by irreversible binding, rendering DNMTs inactive once engaged to 5azaC. We plan to test transgenerational inheritance of altered DNAme levels but the experiments related to this story are not finished as of yet.

1.8 Polyadenylation of RNA in Cells

A fundamental process in the cell is transcription, which copies genes in the DNA, creating RNA molecules that may be used as templates for protein synthesis. Importantly, the stability and hence the half life of these copies need to be tightly regulated in order to ensure turn over and allow for adaptations to a changing environment [39,40]. The regulation of RNA stability and, in case of mRNA, the translation of such into protein is considered a posttranscriptional process. A key posttranscriptional change at RNAs involves a process known as polyadenylation (polyA) that extends about 50-300 adenine nucleotides at the 3' prime end of the transcript. PolyA changes at mRNAs have been implicated to stabilize transcripts[41,42] as well as promote translational efficiency[43,44]. It has been shown that the default mode for RNA biogenesis is the co-transcriptionally polyadenylation in the nucleus[45]. However, a notable exception in development of maturing oocytes is the mechanism of cytosolic polyadenylation[46-49].

Briefly, transcripts that are used for late stage oogenesis or in the early developing embryo, a period in time where the genome is transcriptionally inactive and no new RNA transcripts are generated (Figure 1.4), are shown to skip nuclear polyadenylation and instead gain polyA in the cytosol. The regulatory element involved in mediating cytosolic polyadenylation, CPE (cytosolic polyadenylation element), has been shown to be part of the primary RNA sequence[50-52]. Further, binding proteins called CPEBs (cytosolic polyadenylation element binding protein) can recognize the CPE allowing for accurate timing of cytosolic polyadenylation upon their phosphorylation[47,50,51,53,54].

Most notably, the physiological process of cytosolic polyA gain has an essential role during the maturation of oocytes and in early embryos, before embryonic genome activation (EGA). EGA is the key point in embryonic development, when the genome in the embryo becomes transcriptionally active,

producing newly synthesized RNAs with new polyA tails. At these developmental stages, cytosolic polyA has been extensively studied in Xenopus and Drosophila oocytes and early embryos [46,49,51,52,55-59].

The polyA tail length is controlled by several enzymes that act on the 3'UTR by adding (PolyA Polymerase, PAP) or removing (PolyA Ribonuclease, PARN) adenenine ribonucleotides. After transcription, the RNA messages are transported out of the nucleus where they gain a longer polyA tail in the cytoplasm (Figure 1.5). This regulation requires a sequence element to be present upstream of the polyA site know as the cytosolic polyadenylation element (CPE). This CPE has a sequence motif of 4-6 Ts, followed by 1-2 As and a T. Importantly, CPE binding proteins (CPEBs) exist to recognize the aforementioned sequence motif and serve as the platform for other proteins such as PAP and PARN to bind. In the cytoplasm, protein kinases such as calmodulin-dependent kinase 2 alpha (CAMK2α)[60,61] and aurora A kinase (AURKA/Eg2)[62] are known to target CPEB and phosphorylate tyrosine residue 174 which in turn changes the balance of the PAP/PARN complex at the 3' UTR in favor of the PAP activity (Figure 1.5). Thus, 3' elongation of polyA can be observed after phosphorylation of CPEB. The increased polyA tail at the mRNA then promotes higher translational efficiency, resulting in an increase in protein synthesis of this particular message. Importantly, the messages involved in maturing oocyte and preparing the early embryo development need to also be negatively regulated in order to stop the production of proteins and conserve energy. This will ensure that enough nutrients and energy remain available for the embryo to attach to the uterus wall and develop a placenta. While the process of cytosolic polyA has been studied in great detail in Xenopus and Drosophila, a transcript wide analysis of polyA changes at that the stage of oogenesis and early embryo development in mouse or human remains elusive. Here we show a transcript wide analysis of polyA changes in human oocytes and early human

embryo with details discussed in Chapters 3 and 4.

1.9 Dissertation Overview

Overall, our goal was to study the epigenome of male and female germ cells in the context of aging and infertility for sperm cells and in a normal physiological context for human oocytes. Chapter 2 details my contributions in understanding the effect of aging on the sperm methylome. I wrote software and analyzed sperm cells regarding subpopulations and detailed changes in DNAme patterns in aged donors. Our findings suggest a link between hypomethylated regions in aged sperm from human donors and neuropsychiatric disorders. Future studies possibly involving mouse models may be able to show a causal link. Right now, loci-specific DNA hypomethylation in sperm of aged donors is only correlated with possible negative effects for the offspring.

Appendix A describes human sperm DNA methylation aberrations in infertile patients. Similarly to Chapter 2, I wrote software to analyze targeted bisulfite sequencing for candidate loci. The findings of this publication suggest a correlation of DNAme aberrations at developmentally important loci. It is important to note that this is not a causal link, but it suggests that improper DNAme levels at developmentally important regions may have an influence on male fertility.

Appendix B details a critical improvement to the sperm preparation protocol in order to abolish DNAme contaminations from somatic cells. It is paramount that only pure sperm samples are compared to each other in order to ensure appropriate conclusions. My contribution was the development of a robust and stringent sperm purification protocol that was tested and succeeded with the deliberate contamination of sperm cells with white blood cells. The results and the detailed protocol can be found in Appendix B.

In Chapters 3 and 4, we explore transcriptional changes of human oocytes

and early human embryos. The unique aspect of our work is to understand more about the connection between oocytes that are transcriptionally inactive and RNA regulation during the final stages of oocyte maturation. Notably, oocytes are extremely difficult to study since the material is very limited in amount. However, we generated RNAseq datasets that are unbiased for the polyA status of the transcripts present in oocytes. We developed a software tool called PANDA, detailed in Chapter 3, which is capable of detecting polyA changes of transcripts in both oocyte maturation as well as early embryo development. The software methods paper is written as a manuscript ready for submission and its content is part of Chapter 3. The results from applying PANDA towards the human oocyte and human early embryo data are part of Chapter 4. We were able to identify key transcripts that are subjected to polyA gain with importance in oocyte maturation and early embryo development. We also list transcription factors and DNA binding proteins that have not been previously shown to be important during that developmental time point. Our findings have been documented in Chapter 4, which is currently a manuscript in preparation to be submitted for publication.

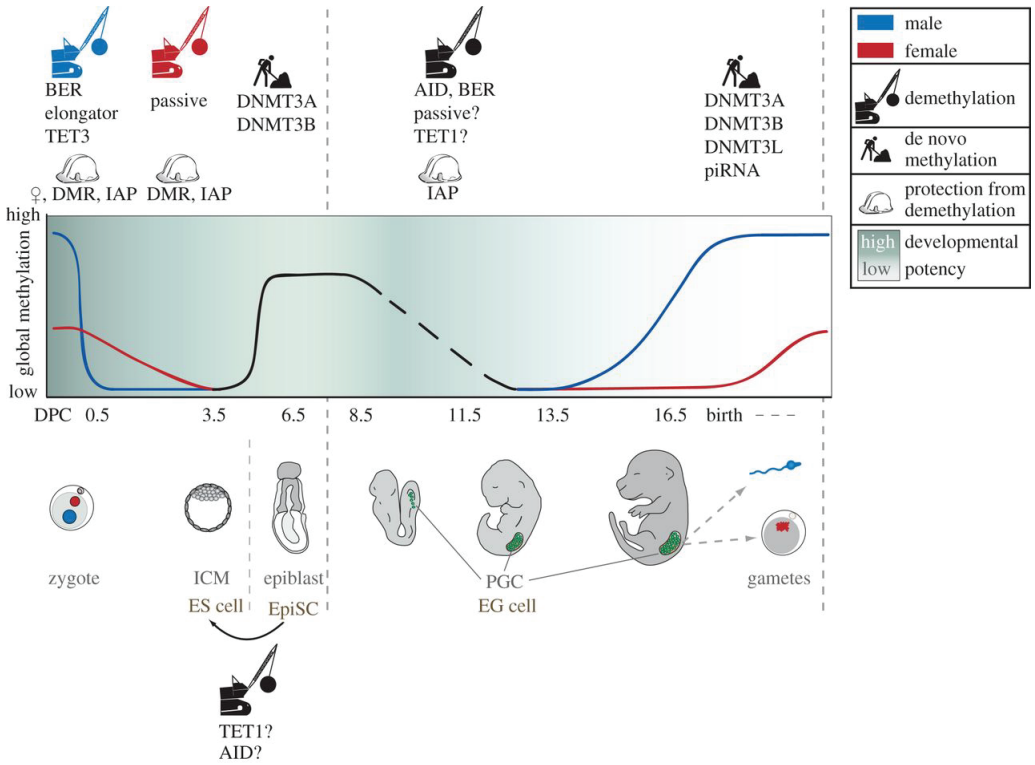


Figure 1.1: Overview of active and passive DNA demethylation druing mouse development in the paternal and maternal genome, respectively. Figure taken from Seisenberger et al., Philosophical transactions of the Royal Society of London Series B, Biological sciences. (2013).

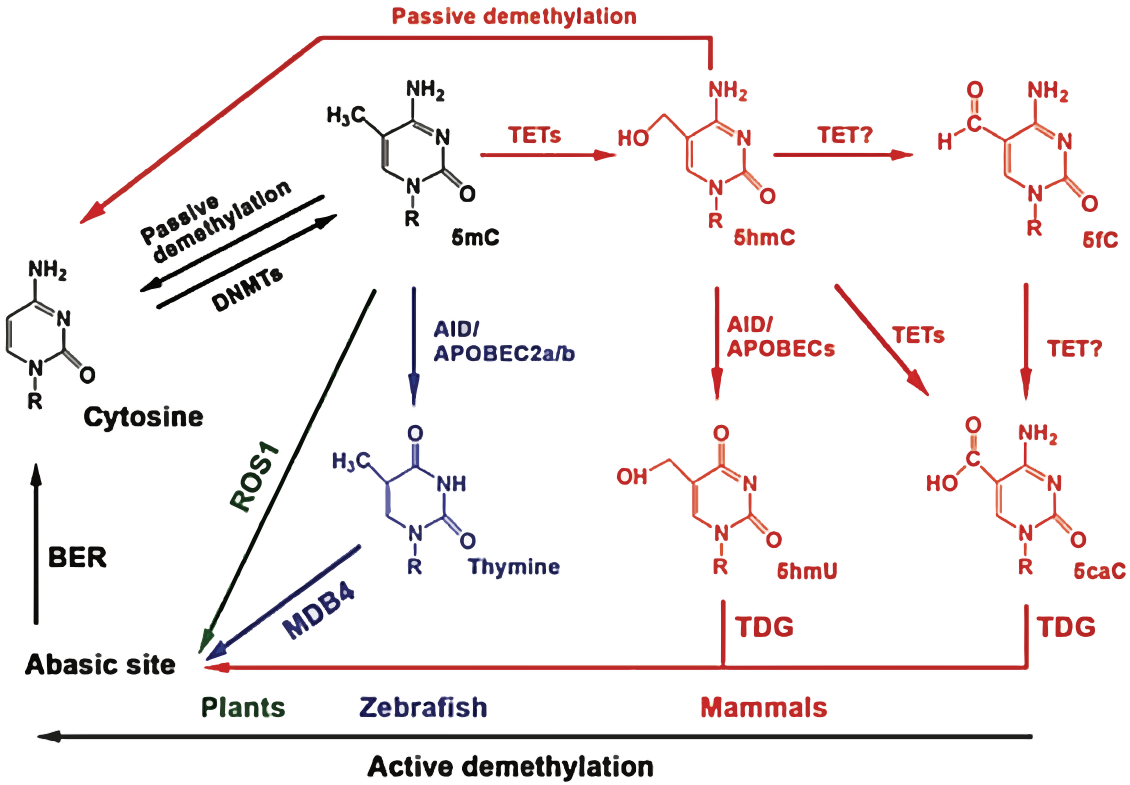


Figure 1.2: Pathway overview and chemistry of active and passive DNA demethylation in in mammals, zebrafish and plants. Any active DNA demethylation leads through an abasic site which is repaired by the base excision repair (BER) machinery. Figure was taken from Gong and Zhu, Cell Research (2011).

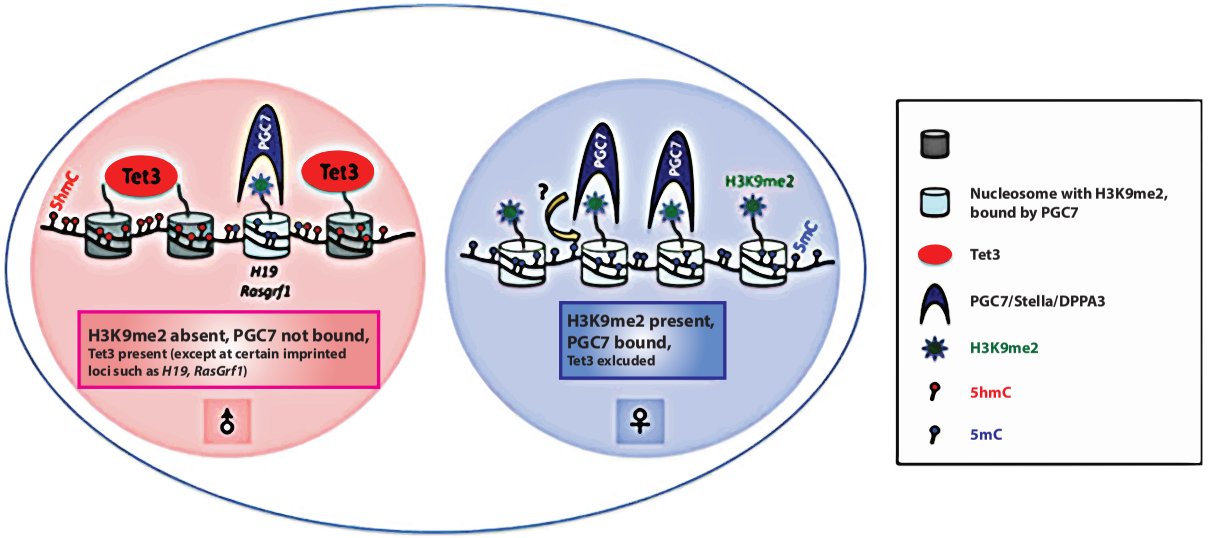


Figure 1.3: Pathway overview and chemistry of active and passive DNA demethylation in in mammals, zebrafish and plants. Any PGC7/DPPA3/Stella protects the maternal genome from Tet3-mediated active DNA demethylation. Histone H3K9me2 methylation plays an important role in DPPA3 recruitment. The paternal genome lacks H3K9me2-modified histones, resulting in lack of DPPA3 recruitment, and hence is preferentially subjected to Tet3-mediated DNA demethylation. Figure taken from Kang et al., Cell Res. (2013).

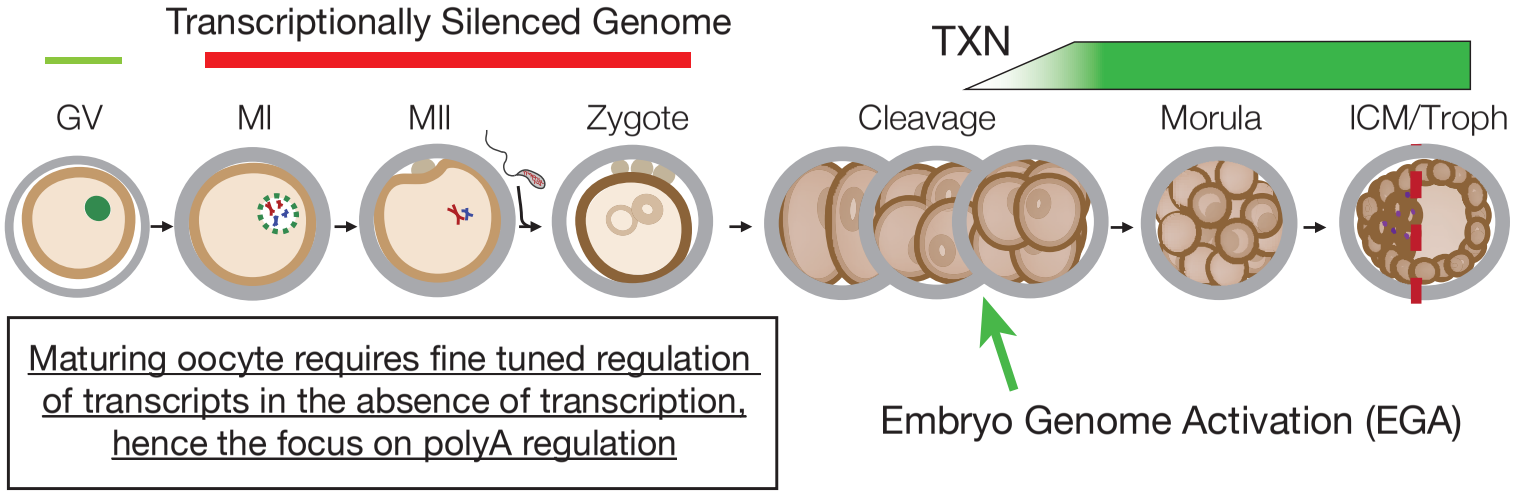


Figure 1.4: Late stage oocyte development and early embyro development lack transcriptional activity. Regulation of transcripts is mediated by modulating polyA levels in the cytosol. The mode of cytosolic polyA regulation is critical until embryo genome activation, when transcription starts again and new RNAs are synthesized. Transcriptional activity is shown in shades of green and transcriptional inactivity in red. Embyro genome activation (EGA) in humans occurs between the 4 and 8 cell stage.

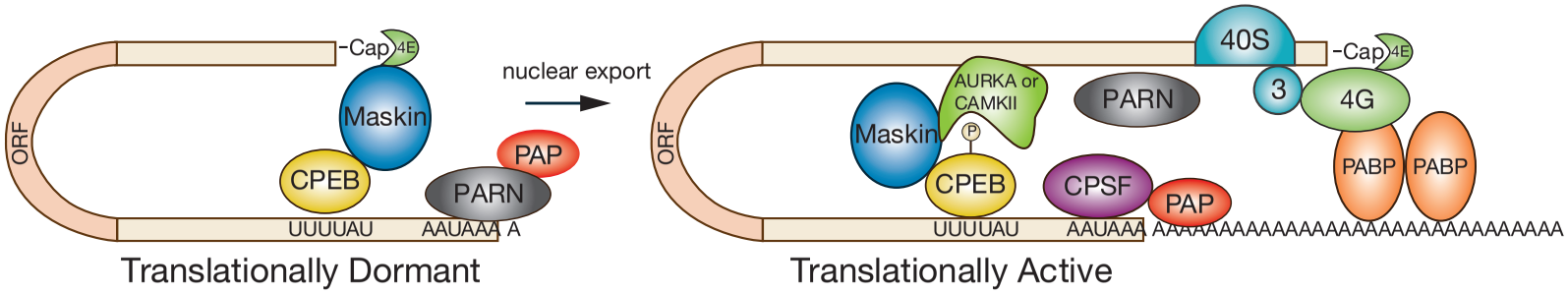


Figure 1.5: Cytosolic polyadenylation regulation. Transcripts containing CPE element upstream of the polyadenylation site are bound by CPEB and nuclear polyadenylation is supressed. Upon nuclear export and phosphorylation of CPEB at Tyr147 by AURKA or CAMKII, cytosolic polyA is activated, resulting in translational activity by means of ribosome engagement and protein synthesis. Figure adapted from Mendez, R. & Richter, J. D., Nat. Rev. Mol. Cell Biol. (2001).

1.10 References

1 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. Science 184, 868-871 (1974).

2 Waddington, C. H. The epigenotype. 1942. Int J Epidemiol 41, 10-13, doi:10.1093/ije/dyr184 (2012).

3 Holliday, R. Epigenetics: a historical overview. Epigenetics 1, 76-80 (2006).

4 Uberbacher, E. C. & Bunick, G. J. Crystallographic structure of the octamer histone core of the nucleosome. Science 229, 1112-1113, doi:10.1126/science.229.4718.1112 (1985).

5 Moudrianakis, E. N., Love, W. E. & Burlingame, R. W. Crystallographic structure of the octamer histone core of the nucleosome. Science 229, 1113, doi:10.1126/science.229.4718.1113 (1985).

6 Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. Cell 128, 707-719, doi:10.1016/j.cell.2007.01.015 (2007).

7 Siegfried, Z. et al. DNA methylation represses transcription in vivo. Nat Genet 22, 203-206, doi:10.1038/9727 (1999).

8 Jin, B., Li, Y. & Robertson, K. D. DNA methylation: superior or subordinate in the epigenetic hierarchy? Genes Cancer 2, 607-617, doi:10.1177/1947601910393957 (2011).

9 Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11, 204-220, doi:10.1038/nrg2719 (2010).

10 Bestor, T. H. The DNA methyltransferases of mammals. Hum Mol Genet 9, 2395-2402 (2000).

11 Jones, P. A. & Takai, D. The role of DNA methylation in mammalian epigenetics. Science 293, 1068-1070, doi:10.1126/science.1063852 (2001).

12 Eden, S. & Cedar, H. Role of DNA methylation in the regulation of transcription. Curr Opin Genet Dev 4, 255-259 (1994).

13 Hammoud, S. S. et al. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell 15, 239-253, doi:10.1016/j.stem.2014.04.006 (2014).

14 Habu, Y. et al. Epigenetic regulation of transcription in intermediate heterochromatin. EMBO Rep 7, 1279-1284, doi:10.1038/sj.embor.7400835 (2006).

15 Ahmed, S. & Brickner, J. H. Regulation and epigenetic control of transcription at the nuclear periphery. Trends Genet 23, 396-402, doi:10.1016/j.tig.2007.05.009 (2007).

16 John, R. M. & Surani, M. A. Imprinted genes and regulation of gene expression by epigenetic inheritance. Curr Opin Cell Biol 8, 348-353 (1996).

17 Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. Carcinogenesis 31, 27-36, doi:10.1093/carcin/bgp220 (2010).

18 Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69, 915-926, doi:0092-8674(92)90611-F [pii] (1992).

19 Hammoud, S. S. et al. Distinctive chromatin in human sperm packages genes for embryo development. Nature 460, 473-478, doi:nature08162 [pii]10.1038/nature08162 (2009).

20 Kobayashi, H. et al. Aberrant DNA methylation of imprinted loci in sperm from oligospermic patients. Human Molecular Genetics 16, 2542-2551, doi:10.1093/hmg/ddm187 (2007).

21 Kawamura, N. et al. Elevation of serum IgE level and peripheral eosinophil count during T lymphocyte-directed gene therapy for ADA deficiency: implication of Tc2-like cells after gene transduction procedure. Immunol Lett 64, 49-53, doi:S0165-2478(98)00083-2 [pii] (1998).

22 Marques, C. J. et al. Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia. Molecular human reproduction 14, 67-74, doi:10.1093/molehr/gam093 (2007).

23 Hammoud, S. S., Purwar, J., Pflueger, C., Cairns, B. R. & Carrell, D. T. Alterations in sperm DNA methylation patterns at imprinted loci in two classes of infertility. Fertility and sterility 94, 1728-1733, doi:10.1016/j.fertnstert.2009.09.010 (2010).

24 Wei, J. & Hemmings, G. P. TNXB locus may be a candidate gene predisposing to schizophrenia. Am J Med Genet B Neuropsychiatr Genet 125B, 43-49, doi:10.1002/ajmg.b.20093 (2004).

25 Lung, F. W., Tzeng, D. S. & Shu, B. C. Ethnic heterogeneity in allele variation in the DRD4 gene in schizophrenia. Schizophr Res 57, 239-245 (2002).

26 Arpanahi, A. et al. Endonuclease-sensitive regions of human spermatozoal chromatin are highly enriched in promoter and CTCF binding sequences. Genome Res 19, 1338-1349, doi:gr.094953.109 [pii]

10.1101/gr.094953.109 (2009).

27 Balhorn, R., Brewer, L. & Corzett, M. DNA condensation by protamine and arginine-rich peptides: analysis of toroid stability using single DNA molecules. Mol Reprod Dev 56, 230-234, doi:10.1002/(SICI)1098-2795(200006)56:2+<230::AID-MRD3>3.0.CO;2-V (2000).

28 Ward, W. S. & Coffey, D. S. DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. Biol Reprod 44, 569-574 (1991).

29 Jenkins, T. G., Aston, K. I., Pflueger, C., Cairns, B. R. & Carrell, D. T. Age-associated sperm DNA methylation alterations: possible implications in offspring disease susceptibility. PLoS Genet 10, e1004458, doi:10.1371/journal.pgen.1004458 (2014).

30 Kaati, G., Bygren, L. O., Pembrey, M. & Sjostrom, M. Transgenerational response to nutrition, early life circumstances and longevity. Eur J Hum Genet 15, 784-790, doi:5201832 [pii]

10.1038/sj.ejhg.5201832 (2007).

31 Pembrey, M. E. et al. Sex-specific, male-line transgenerational responses in humans. Eur J Hum Genet 14, 159-166, doi:5201538 [pii]

10.1038/sj.ejhg.5201538 (2006).

32 Carone, B. R. et al. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. Cell 143, 1084-1096, doi:S0092-8674(10)01426-1 [pii]10.1016/j.cell.2010.12.008 (2010).

33 Lai, F. & Shiekhattar, R. Where long noncoding RNAs meet DNA methylation. Cell Res 24, 263-264, doi:10.1038/cr.2014.13 (2014).

34 Rinn, J. L. et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

35 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322, 750-756, doi:10.1126/science.1163045 (2008).

36 Nagano, T. et al. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322, 1717-1720, doi:10.1126/science.1163802 (2008).

37 Redrup, L. et al. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development 136, 525-530,

doi:10.1242/dev.031328 (2009).

38 Kim, J. & Kim, H. Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. ILAR J 53, 232-239, doi:10.1093/ilar.53.3-4.232 (2012).

39 Goldstrohm, A. C. & Wickens, M. Multifunctional deadenylase complexes diversify mRNA control. Nat Rev Mol Cell Biol 9, 337-344, doi:10.1038/nrm2370 (2008).

40 Weill, L., Belloc, E., Bava, F. A. & Mendez, R. Translational control by changes in poly(A) tail length: recycling mRNAs. Nat Struct Mol Biol 19, 577-585, doi:10.1038/nsmb.2311 (2012).

41 Guzeloglu-Kayisli, O. et al. Embryonic poly(A)-binding protein (EPAB) is required for oocyte maturation and female fertility in mice. The Biochemical journal 446, 47-58, doi:10.1042/BJ20120467 (2012).

42 Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. Gene 265, 11-23 (2001).

43 Beilharz, T. H. & Preiss, T. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. RNA 13, 982-997, doi:10.1261/rna.569407 (2007).

44 Preiss, T., Muckenthaler, M. & Hentze, M. W. Poly(A)-tail-promoted translation in yeast: implications for translational control. RNA 4, 1321-1331 (1998).

45 Bentley, D. L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol 17, 251-256, doi:10.1016/j.ceb.2005.04.006 (2005).

46 Barkoff, A., Ballantyne, S. & Wickens, M. Meiotic maturation in Xenopus requires polyadenylation of multiple mRNAs. The EMBO journal 17, 3168-3175, doi:10.1093/emboj/17.11.3168 (1998).

47 Charlesworth, A., Cox, L. L. & MacNicol, A. M. Cytoplasmic polyadenylation element (CPE)- and CPE-binding protein (CPEB)-independent mechanisms regulate early class maternal mRNA translational activation in Xenopus oocytes. J Biol Chem 279, 17650-17659, doi:10.1074/jbc.M313837200 (2004).

48 Charlesworth, A., Ridge, J. A., King, L. A., MacNicol, M. C. & MacNicol, A. M. A novel regulatory element determines the timing of Mos mRNA translation during Xenopus oocyte maturation. The EMBO journal 21, 2798-2806, doi:10.1093/emboj/21.11.2798 (2002).

49 McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B. & Richter, J. D. Poly(A)

elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. Genes Dev 3, 803-815 (1989).

50 Hodgman, R., Tay, J., Mendez, R. & Richter, J. D. CPEB phosphorylation and cytoplasmic polyadenylation are catalyzed by the kinase IAK1/Eg2 in maturing mouse oocytes. Development 128, 2815-2822 (2001).

51 Mendez, R., Barnard, D. & Richter, J. D. Differential mRNA translation and meiotic progression require Cdc2-mediated CPEB destruction. The EMBO journal 21, 1833-1844, doi:10.1093/emboj/21.7.1833 (2002).

52 Paris, J. & Richter, J. D. Maturation-specific polyadenylation and translational control: diversity of cytoplasmic polyadenylation elements, influence of poly(A) tail size, and formation of stable polyadenylation complexes. Mol Cell Biol 10, 5634-5645 (1990).

53 Groisman, I., Huang, Y. S., Mendez, R., Cao, Q. & Richter, J. D. Translational control of embryonic cell division by CPEB and maskin. Cold Spring Harb Symp Quant Biol 66, 345-351 (2001).

54 Mendez, R. & Richter, J. D. Translational control by CPEB: a means to the end. Nat Rev Mol Cell Biol 2, 521-529, doi:10.1038/35080081 (2001).

55 Richter, J. D. Cytoplasmic polyadenylation in development and beyond. Microbiol Mol Biol Rev 63, 446-456 (1999).

56 Roy, L. M. et al. The cyclin B2 component of MPF is a substrate for the c-mos(xe) proto-oncogene product. Cell 61, 825-831 (1990).

57 Song, J. et al. The type II activin receptors are essential for egg cylinder growth, gastrulation, and rostral head development in mice. Dev Biol 213, 157-169, doi:10.1006/dbio.1999.9370 S0012-1606(99)99370-3 [pii] (1999).

58 Yamashita, M. Molecular mechanisms of meiotic maturation and arrest in fish and amphibian oocytes. Semin Cell Dev Biol 9, 569-579, doi:10.1006/scdb.1998.0251 (1998).

59 Salles, F. J., Lieberfarb, M. E., Wreden, C., Gergen, J. P. & Strickland, S. Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs. Science 266, 1996-1999 (1994).

60 Atkins, C. M., Nozaki, N., Shigeri, Y. & Soderling, T. R. Cytoplasmic polyadenylation element binding protein-dependent protein synthesis is regulated by calcium/calmodulin-dependent protein kinase II. J Neurosci 24, 5193-5201, doi:10.1523/JNEUROSCI.0854-04.2004 (2004).

61 Charlesworth, A., Meijer, H. A. & de Moor, C. H. Specificity factors in cytoplasmic polyadenylation. Wiley Interdiscip Rev RNA 4, 437-461, doi:10.1002/wrna.1171 (2013).

62 Mendez, R. et al. Phosphorylation of CPE binding factor by Eg2 regulates translation of c-mos mRNA. Nature 404, 302-307, doi:10.1038/35005126 (2000).

CHAPTER 2

AGE-ASSOCIATED SPERM DNA ALTERATIONS: POSSIBLE IMPLICATIONS IN OFFSPRING DISEASE SUSCEPTIBILITY

Reprinted with permission from Plos Genetics. Jenkins TG, Aston KI, Pflueger C, Cairns BR, Carrell DT (2014) Age-associated sperm DNA alterations: possible implications in offspring disease susceptibility. Plos Genetics 10(7): e1004458.

Chapter 2 is a published article. My contribution to this work involved performing the base pair resolution DNA methylation analysis from the samples that were sequenced using the MiSeq technology.

OPEN ACCESS Freely available online

PLOS | GENETICS

Age-Associated Sperm DNA Methylation Alterations: Possible Implications in Offspring Disease Susceptibility

Timothy G. Jenkins[1], Kenneth I. Aston[1], Christian Pflueger[2], Bradley R. Cairns[2,3]*, Douglas T. Carrell[1,4,5]*

1 Andrology and IVF Laboratories, Department of Surgery, University of Utah School of Medicine, Salt Lake City, Utah, United States of America, **2** Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah, United States of America, **3** Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America, **4** Department of Genetics, University of Utah School of Medicine, Salt Lake City, Utah, United States of America, **5** Department of Obstetrics and Gynecology, University of Utah School of Medicine, Salt Lake City, Utah, United States of America

Abstract

Recent evidence demonstrates a role for paternal aging on offspring disease susceptibility. It is well established that various neuropsychiatric disorders (schizophrenia, autism, etc.), trinucleotide expansion associated diseases (myotonic dystrophy, Huntington's, etc.) and even some forms of cancer have increased incidence in the offspring of older fathers. Despite strong epidemiological evidence that these alterations are more common in offspring sired by older fathers, in most cases the mechanisms that drive these processes are unclear. However, it is commonly believed that epigenetics, and specifically DNA methylation alterations, likely play a role. In this study we have investigated the impact of aging on DNA methylation in mature human sperm. Using a methylation array approach we evaluated changes to sperm DNA methylation patterns in 17 fertile donors by comparing the sperm methylome of 2 samples collected from each individual 9–19 years apart. With this design we have identified 139 regions that are significantly and consistently hypomethylated with age and 8 regions that are significantly hypermethylated with age. A representative subset of these alterations have been confirmed in an independent cohort. A total of 117 genes are associated with these regions of methylation alterations (promoter or gene body). Intriguingly, a portion of the age-related changes in sperm DNA methylation are located at genes previously associated with schizophrenia and bipolar disorder. While our data does not establish a causative relationship, it does raise the possibility that the age-associated methylation of the candidate genes that we observe in sperm might contribute to the increased incidence of neuropsychiatric and other disorders in the offspring of older males. However, further study is required to determine whether, and to what extent, a causative relationship exists.

Citation: Jenkins TG, Aston KI, Pflueger C, Cairns BR, Carrell DT (2014) Age-Associated Sperm DNA Methylation Alterations: Possible Implications in Offspring Disease Susceptibility. PLoS Genet 10(7): e1004458. doi:10.1371/journal.pgen.1004458

Editor: John M. Greally, Albert Einstein College of Medicine, United States of America

Received November 21, 2013; **Accepted** May 9, 2014; **Published** July 10, 2014

Funding: An internal University of Utah small grant from the "University of Utah Center on Aging" was used for this study. Additionally, clinical funds were used for this study. No outside grant agency funds were applied. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: Brad.Cairns@hci.utah.edu (BRC); douglas.carrell@hsc.utah.edu (DTC)

Introduction

The effects of advanced paternal age have only recently become of interest to the scientific community as a whole. This interest has likely arisen as a result of recent studies that suggest an association with increased incidence of diseases and abnormalities in the offspring of older fathers. Specifically, offspring sired by older fathers have been shown to have increased incidence of neuropsychiatric disorders (autism, bipolar disorder, schizophrenia, etc.) [1–3], trinucleotide repeat associated diseases (myotonic dystrophy, spinocerebellar atixia, Huntington's disease, etc.) [4–7], as well as some forms of cancer [8–11]. Though these are intriguing data, we know very little about the etiology of the increased frequency of diseases in the offspring of older fathers. Among the most likely contributing factors to this phenomenon are epigenetic alterations in the sperm that can be passed on to the offspring.

These studies are in striking contrast to the previously held dogma that the mature sperm are responsible only for the safe delivery of the paternal DNA. Intriguingly, with increased investigation has come mounting evidence that the sperm epigenome is not only well suited to facilitate mature gamete function but is also competent to contribute to events in embryonic development. It has been established that even through the dramatic nuclear protein remodeling that occurs in the developing sperm, involving the replacement of histone proteins with protamines, some nucleosomes are retained [12]. Importantly, histones are retained at promoters of important genomic loci for development, suggesting that the sperm epigenome is poised to play a role in embryogenesis [12]. In addition, recent reports suggest that hypomethylated regions with high CpG density also appear to drive nucleosome retention [13]. Similarly, DNA methylation marks in the sperm have been identified that likely contribute to embryonic development as well [12,14]. These data strongly support the hypothesis that the sperm epigenome is not only well suited to facilitate mature sperm function, but that it also contributes to events beyond fertilization.

Looking past fertilization and embryogenesis, sperm appear to contribute to events manifesting later in life. The remarkable claim that sperm, independent of gene mutation, may be capable of affecting phenotype in the offspring was initially proposed as a result of large retrospective epidemiological studies observing

Author Summary

There is a striking trend of delayed parenthood in developed countries due to secular and socioeconomic pressures. As a result, physicians commonly consult with concerned patients inquiring about the impact of advanced age on their ability to conceive healthy offspring. The concern has more frequently surrounded the effects of advanced maternal age, but recent evidence suggests negative effects of advanced paternal age as well. Specifically, studies have demonstrated increased incidence of neuropsychiatric and other disorders in the offspring of older males. In this study we have investigated a commonly hypothesized mechanism for this effect, namely sperm DNA methylation alteration. Our data indicate that specific genomic regions of DNA methylation are commonly altered with age, suggesting that some regions of the sperm genome are more susceptible than others to age-related epigenetic changes. Importantly, a significant portion of these alterations occur at genes known to be associated with schizophrenia and bipolar disorder, both of which display increased incidence in the offspring of older fathers. These data will be important in driving future studies aimed at determining the impact that these methylation alterations may have on offspring health and will thus enable couples at advanced reproductive ages to be more informed of possible risks.

changes in the frequency of diseases in the offspring of fathers who were exposed to famine conditions in the early 19th century [15,16]. Recently, many studies utilizing animal models have discovered similar patterns that comport with the epidemiological data. Specifically, in male animals fed a low protein diet, offspring display altered cholesterol metabolism in hepatic tissue [17]. However, the etiology of this phenomenon is poorly understood. Despite this, there are multiple likely candidates that may drive these effects, such as DNA methylation.

Methylation marks at cytosine residues, typically found at cytosine phosphate guanine dinucleotides (CpGs), in the DNA are capable of regulatory control over gene activation or silencing. These roles are dependent on location relative to gene architecture (promoter, exon, intron, etc.). Since these heritable marks are capable of driving changes that may affect phenotype, they represent a possible mechanism to explain the increased disease susceptibility in the offspring of older fathers. Additionally, in both sexes, aging alters DNA methylation marks in most somatic tissues throughout the body. In one of the first large studies to address the question of age-associated methylation alterations, Christensen et al. identified over 300 different CpG loci with age-associated methylation alterations in many tissues [18]. One recent study compared age-associated DNA methylation alterations in blood, brain, kidney and muscle tissue and identified both common and unique methylation alterations between different tissues [19]. Additionally, recent work suggests that DNA methylation can be used to predict the age of an organism based on tissue methylation profiles [20]. This study also supports previous reports which identify global hypomethylation as a hallmark of aging in most somatic tissues [21]. Because of its prevalence in other cell types, age-associated DNA methylation alteration is likely to occur in sperm as well. In further support of this idea is work demonstrating that frequently dividing cells typically have more striking methylation changes associated with age than do cells which divide less often [22]. In this study we have analyzed the age associated sperm DNA methylation alterations that are common among the individuals in our study population to determine the magnitude of sperm DNA methylation changes over time and whether specific regions are consistently altered with age.

Results

Our study includes 17 sperm donors (of known fertility) that collected an ejaculate in the 1990's. These donors were asked to provide an additional semen sample in 2008, enabling the evaluation of intra-individual changes to sperm DNA methylation with age. These samples are referred to as young (1990's collection) and aged (2008 collection) respectively. The age difference between each collection varied between 9 and 19 years, and the age at first collection ("young" sample) was between 23 and 56 years of age. Table 1 describes the donor demographics within both categories.

Global methylation analysis

To assess global methylation in the samples in question we performed pyrosequencing analysis of long interspersed elements (LINE), a commonly used tool for the analysis of global methylation in many tissues [23,24]. We identified significant global hypermethylation with age in sperm DNA as previous data from our lab suggests (Figure 1) [25]. Specifically, there was significant hypermethylation with age based on a paired analysis ($p = 0.028$) or by stratifying the samples by age alone and performing linear regression analysis ($p = 0.0062$).

Array analysis

In addition to the global analysis, we performed a high resolution (CpG level) analysis of methylation alterations with age. To perform this we utilized Illumina's Infinium HumanMethylation 450K array. Each sample was hybridized and analyzed on an array and the results were compared to detect changes in methylation that are consistent with age. We utilized a sliding window analysis, coupled with regression analysis (average methylation at identified window versus the age at collection) as an additional filter (any window whose regression p-value was >0.05 was excluded from downstream analysis), to compare changes that are common between paired samples. A Benjamini Hochberg corrected Wilcoxon Signed Rank Test FDR of $<= 0.0001$ and an absolute log2 ratio $>= 0.2$ (effectively a change in methylation of approximately 10% or greater) was used as our threshold of significance. Raw FDR values have been transformed for visualization in figures and reference in this text ($(-10 \log_{10}$ (q-value FDR)), such that a transformed FDR value of $13 = 0.05$, $20 = 0.01$, $25 = 0.003$, $30 = 0.001$, and $40 = 0.0001$. With this approach we have identified multiple age-associated intra-individual regional methylation alterations that consistently occur within the same genomic windows in most or all of the donors screened. Specifically, we identified a total of 139 regions that are significantly hypomethylated with age (Log_2 ratio ≤ -0.2) and 8 regions that are significantly hypermethylated with age (Log_2 ratio ≥ 0.2; Table S1). The average significant window is approximately 887 base pairs in length and contains an average of 5 CpGs with no fewer than 3 in any significant window. Of the 139 hypomethylated regions, 112 are associated with a gene (at either the promoter or the gene body), and of the 8 hypermethylated regions 7 are gene associated. The 8 hypermethylated regions that were found did change in all donor samples, however they did not increase DNA methylation levels beyond 0.1 fraction methylation. In one case we identified 3 significantly hypomethylated windows within a single gene (PTPRN2). Thus there were a total of 110 genes with age-associated hypomethylation.

Table 1. Donor demographics.

Parameter	Young (±SEM)	Aged (±SEM)	Significance
Age	37.7 (±2.12)	50.3 (±2.1)	N/A
Volume	3.78 (±0.46)	2.85 (±0.45)	p = 0.0142
Million/ml	125.4 (±9.16)	145.56 (±15.57)	P>0.05
Total count	434.32 (±53.67)	424.67 (±88.69)	P>0.05
Total motile	63.38 (±1.64)	61.25 (±4.34)	P>0.05
% live	69.08 (±1.47)	61.0 (±3.93)	P>0.05

doi:10.1371/journal.pgen.1004458.t001

A previous report analyzing multiple somatic tissues suggests that the magnitude of DNA methylation alterations that occurs with age is fairly subtle with an average percent change per year (measured as slope) at a single CpG of approximately 0.05% to 0.15% [19]. Our data, while still subtle, suggest that there may be a stronger effect of age on the methylation alterations in sperm compared with somatic cells. Briefly, in the four tissues screened by Day et al. (blood, brain, kidney and muscle) they identified a total of 8 individual CpGs with a methylation change per year of >0.4% and a single CpG with a yearly change of >0.5%. By comparison, our data have revealed a total of 26 genomic windows (not just individual CpGs) whose average fraction methylation change is >0.4% per year and 13 genomic windows with an average fraction methylation change per year of >0.5% (Figure 2A–B). Specifically in hypermethylated regions, the average fraction methylation change was 0.304% per year (ranging from 0.08% to 0.95% per year). In hypomethylated regions the average fraction methylation change was 0.279% per year (ranging from 0.08% to 0.92% per year). Considering the entire reproductive lifespan of a male, it is not unreasonable to anticipate an average change of 10–12% at these significantly altered regions. Importantly, these alterations all occur in windows with an average initial fraction methylation of <0.6 at the first collection and the majority (67% of altered regions) are also considered to have intermediate methylation based on conventional standards (fraction DNA methylation levels between 0.2 and 0.8; Figure 2B). Despite the increased magnitude of age-associated alterations in sperm when compared to somatic cells these changes are still quite subtle when considering the possible biological impacts at the 119 regions of age-associated alteration that are found at genes (gene bodies, promoters). Gene promoters were defined based on Illumina's array annotation, in general these fall within 1000 bps of the associated gene.

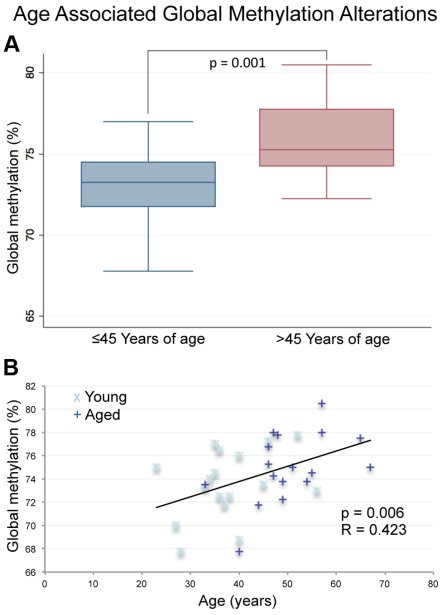


Figure 1. Pyrosequencing results for the LINE-1 global methylation assay. (A) The box plot depicts significantly increased average global methylation with age based on a non-paired t-test of all samples ≤45 (n = 17) years of age vs. all samples >45 (p = 0.001; n = 17). Global methylation was also stratified based only on age at the time of collection for each sample from all 17 donors (a total of 34 samples with each donor represented twice). (B) Linear regression analysis confirmed the significant increases in global sperm DNA methylation with age (p = 0.0062).
doi:10.1371/journal.pgen.1004458.g001

The significant loci identified in our analyses are located at various genomic features. The majority of regions that undergo age-associated hypomethylation occurred at CpG shores, whereas hypermethylation events are more commonly associated with CpG islands, and these differences are significant in both cases (p = 0.0015 and p = 0.0056 respectively; Figure 2C). It should be noted that while we did observe these significant changes there are slight differences in the baseline fraction methylation at islands and shores between regions with hypomethylation events and those with hypermethylation events (at the highest an absolute fraction methylation change of 0.16). We additionally analyzed the co-localization of windows of age associated methylation alterations with known regions of nucleosome retention in the mature sperm, as well as regions where specific histone modifications are found based on previous work from our laboratory [12]. We found that approximately 88% of regions that are hypomethylated with age are found within 1 kb of known nucleosome retention sites in the mature sperm (Figure 2D). Interestingly, loci that are hypermethylated with age are far less frequently found in regions of histone retention, with only approximately 37.5% being associated with sites where nucleosomes are found, though there are only 8 regions of significance on which to base this analysis. This difference was significant based on a fisher's exact test (p = 0.002). Similarly, 23% of loci with age-associated hypomethylation are associated with H3K4 methylation and 45.3% are associated with H3K27

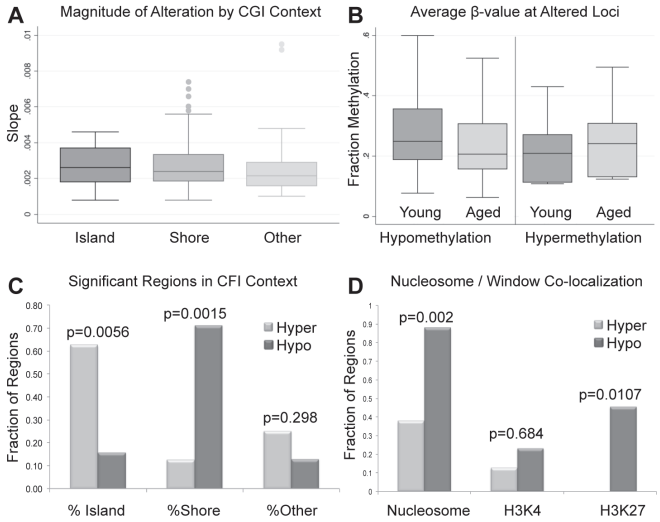


Figure 2. (A) The magnitude of alterations in terms of amount of change per year (reported as slope magnitude) for all regional changes that occur at CpG islands, shores and outside of these regions (other). Average alterations per year were approximately 0.281%. (B) Average β-values for all significant windows (hypomethylation and hypermethylation events) for both aged and young. Average decrease in β-value was approximately 3.9% for intra-individual hypomethylation events and 3.2% for hypermethylation events. (C) the percent of regions of hypermethylation and hypomethylation at CpG islands, shores and outside of these regions (Other). Hypermethylation events were significantly more enriched at islands than were hypomethylation events based on a fisher exact test ($p = 0.0056$). Hypomethylation events were significantly more enriched at shores in comparison to hypermethylation events ($p = 0.0015$). Hypermethylation and hypomethylation events were similarly enriched in regions outside of islands and shores. (D) We also investigated the co-localization of nucleosomes (every region of known histone retention) as well as histone modifications (H3K4 methylation, and H3K27 methylation) with our windows of interest. Hypermethylation events were less frequently associated with all retained histones (nucleosomes) or loci with H3K27 methylation when compared to hypomethylation events based on Fisher's Exact Test ($p = 0.002$; $p = 0.0107$). Co-localization of hypermethylation or hypomethylation events with H3K4 methylation was statistically similar.
doi:10.1371/journal.pgen.1004458.g002

methylation. The same co-localization is very rare with hypermethylation events ($p = 0.0107$). Additionally, we analyzed chromosomal enrichment of these marks to determine if there are specific chromosomal regions that are more susceptible to age-related methylation alterations. We found a random distribution of significant age-associated methylation alterations throughout the entire genome with what appears to be enrichment at telomeric and sub-telomeric loci, however this apparent enrichment failed to reach significance (Figure 3).

Sequencing analysis

To confirm our array data we selected 21 regions found to be significant by our array analysis and subjected them to targeted bisulfite sequencing (on the MiSeq platform) to confirm that the CpGs tiled on the array reflected the entire CpG content within the windows of interest. Specifically, we amplified via PCR, bisulfite converted DNA from each donor (young and aged collections). The PCR was designed to produce amplicons of approximately 300–500 bp that were located within 21 of the regions of significant methylation alteration we identified by array. Our depth of sequencing was quite robust with an average of 2,252 (SE ±371.6) reads per amplicon in each sample. The minimum number of average reads for any one amplicon was 313. In 20 of the 21 gene regions that were analyzed, the array and MiSeq data were similar in both direction and relative magnitude (Figure 4A). In the one case that did not show a similar trend (hypomethylation with age by array and no change by MiSeq) the amplicon was outside the region of the two CpGs that drove the significance of the window. When comparing the methylation of the approximately 300 bp amplicon to the CpG tiled on the array in that same region only, and not the array CpGs over the entire 1000 bp window, the data are in agreement. Taken together, the sequencing run confirmed that our array data is a good representation of the methylation status at all CpGs in our regions of interest.

Independent cohort analysis at identified regions of interest

To confirm that the sites identified on the array were not only altered in the samples we investigated, but that these loci are also altered with age in the sperm of non-selected individuals in the general population, we have performed an analysis on an independent cohort of individuals from two age groups: young, defined as <25 years of age ($n = 47$), and aged, defined as ≥45 years of age ($n = 19$). Average age in the young cohort was 20.46 years of age (SE ±0.18), and in the aged cohort 47.71 years of age (SE ±0.77). We performed a multiplex sequencing run on sperm DNA from these individuals to probe for 15 different regions of interest that were identified with the array data. Briefly, we PCR amplified 15 regions (using bisulfite converted DNA) from each

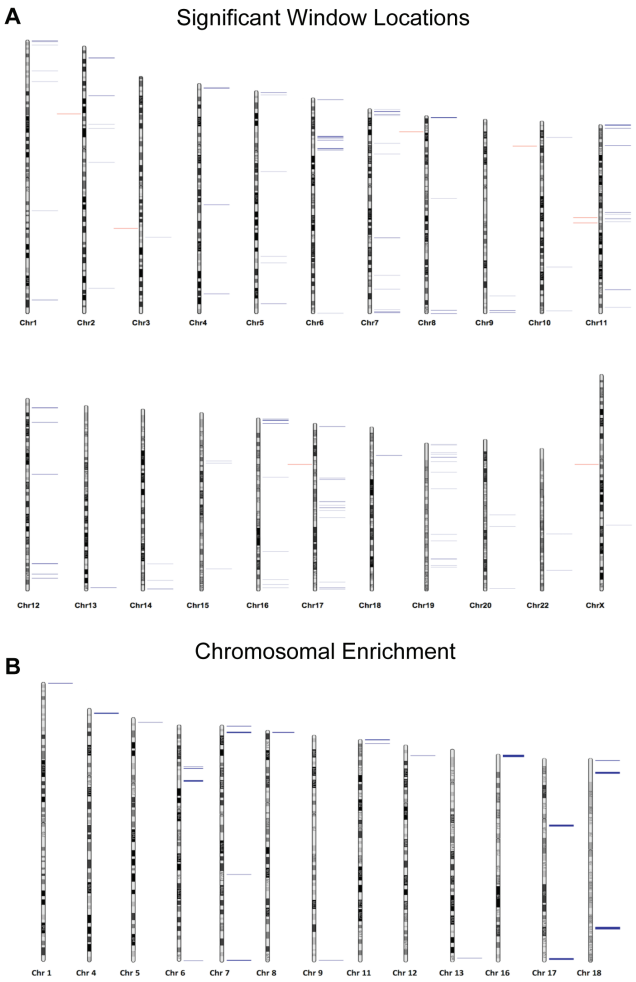


Figure 3. (A) Chromosomal loci of each altered region are depicted where blue marks represent hypomethylation events and red marks hypermethylation events. (B) The Correlation Maps app on the USeq platform was used to locate any specific chromosomal enrichment of altered methylation windows. Specifically, the application called any 100 kb region where at least two significantly altered methylation marks were found. All called chromosomal enrichment regions are displayed though none were found to be significantly enriched over the background.
doi:10.1371/journal.pgen.1004458.g003

individual (47 young, and 19 aged). The PCR was designed to produce amplicons of approximately 300–500 bp that were located within 15 regions of significant methylation alteration identified by array. Our depth of sequencing was, again, quite robust with approximately 3,645 (SE ±853.4) reads per amplicon in each sample with a minimum average number of reads for any one amplicon of 263. From these data we have confirmed that these genomic regions clearly undergo age-associated methylation alterations (Figure 4B). Interestingly, the average magnitude of alteration is also much higher in our independent cohort than in our initial paired donor sample study (approximately 2.2 times greater on average). This is particularly remarkable when considering that the average age difference in the independent cohort study was 27.2 years, effectively 2.3 times greater than the average age difference of 12.6 years seen in the paired donor analysis. This further supports our regression data in the paired

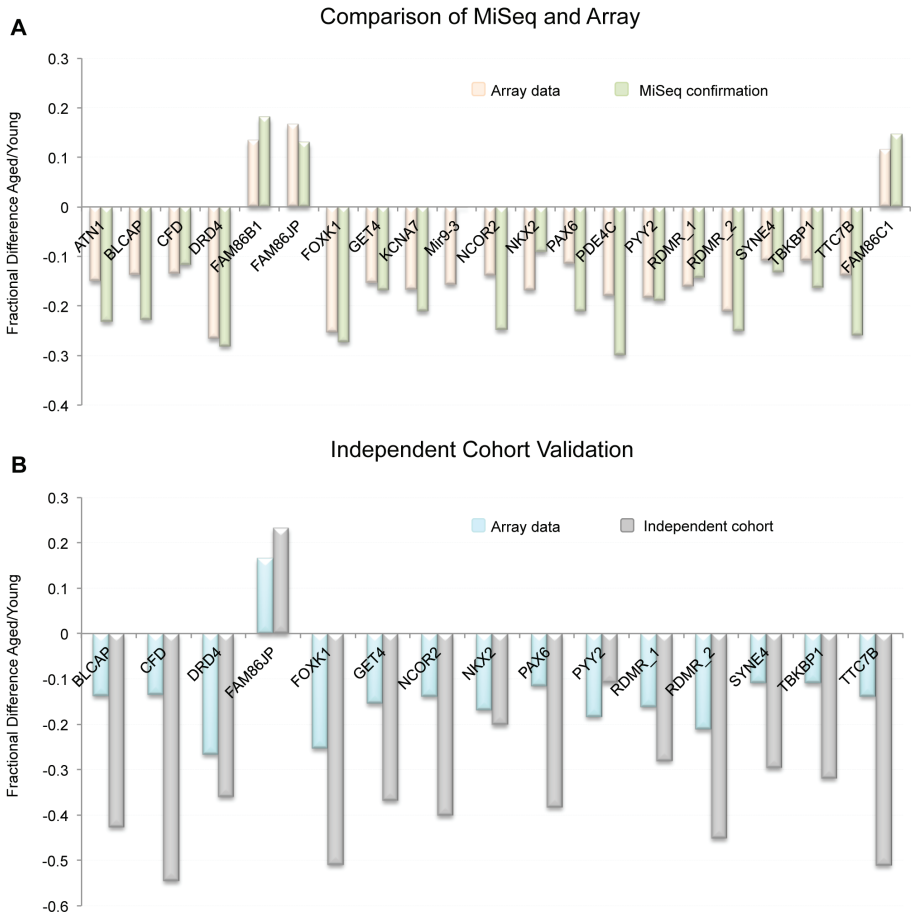


Figure 4. (A) Comparison of MiSeq results to our array results at 21 representative regions. Because beta-values and fraction methylation are generated in a different manner (array vs. sequencing respectively) they are not directly comparable. For this reason we compared the fractional difference for each loci and each technology. This is accomplished by the following equation: fractional difference = (aged value/young value) − 1. (B) the fractional difference between young and aged samples at 15 selected loci as measured by array in the 17 donor samples as well as in the independent cohort (19 samples from individuals >= 45 years of age and 47 samples from individuals <25 years of age taken from the general population). On average the fractional difference identified in the independent cohort (as measured by sequencing) was approximately 2.2 times greater in magnitude than was identified in the 17 donors.
doi:10.1371/journal.pgen.1004458.g004

donor study, which generally suggest a linear relationship of methylation alterations with age at most of the identified genomic loci.

Single molecule analysis of targeted sequencing

To address the question of the dynamics of sperm population changes associated with the approximately 0.281% change per year identified in this study we subjected our next generation sequencing data from the paired donor samples to a novel analysis where we compared the sperm population shifts between the young and aged samples. Because the MiSeq platform produces data for each single nucleotide sequence (each representing the methylation status in a single sperm) we are able to determine average methylation at each region for all of the amplicons analyzed. We identified 3 general patterns in methylation profile population shifts that resulted in the age–associated methylation

alterations we identified. First, we identified regions whose methylation at an age <45 was strongly hypomethylated, and the methylation profile in individuals >45 years of age is virtually the same, though it is more strongly hypomethylated. In these cases the change is still strikingly significant, but the magnitude of fraction DNA methylation change is minimal. Second, we see a single population in samples collected at <45 years of age that is shifted toward more hypomethylation in samples collected at >45 years of age. Last, we identified a bimodal distribution in samples collected <45 years of age that, in samples >45 years of age, is stabilized into a single population (Figure 5). This could be indicative of at least two sperm subpopulations, which are biased to a single, more hypomethylated sperm population with age. In every case the results suggest that all of the alterations we detected with the array are the result of the entire sperm population being altered in similar subtle ways and not a result of a dramatic alteration in a small portion of the sperm population.

GO term, Pathway and disease association analysis

The genes affected by the age associated methylation alterations (those that have alterations that occur at their promoter, or gene body) were analyzed by Pathway, GO and disease association analysis. The results indicate that no one GO term or Pathway is significantly altered in our gene group. Similarly, there were no significant diseases or disease classes associated with the genes identified in this study based on results of the disease association tool on DAVID. However the most significant disease hits (those that were significant prior to multiple comparison correction) have both been suggested to have increased incidence in the offspring of older fathers, namely myotonic dystrophy and schizophrenia [2,7].

To directly investigate the disease association(s) in our set of genes we searched the National Institute of Health's (NIH) genetic association database (GAD). We investigated all 117 genes that were determined to have age associated methylation alterations (110 hypomethylated; 7 hypermethylated) for their various disease associations. From these a total of 46 genes have been confirmed to be associated with either a phenotypic alteration or a disease based on GAD annotation. We identified 4 diseases that were most commonly associated with our set of genes (those disease that are associated with at least 2 genes identified in our study; diabetes mellitus, hypertension, bipolar disorder and schizophrenia). To further investigate these associations, we analyzed the frequency of genes associated with these 4 diseases in our gene set and compared it to their frequency in all 11,306 genes known to be associated with either a phenotypic alteration or a disease. Only bipolar disorder appeared to be more frequently associated with our identified genes than the background set of genes, based on chi-squared analysis with multiple comparison correction (Bonferonni) of the 117 age associated genes identified in our analyses ($p = 0.012$). Interestingly, schizophrenia also appeared to trend

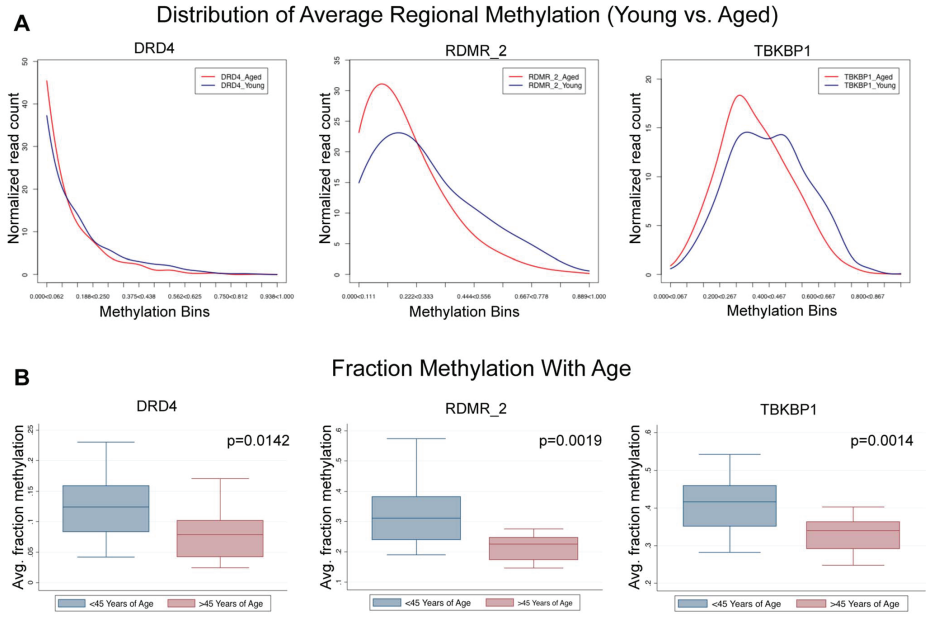


Figure 5. Single molecule analysis reveled 3 distinct alterations that occur with age. (A) DRD4 has only slight alterations associated with age because the young cohort (<45) is strongly hypomethylated initially, and aging simply amplifies this. RDMR_2 is representative of many alterations observed in this analysis which had a strong population shift from moderately hypomethylated to hypomethylated. TBKBP1 is representative of sites that had a bimodal distribution methylation patterns in the young group that becomes stabilized with age. (B) in every case (DRD4, RDMR_2, TBKBP1) each region has significant demethylation with age though the magnitude of change varies.
doi:10.1371/journal.pgen.1004458.g005

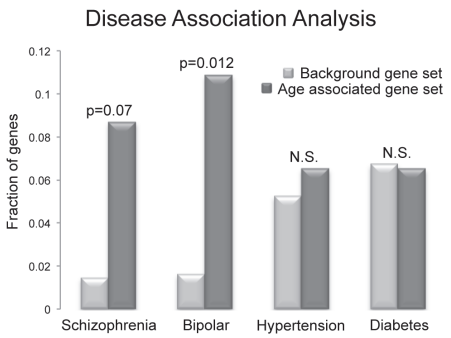


Figure 6. The frequency of disease associations within our gene set was analyzed and compared to the frequency of disease associations for all genes known to be associated with at least a single disease based on GAD annotation. Schizophrenia, bipolar disorder, diabetes mellitus and hypertension were selected, as there were at least 3 genes in our small set of identified genes that are associated with these diseases. Only bipolar disorder was more frequently associated with our identified genes than the background set of genes, based on chi-squared analysis with multiple comparison correction (Bonferonni) of the 117 age associated genes analyzed ($p = 0.012$), and schizophrenia also trended toward increased frequency ($p = 0.07$). However, these are not considered significant enrichments if considering all genes in the genome (omitting the filter for a disease connection). The frequency of genes associated with hypertension and diabetes mellitus in the two groups was statistically similar.
doi:10.1371/journal.pgen.1004458.g006

toward increased frequency ($p = 0.07$; figure 6). However, it is important to note that these are not considered significant enrichments if considering correction for comparisons with all genes in the genome (omitting the filter for a disease connection). The frequency of genetic association between our gene set and the background gene set was statistically similar for both hypertension and diabetes mellitus.

Discussion

Herein we report alterations to sperm DNA methylation associated with age. Interestingly, our data are in contrast with previous reports of age-associated methylation alterations in somatic cells. Recent literature suggests age-associated global hypomethylation with regional (gene associated) hypermethylation in somatic tissue [20,21]. In contrast, our data reveal age-associated hypermethylation globally with a strong bias toward hypomethylation regionally. This is less surprising when we take into account the fact that sperm are known to have other age-associated modifications that defy convention (i.e. telomere length) [26–28]. Intriguingly, while the methylation alterations reported herein are relatively subtle, they are strikingly significant and are common among individuals at various ages and intervals between collections, suggesting that these regions are consistently altered over time in a linear fashion. Importantly, it appears that many significantly altered regions are at loci that may contribute to various diseases known to have increased incidence in the offspring of older fathers. Coupling these with our data demonstrating that no one GO term or pathway is up or down-regulated in the sperm, as a result of the aging process, suggests that the alterations we observed are the result of regional genomic susceptibility to methylation alteration and not the activation or inhibition of any one cellular program. This hypothesis also comports well with the linear nature of the alterations we observed at most loci. While the nature of this susceptibility is difficult to elucidate, it may be related to chromatin architecture.

It should be noted that while we have identified many intriguing alterations to the sperm methylome, these likely do not represent all of the consistently altered regions that occur in the sperm over time. Our approach was to identify alterations with the use of a "promoter array" which has inherent biases. Specifically, the array is tiled with a higher density of probes at promoter or gene dense regions. Taken together with the use of a restricted window size (1000 bps) for searching the genome, this results in a bias toward identifying regions that are well covered on the array. While this bias is real, it does not invalidate the regions we have identified, but it suggests that more regions may be affected that are poorly covered on the array. Additionally, there are inherent concerns with the collection of tissues over time. One such concern is the difference in freezing methods over the years and the role this may play in methylation profiles. While it is unlikely that this alone could affect methylation profiles, the variances over time should still be reported. Our laboratory's protocol for freezing samples has been consistent for the times of collection of all samples included in this study. We have used the same cryomedium, test yolk buffer, over these years. It is unlikely then, that the methylation patterns in these samples are affected based on any of these variables, and thus the perturbations identified herein represent a true biological change to the sperm epigenome. This is supported by our replication dataset in which young and aged samples were collected concurrently.

Localization of altered regions

To investigate the attributes of regions that we determined to be most susceptible to methylation alterations, we evaluated the co-localization of significantly altered loci in our study with regions of nucleosome retention in the mature sperm. It appears that hypomethylation events are most commonly associated with sites of nucleosome retention. It should be noted that our criteria for sites of nucleosome retention is simply that our sites of alteration occur within 1 kb of known retention sites and thus there may be a greater degree of complexity in the actual sites of methylation alteration than we have identified. The actual nature of methylation patterns at a higher resolution in these regions (whether the affected regions are flanking or directly associated with histones) is difficult to elucidate due to the nature of array tiling in many of the loci we identified. Interestingly, this same co-localization was not seen with hypermethylation events. Though co-localization patterns are significantly different between the hypomethylation and hypermethylation events, it should be noted that the sample size is quite small in the hypermethylation group (8 significant windows). It should also be noted that while the co-localization of histones and the hypomethylation events we observed in our study are significant, the methylation marks observed are likely established earlier in spermatogenesis and thus may not be affected by the nucleosome architecture in the fully matured sperm. In addition, the alterations identified in this study are not localized everywhere that histones are retained, thus nucleosome retention alone can't be the independent driving force of regional susceptibility to methylation alterations. It should be further noted that our approach was not targeted to observe changes in chromatin co-localization patterns and as such this represents a secondary analysis of these patterns with the use of a "promoter array." As a result of observing only a selected portion of the genome, there are

clear biases that are introduced that should be taken into account when considering these findings.

Recent literature suggests an interesting hypothesis of "selfish spermatogonial selection" that may have application in this study as well [29]. Briefly, the hypothesis states that some gene mutations that are causative of abnormalities in the offspring are beneficial to spermatogenesis and become enriched throughout the aging process in spermatogonial stem cells. Thus, sperm carrying these mutations become more frequent in the population to the detriment of the offspring. Similarly, it is possible that the age-associated methylation alterations we have identified may be in regions that are important to spermatogenesis and thus would be selected for. While the genes identified herein are not well known spermatogenesis hotspots, they may lie close to other genes that are important in development and thus may be subject to a looser chromatin state leaving these genes more susceptible to methylation perturbations.

The hypomethylation events we identified could occur as a result of either active or passive demethylation. For example, regional transcription activity at loci important in spermatogenesis would likely be accompanied by a relaxed chromatin structure that could result in increased frequency of DNA damage over time. Established methylation marks located within this region could then be passively removed through repair mechanisms in the developing sperm. If the removal of this mark is either beneficial or has no effect on spermatogenesis it will persist, and over time similar marks could accumulate at nearby CpGs ultimately leading to the profile we identified in our study. It should also be noted that the accumulation of de novo mutations could lead to a similar profile. It is clear that the number of mutations in the sperm increase with age, and if these mutations involve deamination of cytosine residues the resulting sequence could appear as a loss of methylation with the technologies utilized herein. However, the mutation load, and specifically these C to T transitions, in sperm are stochastic in nature and thus cannot be the primary driving factor for the genomic hotspots of age-associated hypomethylation seen in virtually all of the individuals screened [30]. Alternatively, active enzymatic removal of methylation marks in the sperm might drive age-associated methylation changes. For this to be mechanistically plausible we would have to assume that hypomethylation in the windows we identified is always beneficial to spermatogenesis. While either of these mechanisms is plausible, it is likely that the effects we have identified involve some combination of both.

The mechanics of hypermethylation events with age are more difficult to elucidate, as this, by definition, has to be an active targeted process involving methyltransferase enzymes. However, some evidence from this study indicates DNA sequence may be an important driver of age-related hypermethylation. Of the 7 windows that we identified with gene-associated hypermethylation with age, 4 are associated with the FAM86 family of genes that are categorized not by protein function or genomic location but sequence similarity. This strongly suggests that, at least in part, age associated hypermethylation events at specific loci are driven, either directly or indirectly, by DNA sequence. Interestingly, this family of genes (FAM86) with unknown function has recently been categorized with a larger family of methyltransferase genes, though it remains unclear what the FAM86 target(s) is/are (DNA, Histone, other proteins, etc.). It is important to note that in addition to these regional hypermethyaltion events, globally DNA methylation is markedly increased as well. The possible role of chromatin modifications (histone tail modifications, etc) in this process is also important to note, as what we have identified may be linked to regional histone methylation, acetylation, etc. Such histone modifications may reflect underlying transcriptional changes during spermatogenesis. Taken together, the mechanisms that drive age-related methylation alterations in the sperm remain elusive, but likely involve both active and passive methylation modification.

Biological significance

It is important to consider two questions in determining the biological impacts of the identified methylation changes in this study. First, are the methylation changes described herein capable of transcriptional alterations? Second, are these methylation changes capable of avoiding embryonic methylation reprogramming? Regardless of the mechanism by which these methylation marks are altered in the sperm over time, it is striking that these changes occur with such consistency between individuals and have such a tight association with age that was seen in both the paired donor analysis and the independent cohort analysis. This is in stark contrast to the relative stability of the sperm methylome seen over time within each individual in the majority of the genome. One limitation of these findings, however, is the magnitude of alterations we have discovered. As described earlier the average fraction methylation alteration per year was approximately a change of 0.281%. Though this seems relatively small, when expanded to include the possible reasonable reproductive years in a male the change would be 10–12%. The increased magnitude of change with increasing age is strongly supported by our independent cohort study where an increase in the age difference between two groups was directly correlated with an increase in the magnitude of methylation alterations at virtually every locus screened in a relatively linear manner. Importantly, based on our analysis of complete nucleotide sequences from our sequencing data it appears that this decrease of 10–12% reflects changes to random CpGs within windows of susceptibility in all sperm, which would manifest in an individual sperm as a mosaically methylated region. The resultant 10–12% change in methylation within every individual sperm (effectively 1 out of every 10 CpGs are demethylated) suggests that every sperm carries similar, more subtle, alterations within these regions on average.

It is important to note that because we only investigated a portion of the regions of interest in our sequencing run (used for confirmation of array results) and the amplicons we probed made up only a portion of the regions of interest, we can not make a definitive overarching statement about the dynamics of methylation profile population shifts in sperm as a result of age. Despite this, the consistency of population shifts in the regions we were able to observe suggests that other regions of interest would likely follow similar patterns. Regardless, the resultant age-associated epigenetic landscape alterations may contribute to disease susceptibility in the offspring despite the small degree of change though the increased risk to the offspring may be relatively small. Figure 7 illustrates the alterations seen at two representative loci from our analysis, Dopamine receptor D4 (DRD4; ENSG00000170956) and tenascin XB (TNXB; ENSG00000168477).

The heritability of such marks is more difficult to elucidate mainly because the current study does not directly address this question. However, this issue needs to be addressed as the identified age-associated methylation alterations in the mature sperm may be removed through the embryonic demethylation wave. Despite the fact that there is no direct evidence of methylation alteration heritability at the specific loci presented in this work, the observed age-associated changes at regions known to be of significance in diseases with increased incidence in the offspring of aged males is striking and warrants further study. The intriguing localization of these alterations suggests that the

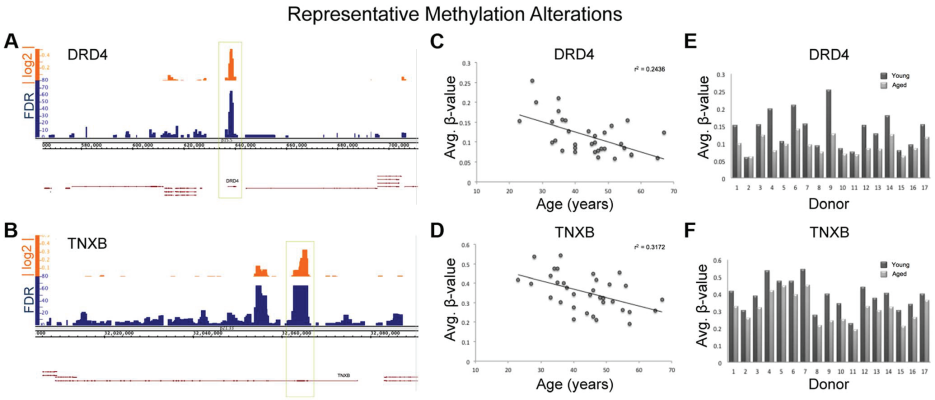


Figure 7. Various descriptive statistics are presented for both TNXB and DRD4; 2 regions of representative methylation alterations. (A,B) The alignment track for each gene is displayed in Integrated Genome Browser (IGB) with the associated false discovery rate (FDR) denoting the significance of the change and the absolute log 2 ratio reflecting the magnitude of the alteration. (C,D) Scatter plots for each sample from all 17 donors (a total of 34 samples with each donor represented twice) with linear regression lines and associated r^2 values were generated. Regression analysis revealed a significant decrease in methylation with age at both DRD4 and TNXB ($p = 0.0005$ and $p = 0.003$ respectively). (E,F) The average methylation within each window (DRD4 and TNXB) was plotted for each paired sample set and is displayed for each donor.
doi:10.1371/journal.pgen.1004458.g007

methylation profile in the mature sperm, at specific loci, either contributes to the increased incidence of associated abnormalities in the offspring or that they reflect (are downstream of) changes that are actually causative of the associated abnormalities in the offspring. Moreover, it has been previously proposed that epigenetic alterations are among the most likely candidates to transmit such transgenerational effects, and we have identified methylation alterations that appear capable of contributing to the various pathologies associated with advanced paternal age. Despite this, future work must still be performed to determine the real impact these marks have on transcription and thus phenotype and disease. Taken together, these subtle yet highly significant, age-associated alterations to the sperm methylation profile are intriguing because of their location and consistency, but more work is required to elucidate the biological impact of these marks.

There are many genes identified in our study that, if biologically affected, may result in pathologies in the offspring. DRD4 is one of the most widely implicated genes in the pathology of both schizophrenia and bipolar disorder as well as many other neuropsychiatric disorders [31,32]. Interestingly, the entire DRD4 gene itself is hypomethylated with age (Figure 7). TNXB has also been suggested to be associated with schizophrenia based on multiple studies, though the data are controversial [33,34], and virtually the entire 1st exon of TNXB is hypomethylated with age. Additionally, DMPK (ENSG00000104936), a gene identified in our study, is known to be associated with myotonic dystrophy, a disease for which advanced paternal age is a risk factor [7]. In fact, increases in trinucleotide repeats in DMPK are believed to be the cause of myotonic dystrophy type 1. Importantly, previous data suggests that altered methylation marks may affect trinucleotide instability [35]. These examples represent only a portion of the genes that were identified in our study and support the hypothesis that age-associated DNA methylation alterations in sperm may play a role in the etiology of various diseases associated with advanced paternal age.

Future directions

There are two important findings in this study. First, that there are any age-associated alterations common among such a varied study population (in terms of the age at collection) is remarkable. Specifically, age-associated methylation alterations occur in the sperm regardless of whether the ages between collections are approximately 20 to 30 years of age or 50 to 60 years of age. Second, the increased frequency of genes associated with bipolar disorder and schizophrenia in our study when compared to all genes associated with disease provides intriguing insight into the increased susceptibility of these specific disorders in the offspring of older fathers. Though frequently hypothesized, this work comprises, to the best of our knowledge, the first direct evidence suggesting the plausibility of epigenetic alterations in the sperm of aged fathers influencing, or even causing, disease in the offspring. Because of the nature of the unique sample set we have utilized in this study future work is needed to directly address a number of questions. Are methylation alterations, similar to those seen in our study, causative of neuropsychiatric disease? Can the methylation marks we observe in our study avoid embryonic demethylation? Future targeted work is required to directly address these questions to enable us to determine the role that these altered methylation marks play in the increased incidence of various diseases seen in the offspring of older fathers.

Methods

Ethics statement

The Institutional Review Board at the University of Utah approved this study. Written informed consent was obtained from all participants for their tissues to be utilized for this work.

Study group

Under an Institutional Review Board approved study our laboratory has accessed samples from 17 sperm donors (of known

fertility) that were collected in the 1990's. These donors provided an additional semen sample in 2008, enabling the evaluation of intra-individual changes to sperm DNA methylation with age. These samples are referred to as young (1990's collection) and aged (2008 collection) samples. The age difference between each collection varied between 9 and 19 years, and the age at first collection ("young" sample) was between 23 and 56 years of age.

At every collection donors were required to strictly follow the University of Utah Andrology Laboratory collection instructions, which includes abstinence time of between 2 and 5 days. The whole ejaculate (no sperm selection method was employed) collected at each visit was frozen in a 1:1 ratio with Test Yolk Buffer (TYB; Irvine Scientific, Irvine, CA) and stored in liquid nitrogen prior to DNA isolation. Samples were thawed and the DNA was extracted simultaneously to decrease batch effects. Sperm DNA was extracted with the use of a sperm-specific extraction protocol used routinely in our laboratory [36]. Briefly, sperm DNA was isolated by enzymatic and detergent-based lysis followed by treatment with RNase and finally DNA precipitation using isopropanol and salt, with subsequent DNA cleanup using ethanol. To ensure the absence of potential contamination from somatic cells the samples were visually inspected prior to DNA extraction. Additionally, we analyzed our sequencing results in an attempt to identify reads that did not match the methylation profile of sperm but instead reflected that of leukocytes. We also analyzed imprinted regions from our array data in an attempt to identify fraction methylation values that were inconsistent with previous reports of sperm DNA methylation patterns. Specifically, at a region of the IGF-2 locus that is tiled on the 450K array, it has been previously shown that sperm DNA is strongly hypermethlyated with a fraction methylation of approximately 0.8–0.85 and in leukocytes this same region is strongly demethylated with a fraction methylation of <0.1 [37]. Our array data indicate average methylation in every sample screened at these sites is approximately 0.844. In summary, with neither approach did we identify any signal that indicated leukocyte or other somatic cell contamination.

Pyrosequencing analysis

Each sample was subjected to pyrosequencing analysis of a portion of the LINE-1 repetitive element for the purpose of confirming previously determined global methylation changes with age. Briefly, isolated sperm DNA samples were submitted to EpigenDx (Hopkinton, MA) for pyrosequencing analysis. Qiagen's PyroMark LINE1 kit was used to determine methylation status at each CpG investigated with the assay. The experiment was performed based on manufacturer recommendations. The resultant values for each CpG are reported as fraction methylation, or the percent of methylated cytosines at that specifc CpG position. The average of these values was calculated for each individual (young and aged), and the values were compared both by linear regression and by a paired t-test.

Microarray analysis

Each of the paired samples for the 17 donors (a total of 34 samples) was subjected to array analysis using the Infinium HumanMethylation 450 Bead Chip micro-array (Illumina, San Diego CA). Extracted sperm DNA was bisulfite converted with EZ-96 DNA Methylation-Gold kit (Zymo Research, Irvine CA) according to manufacturer's recommendations. Converted DNA was then hybridized to the array and analyzed according to Illumina protocols at the University of Utah genomics core facility. Once scanned and analyzed for methylation levels at each CpG a β-value was generated by applying the average methylated and unmethylated intensities at each CpG using the calculation: β-value = methylated/(methylated+unmethylated). The resultant β-value ranges from 0 to 1 and indicates the relative levels of methylation at each CpG with highly methylated sites scoring close to 1 and unmethylated sites scoring close to 0.

The raw data were subjected to normalization to ensure the removal of poorly performing probes from the downstream analysis (probes with a QC $p<0.05$). Batch effect correction and basic descriptive analyses of the microarray data were performed using Partek (St. Louis MO). More in depth analysis was performed using the USeq platform with the application Methylation Array Scanner which identifies regions of altered methylation that are common among individuals utilizing a sliding window analysis. Briefly, paired data from each donor (young and aged) was subjected to a 1000 base pair sliding window analysis where regions of altered methylation with age that are common among donors were called by Wilcoxon Signed Rank Test. To diminish the influence of outliers in the data set, methylation for a specific window was reported as a pseudo-median and differences between the young and aged sample are reported as log 2 ratios. Two thresholds were applied to identify windows with significant differential methylation. A Benjamini Hochberg corrected Wilcoxon Signed Rank Test FDR of $<=0.0001$ ($>=$ transformed FDR of 40) and an absolute log2 ratio $>=0.2$ was used as our threshold for significance. Raw FDR values were transformed for visualization in figures and reference in this text ($(-10\ \log_{10}$ (q-value FDR)), such that a transformed FDR value of $13=0.05$, $20=0.01$, $25=0.003$, $30=0.001$, and $40=0.0001$, etc. We selected this robust level of significance, as opposed to an FDR of $>=13$ (corrected p-value of 0.05), to ensure that we selected only the most striking alterations that are consistently perturbed in most or all of the individuals screened. To confirm the significance of each of the called windows we subjected the mean β-value within the window for each donor (young and aged samples) to a paired t-test. Following this initial filter we additionally subjected each significant window to a regression analysis (age at time of collection versus average methylation within significant windows) to determine the relationship between age and mean methylation within each window. Regression analysis and paired t-tests were performed using STATA 11 software package. A p-value of <0.05 was considered significant for these analyses.

Sequencing analysis

We performed multiplex sequencing in a replication cohort as a confirmation that the alterations identified in the paired donors via array represent methylation alterations that are common in human sperm with age.

First, each donor sample used in the array study was additionally subjected to targeted bisulfite sequencing at loci determined to be most consistently altered based on the window analysis. This step was designed to confirm the array results and to provide greater depth of coverage of the CpGs in the windows of interest. Primers for 21 loci were designed using MethPrimer (Li Lab, UCSF). PCR was performed on samples following sperm DNA bisulfite conversion with EZ-96 DNA Methylation-Gold kit (Zymo Research, Irvine CA). PCR products were purified with QIAquick PCR Purification Kit (Qiagen, Valencia CA) and were pooled for each sample. The pooled products were delivered to the Microarray and Genomic Analysis core facility at the University of Utah for library prep which included shearing of the DNA with a Covaris sonicator to generate products of approximately 300 base pairs, in preparation for 150 bp paired end sequencing, and the addition of sample-specific barcodes for all 34 samples. Multiplex

sequencing was then performed on a single lane on the MiSeq platform (Illumina, San Diego CA).

Second, 19 sperm DNA samples from an independent, unselected cohort of general population donors who were ≥45 of ages were selected and compared to 47 sperm DNA samples from general population donors who were <25 years of age. These samples underwent the same preparation as described above for multiplex sequencing, though only 15 amplicons were targeted in this study of larger sample size. Average fraction methylation for each window was determined and was subjected to unpaired t tests between the young and aged groups.

Single molecule analysis of targeted bisulfite sequencing data

Bisulfite sequencing data was aligned against the human reference genome Hg19 using Novoalign. The aligned reads were processed using Novoalign Bisulfite Parser, BisStat and Parse Point Data Context for CG from the USeq package. The binned CpG graphs were generated using a modified version of the Allelic Methylation Detector from the USeq package. In short, all reads were queried for their number of CpGs. A consensus CpG number was then taken based on the highest number of CpGs per read and a minimum of 10% of all aligned reads (approximately 100 reads per region) must cover said number of CpGs. The consensus CpG number then served as the basis for the number of bins per region. Samples that were donated at an age of 45 years or older were coalesced *in silico* in the "aged donor group". Conversely, samples younger than 45 years were grouped in the "young donor group". All reads for the consensus CpG count were summed up based on their age group and then normalized to a 100 reads total. The graphs plotting normalized reads to methylation bins were then generated using the spline function from the R package.

GO term/Pathway/disease association analysis

GO term Analysis was performed with Gene Ontology Enrichment Analysis and Visualization Tool (GOrilla; cbl-gorilla.cs.technion.ac.il). Pathway and disease association analysis was performed on the Database of Annotation, Visualization, and Integrated Discovery (DAVID; david.abcc.ncifcrf.gov) v6.7. Additional disease association analysis was performed directly on the National Institute of Health's Genetic Association Database (GAD; geneticassociationdb.nih.gov).

Additional statistical analyses

Fishers exact test was used to determine the differences in frequencies of genes associated with particular diseases between our significant gene group and a background group. This analysis was also used to detect the differences in frequencies of windows that were found in regions of histone retention in the hypomethylation group and the hypermethylation group. Additionally, regression analysis was utilized to determine relationships between age and methylation status at various loci. STATA software package was used to test for significance with a $p<0.05$ being considered a significant finding.

Supporting Information

Table S1 Genomic features of significantly altered windows. Represented in this table are the windows of significance that were identified in our study as well as their transformed FDR, log 2 ratio, association to genes, association to known DMR, and CpG Island context.
(DOCX)

Author Contributions

Conceived and designed the experiments: DTC BRC TGJ KIA. Performed the experiments: TGJ KIA. Analyzed the data: CP TGJ BRC DTC KIA. Contributed reagents/materials/analysis tools: DTC BRC. Wrote the paper: TGJ CP BRC DTC KIA.

References

1. Hare EH, Moran PA (1979) Raised parental age in psychiatric patients: evidence for the constitutional hypothesis. Br J Psychiatry 134: 169–177.
2. Miller B, Messias E, Miettunen J, Alaraisanen A, Jarvelin MR, et al. (2011) Meta-analysis of paternal age and schizophrenia risk in male versus female offspring. Schizophr Bull 37: 1039–1047.
3. Frans EM, Sandin S, Reichenberg A, Lichtenstein P, Langstrom N, et al. (2008) Advancing paternal age and bipolar disorder. Arch Gen Psychiatry 65: 1034–1040.
4. Goldberg YP, Kremer B, Andrew SE, Theilmann J, Graham RK, et al. (1993) Molecular analysis of new mutations for Huntington's disease: intermediate alleles and sex of origin effects. Nat Genet 5: 174–179.
5. Andrew SE, Goldberg YP, Kremer B, Telenius H, Theilmann J, et al. (1993) The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet 4: 398–403.
6. Brunner HG, Bruggenwirth HT, Nillesen W, Jansen G, Hamel BC, et al. (1993) Influence of sex of the transmitting parent as well as of parental allele size on the CTG expansion in myotonic dystrophy (DM). Am J Hum Genet 53: 1016–1023.
7. Zheng CJ, Byers B, Moolgavkar SH (1993) Allelic instability in mitosis: a unified model for dominant disorders. Proc Natl Acad Sci U S A 90: 10178–10182.
8. Oksuzyan S, Crespi CM, Cockburn M, Mezei G, Kheifets L (2012) Birth weight and other perinatal characteristics and childhood leukemia in California. Cancer Epidemiol 36: e359–365.
9. Murray L, McCarron P, Bailie K, Middleton R, Davey Smith G, et al. (2002) Association of early life factors and acute lymphoblastic leukaemia in childhood: historical cohort study. Br J Cancer 86: 356–361.
10. Hemminki K, Kyyronen P, Vaittinen P (1999) Parental age as a risk factor of childhood leukemia and brain cancer in offspring. Epidemiology 10: 271–275.
11. Yip BH, Pawitan Y, Czene K (2006) Parental age and risk of childhood cancers: a population-based cohort study from Sweden. Int J Epidemiol 35: 1495–1503.
12. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, et al. (2009) Distinctive chromatin in human sperm packages genes for embryo development. Nature 460: 473–478.
13. Erkek S, Hisano M, Liang CY, Gill M, Murr R, et al. (2013) Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. Nat Struct Mol Biol 20: 868–875.
14. Arpanahi A, Brinkworth M, Iles D, Krawetz SA, Paradowska A, et al. (2009) Endonuclease-sensitive regions of human spermatozoal chromatin are highly enriched in promoter and CTCF binding sequences. Genome Res 19: 1338–1349.
15. Kaati G, Bygren LO, Pembrey M, Sjostrom M (2007) Transgenerational response to nutrition, early life circumstances and longevity. Eur J Hum Genet 15: 784–790.
16. Pembrey ME, Bygren LO, Kaati G, Edvinsson S, Northstone K, et al. (2006) Sex-specific, male-line transgenerational responses in humans. Eur J Hum Genet 14: 159–166.
17. Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, et al. (2010) Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. Cell 143: 1084–1096.
18. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, et al. (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet 5: e1000602.
19. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, et al. (2013) Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol 14: R102.
20. Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14: R115.
21. Richardson B (2003) Impact of aging on DNA methylation. Ageing Res Rev 2: 245–261.
22. Thompson RF, Atzmon G, Gheorghe C, Liang HQ, Lowes C, et al. (2010) Tissue-specific dysregulation of DNA methylation in aging. Aging Cell 9: 506–518.
23. Kreimer U, Schulz WA, Koch A, Niegisch G, Goering W (2013) HERV-K and LINE-1 DNA Methylation and Reexpression in Urothelial Carcinoma. Front Oncol 3: 255.

24. Deroo LA, Bolick SC, Xu Z, Umbach DM, Shore D, et al. (2013) Global DNA methylation and one-carbon metabolism gene polymorphisms and the risk of breast cancer in the Sister Study. Carcinogenesis 35: 333–338.
25. Jenkins TG, Aston KI, Cairns BR, Carrell DT (2013) Paternal aging and associated intraindividual alterations of global sperm 5-methylcytosine and 5-hydroxymethylcytosine levels. Fertil Steril 4: 945–951.
26. Unryn BM, Cook LS, Riabowol KT (2005) Paternal age is positively linked to telomere length of children. Aging Cell 4: 97–101.
27. Njajou OT, Cawthon RM, Damcott CM, Wu SH, Ott S, et al. (2007) Telomere length is paternally inherited and is associated with parental lifespan. Proc Natl Acad Sci U S A 104: 12135–12139.
28. Allsopp RC, Vaziri H, Patterson C, Goldstein S, Younglai EV, et al. (1992) Telomere length predicts replicative capacity of human fibroblasts. Proc Natl Acad Sci U S A 89: 10114–10118.
29. Goriely A, Wilkie AO (2012) Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. Am J Hum Genet 90: 175–200.
30. Paul C, Robaire B (2013) Ageing of the male germ line. Nat Rev Urol 10: 227–234.
31. Serretti A, Mandelli L (2008) The genetics of bipolar disorder: genome 'hot regions,' genes, new potential candidates and future directions. Mol Psychiatry 13: 742–771.
32. Lung FW, Tzeng DS, Shu BC (2002) Ethnic heterogeneity in allele variation in the DRD4 gene in schizophrenia. Schizophr Res 57: 239–245.
33. Wei J, Hemmings GP (2004) TNXB locus may be a candidate gene predisposing to schizophrenia. Am J Med Genet B Neuropsychiatr Genet 125B: 43–49.
34. Tochigi M, Zhang X, Ohashi J, Hibino H, Otowa T, et al. (2007) Association study between the TNXB locus and schizophrenia in a Japanese population. Am J Med Genet B Neuropsychiatr Genet 144B: 305–309.
35. Gorbunova V, Seluanov A, Mittelman D, Wilson JH (2004) Genome-wide demethylation destabilizes CTG.CAG trinucleotide repeats in mammalian cells. Hum Mol Genet 13: 2979–2989.
36. Nanassy L, Carrell DT (2011) Analysis of the methylation pattern of six gene promoters in sperm of men with abnormal protamination. Asian J Androl 13: 342–346.
37. Boissonnas CC, Abdalaoui HE, Haelewyn V, Fauque P, Dupont JM, et al. (2010) Specific epigenetic alterations of IGF2-H19 locus in spermatozoa from infertile men. Eur J Hum Genet 18: 73–80.

CHAPTER 3

PANDA: A NOVEL TOOL TO INVESTIGATE POLYADENYLATION CHANGES FROM NEXT-GENERATION TOTAL RNASEQ DATA

Chapter 3 is a manuscript in preparation for publication. The authors of this manuscript are Christian Pflüger, David Nix and Bradley Cairns.

3.1 Abstract

PolyA changes are physiological adaptions to RNAs in cells in order to control RNA stability and translational efficiency in maturing oocytes. It is of great interest to the scientific community to study transcript wide polyA changes in various developmental contexts and disease models. Recently, a polyA sequencing approach was shown to determine polyA tail lengths for a complete transcriptome. However, this sequencing technique is very specialized and requires total RNA amounts of 1-50 ug. Here we detail a bioinformatics approach, coined PANDA, that can detect relative polyA changes between transcripts using data from total RNAseq performed on samples with as low as 5 ng input RNA. PANDA enables the user to detect polyA changes from total RNAseq data that are not polyA selected. We identified that the frequency of adenine homopolymers with a length of >= 7 at the end of sequences in RNAseq datasets is significantly higher statistically than corresponding distribution of adenine homopolymers observed in either the human or mouse genome. This method also shows that changes in the ratio of reads over an exon that contain polyA to reads lacking polyA (polyA ratio) strongly correlate with changes in polyA length over the same exon. PANDA allows for PolyA detection based on a variable PolyA seed length where a shorter seed length comes at the expense of increased computational time. We recommend triplicates per condition sequenced at a minimum depth of 20 million reads. PANDA has the potential to identify transcriptome wide polyA changes for any RNAseq dataset that is not biased for polyA, which could be of particular interest to the field of neurobiology and cancer research where the cells are known to alter their polyA usage.

3.2 Introduction

A hallmark of cellular activity is the transcription of genomic information into RNA, essentially creating copies of the master DNA repository and using these

RNA transcripts for a variety of cellular functions [1]. Importantly, the stability and hence the half life of these copies need to be tightly regulated in order to ensure turn over and allow for adaptations to a changing environment [2,3]. The regulation of RNA stability and, in case of mRNA, the translation of such into protein, is considered a posttranscriptional process. A key posttranscriptional change at RNAs involves a process known as polyadenylation (polyA) that extends about 50-300 adenine nucleotides at the 3' prime end of the transcript. PolyA changes at mRNAs have been implicated to stabilize transcripts[4,5] as well as promote translational efficiency[6,7]. It has been shown that the default mode for RNA biogenesis is the co-transcriptionally polyadenylation in the nucleus[8]. However, a notable exception in development of maturing oocytes is the mechanism of cytosolic polyadenylation[9-12]. Briefly, transcripts that are used for late stage oogenesis or in the early developing embryo, a period in time where the genome is transcriptionally inactive and no new RNA transcripts are generated, are shown to skip nuclear polyadenylation and instead gain polyA in the cytosol. The regulatory element involved in mediating cytosolic polyadenylation, CPE (cytosolic polyadenylation element), has been shown to be part of the primary RNA sequence[13-15]. Further, binding proteins called CPEBs (cytosolic polyadenylation element binding protein) can recognize the CPE allowing for accurate timing of cytosolic polyadenylation upon their phosphorylation[10,13,14,16,17]. Most notably, the physiological process of cytosolic polyA gain has an essential role during the maturation of oocytes and in early embryos, before embryonic genome activation (EGA). EGA is the key point in embryonic development, when the genome in the embryo becomes transcriptionally active, producing newly synthesized RNAs with new polyA tails. At these developmental stages, cytosolic polyA has been extensively studied in Xenopous and Drosohilla oocytes and early embryos [9,12,14,15,18-22]. However, transcript wide analysis of polyA changes has only recently been described using PALseq[23] that requires a fair

amount of RNA input (>1ug) as well as a specialized sequencing setup. Here we describe PANDA (Polyadenylation from Next Generation Total RNAseq Differential Analysis), which can be used to analyze polyA changes between two different conditions. One signicant of this analysis method is that it uses standardized sequencing and library preparation methods. PANDA is part of the USeq package (http://useq.sourceforge.net)[24] and is also written in Java for platform independent usage.

Here we ask the question of whether polyA changes can be studied from total RNAseq data and we present a novel bioinformatics approach of investigating this biologically relevant question.

3.3 Results

PANDA as part of the USeq package was used to process two different datasets. We chose to analyze previously published data from tumor-associated macrophages (TAMs)[25] to serve as a negative control for PANDA since the miRNA manipulations used in that dataset was thought to not result in polyA level changes. The second dataset we used was total RNAseq from human oocytes from our lab which was expected to change polyA during the oocyte maturation phase. The choice for these two datasets was based on two criteria. Firstly, the datasets needed to be derived from total RNA where the library preparation method did not involve biasing for the polyadenylation status. This was achieved by either ribozero treatment to remove ribosomal RNA from the input into the library preparation followed by random hexamer reverse transcription for the TAMs dataset. Similarly, the human oocytes total RNA library was prepared using the Totalscript™ kit (Epicentre) that uses tagmenation with the Tn5 transposases in conjuction with proprietary buffer conditions to reduce ribosomal RNAs. Notably, neither library preparation uses oligodT and hence lacks a bias for the polyA status on the 3-prime

end of the RNAs. Secondly, every sample sequenced in these two datasets was sequenced with either 50bp paired-end or 101bp paired-end sequencing. The advantage of using paired-end sequencing data is to increase the dynamic range of reads that can be uniquely attributed as polyA. For example, reads that have more than 80% polyA present in their sequence can only be mapped uniquely by their read pair. Hence, paired-end sequencing data allows in some cases to attribute a polyA feature based on the mate's read. The choice of datasets analyzed was based on two aspects. The human oocyte dataset revealed significant changes in polyA levels during oocytes maturation. On the contrary, as expected, the TAMs dataset did not exhibit any substantial amount of transcripts with polyA changes.

3.3.1 PolyA Frequency in RNAseq Datasets is Statistically Significant Higher than in Human or Mouse Genomes

To justify the development of PANDA, we wanted to determine the dynamic range allowing us to detect polyA differences in total RNAseq data compared to adenine-homopolymer occurrence in the reference genome. We defined the dynamic range as the difference between polyA length occurrence in the genome and in a total RNAseq dataset. To test this, we analyzed the occurrence of adenine-homopolymer stretches in the human and the mouse reference genome as detailed in Figure 3.1A. As seen in the frequency plot, the separation of polyA length frequency starts at 7 nt and only increases all the way up to 62 nt. Also, neither genome has any adenine-homopolymers longer than 62 nt. In contrast to the genomic adenine-homopolymers, the frequency of polyA stretches in the TAMs dataset reveals a higher occurrence already at 7 or more adenine-homopolymers with a frequency of 7.6% and 7.4% for the human and mouse genome, respectively, and 16.8% for the TAMs dataset. Notably, the difference in distribution is highly statistically significant ($p<0.0001$). Comparing the occurrence of adenine-

homopolymers from human oocytes to the genomes (Figure 3.1B) also reveals a highly statistically different frequency distribution ($p<0.0001$). Remarkably, the unfiltered data from GV to MI oocytes exhibits a statistically significant difference in polyA frequency distribution. Further, there is a spike visible at the 101 nt mark indicating the presence of 101 adenine-homopolymers in the raw sequencing file. These 101 adenine-homopolymers could later be mapped if a read pair exists with a sufficiently mappable sequence.

3.3.2 The PANDA Workflow in Detail

Figure 3.2 illustrates the overall workflow of PANDA. The first step involves parsing the raw sequencing files (FASTQs) using the FastQPolyAParser program. As shown in the top part of Figure 3.2, raw reads are filtered for sequences that contain a adenine-homopolymer stretch at the 3' end or a thymine-homopolymer stretch at the 5' end. The FastqQPolyAParser has an option for the user to select the seed length (minimum adenine homopolymer count) with the default set to 6 or more 'A's. This results in all sequences that adhere to that criteria being saved in a separate FASTQ file. The original read in the FASTQ file that will undergo alignment is trimmed back to remove all adenine homopolymers at the 3' end or thymine homopolymers at the 5' end, respectively, as well as the corresponding sequence quality values. Importantly, if the program encounters 101 adenines or 101 thymines, the trimming process will leave a single A or T in order to not produce a read without any sequence information. Notably, the seed length parameter influences the identification of reads containing polyA albeit at a heavy cost in computational time. Figure 3.3 illustrates that while there is a gain of about 15% in polyA reads when the seed length is reduced from 6 to 5, the computational cost increases exponentially to 5-fold. Importantly, the computational time required for checking reads against the genomic adenine-homopolymer database is largely

dependent on disk I/O (input/output; hard drive or solid state drive speed). Hence, a good compromise for polyA reads identified and CPU time usage is the parameter '6' for the polyA seed length.

The next step involves the alignment of the trimmed FASTQ files. We used Novoalign to obtain reads mapped back in genome coordinate space. Subsequently, the SAM alignment files are subjected to SamTranscriptomeParser, which is also part of the USeq package[24]. This process ensures the conversion of spliced transcript information into genomic coordinates as well as sequence and alignment quality scores can be used to filter out low quality sequence reads. The next step of the PANDAworkflow, named SAMpolyATFilter, ties together the saved PolyA reads from step 1 as well as the BAM alignments from the SamTranscriptomeParser output. In essence, SAMpolyATFilter checks every alignment for the presence of a longer polyA containing read in the FASTQ file that contained the saved, untrimmed reads. If the alignment has a match for potential polyA, then it is checked against coordinates of genomic adenine-homopolymer stretches. Notably, the program checks both the primary read coordinates as well as the mate pair coordinates, provided they exist. In case the read matches the genomic adenine-homopolymer database, the program just moves on to the next read. Conversely, if the read passed the aforementioned genomic 'A' check, it is considered to be posttranscriptionally modified and hence the length of that polyA stretch is determined and saved to the SAM/BAM file as a user-defined tag. For example, if an alignment contained polyA of 25nt length, the SAM alignment is then updated with the tag At:i:25, where 'At' stands for the user defined tag, 'i' defines the value as an integer and the last number indicates the length of the polyA nucleotide stretch. Finally, SAMpolyATFilter produces two BAM files for each input, one for all the reads with the At:i tag set as well as a BAM file that only contains reads that have had the At:i tag set. The latter bam file can then be used

to visually inspect polyA reads in the genome browser. It is important to note that in order to produce bigwig or useq files, the normalization needs to be changed to total reads passing alignment and sequencing thresholds. As shown in Figure 3.4A and 3.4B, polyA containing reads are predominantly visible at the 3' end. An example for general polyA in TAMs is shown in Figure 3.4A and 3.4B, detailing the total read coverage in charcoal and the polyA read coverage in orange for ActinB and Tubulin4b. In contrast, Figure 3.4C depicts altering polyA levels in maturing human oocytes (GV, MI and MII) for the DLST gene.

The final step in the PANDA workflow involves the quantification polyA differences from two different conditions. The program named PolyADifferentialSeq is accepting two conditions with at least two replicates each, a UCSC gene table file for exon information. The user has the option to select if the last exon, all exons or exons and introns should be quantified. After running PolyADifferentialSeq, the output is a tab-delimited file that can easily be loaded up into any spreadsheet program for further analysis or filtering. The quantification is based on relative polyA counts over a given region (e.g. last exon) between two conditions. The log2 ratio is calculated and a p-Value using a chi-squared test is generated followed by multiple test correction using the Benjamini-Hochberg method. The $-10*\log_{10}$ transformed FDR is then reported. Further, stats like polyA length difference, polyA ratio difference as well as all the raw values can be all printed out into the tab-delimited file using the verbose option '-v'. In order to control for the minimum number of polyA counts present in both conditions and their combined replicates, the '-n' option allows the user set their own threshold, while the default is set to 30. Similarly, the '-e' option controls the total read counts of either of the conditions and their respective combined replicates.

3.3.3 PolyA Ratio and PolyA Length are Correlated

PANDA uses the simple measure of polyA ratio to determine changes in polyA levels between two different conditions. The PolyA ratio is determined by counting both the polyA reads as well as the total reads over a defined region (e.g. exon) and simply generating the quotient of both metrics. The fold polyA ratio change is further calculated by the log2 ratio of polyA ratio from condition one versus condition two. When looking into a quantification method to reproducibly detect polyA changes using existing RNAseq data, we noticed that polyA length changes correlated very well with polyA ratio changes (see Figure 3.5) with a Pearson's correlation coefficient of $r = 0.794$ ($p < 0.0001$). Remarkably, when the polyA Length reduced, the log2 polyA ratio followed the same trend. This observation allowed for rapid quantification of polyA counts over a given region.

3.3.4 Sequencing Depth Required for PolyA Analysis

A critical question is the dependency of sequencing depth in order to identify differential polyA. As shown in the coefficient of variation plot for the 45-55% most polyA reads in Figure 3.6, we started out with $4*10^7$ reads per replicate. In order to assess lower read coverage, we performed subsampling four times for each replicate using SAMtools[26]. We covered a broad spectrum cutting down the sequencing depth to $2*10^7$, $4*10^6$, $2*10^6$, $4*10^5$ and $4*10^4$ reads, respectively. As seen in the graph in Figure 3.6, $2*10^7$ reads per replicate still produces acceptable variation values. However, in order to capture even less highly expressed transcripts, deeper sequencing will always yield more information about the polyA status of each individual transcript.

3.4 Discussion

PANDA is a novel tool to investigate polyA changes in a transcriptome that has not been oligodT selected. Here we demonstrated a workflow for using library preparation for RNA-seq in combination with the PANDA software to analyze changes in polyA. Previously, Subtelny et al. published a remarkable method to detect and measure polyA tail length on a transcriptome wide level. However, a fundamental limitation of this technique is the total RNA amount necessary to investigate polyA length changes (1-50 µg total RNA). Tissue or cells that are difficult to obtain in high enough quantities such as oocytes will pose as a hurdle to investigate with the help of PAL-seq. In contrast, PANDA allows for polyA analysis with total RNA amounts as low as 5 ng. An important consideration, however, is the fact that PAL-seq enables the exact measurement of polyA tail lengths. This enables comparison and analysis of polyA tail lengths between different transcripts. PANDA in comparison is working with the parameter of relative polyA tail length change between two conditions for each individual transcript. The biologically important feature of relative polyA change per transcript is also very critical to understand. It enables the investigator to ask the question of whether polyA is added or removed for select transcripts present in the RNAseq data. While PANDA cannot determine the transcriptome wide exact polyA tail length as detailed previously in the PAL-seq method[23], PANDA allows detection of relative polyA changes between two conditions. Notably, the total input RNA amount necessary can be as low as 5 ng as opposed to 1-50 µg used for PAL-seq. This roughly 1000-fold difference in input RNA amounts is absolutely critical when working with hard to obtain tissues like oocytes.

The current maximum polyA length detectable by PANDA is based on sequencing read lengths. Illumina version 3 chemistry is currently limited to 101 nucleotides, which also limits the homopolymeric adenine detection to the same

length. It is noteworthy at this point that Illumina version 4 chemistry is now available, allowing an 25% increase in read lengths to 125 nts with 125 bp paired-end sequencing. This increase in sequencing length will enable an increase in dynamic range, increasing the accuracy of calling polyA differences. In the near future, it is imaginable that longer read lengths may allow detection of full-length polyA tails. An interesting experiment could entail the enrichment of mRNAs followed by 300 bp paired-end MiSeq sequencing. The theoretical maximum sequencing length of 600 nt is in a range that could be sufficient to measure the full spectrum of polyA lengths.

In Figure 3.1, we compared the background adenine homopolymer frequency to the polyA frequency of transcripts sequenced in either human oocytes or TAMs. An interesting observation is the similarity of the human and mouse adenine homopolymer frequency. Not surprisingly, the maximum adenine homopolymer length is around 60 nt for both genomes. A possible caveat, however, might be the assembly of repetitive regions for both genomes, which could possibly harbor more and longer adenine homopolymer stretches. Most importantly though, even if future genome versions will have better annotations of repetitive regions, PANDA's ability to uniquely call polyA for a transcript will not be changed by that fact. As shown in this publication, the possibility of differential polyA analysis with the current genome annotation and sequencing technology is very much possible. This is also strongly supported by the higher frequency of 7 or more adenines at the 3' end of transcripts compared to the occurrence of 7 or more adenines in a homopolymer in the genome. Consequently, it is more likely for PANDA to call a read with 7 or more adenines at the 3' end of a read a true polyA as opposed to a genomic adenine homopolymer. Importantly, with an increase in adenine homopolymer lengths, the chance for true polyA only increases compared to the genome background. Further, the ability of PANDA to call up to 101 nt of polyA,

or longer if a longer better sequencing technology is used, illustrates the dynamic range currently available.

In summary, we have developed PANDA to investigate polyA changes in RNAseq data that are not polyA selected. PANDA is compatible with previously published RNAseq data and can be used in the future for total RNAseq datasets. This will open up opportunities to simultaneously investigate transcript changes as well as polyA changes in a variety of biological questions, including development and cancer.

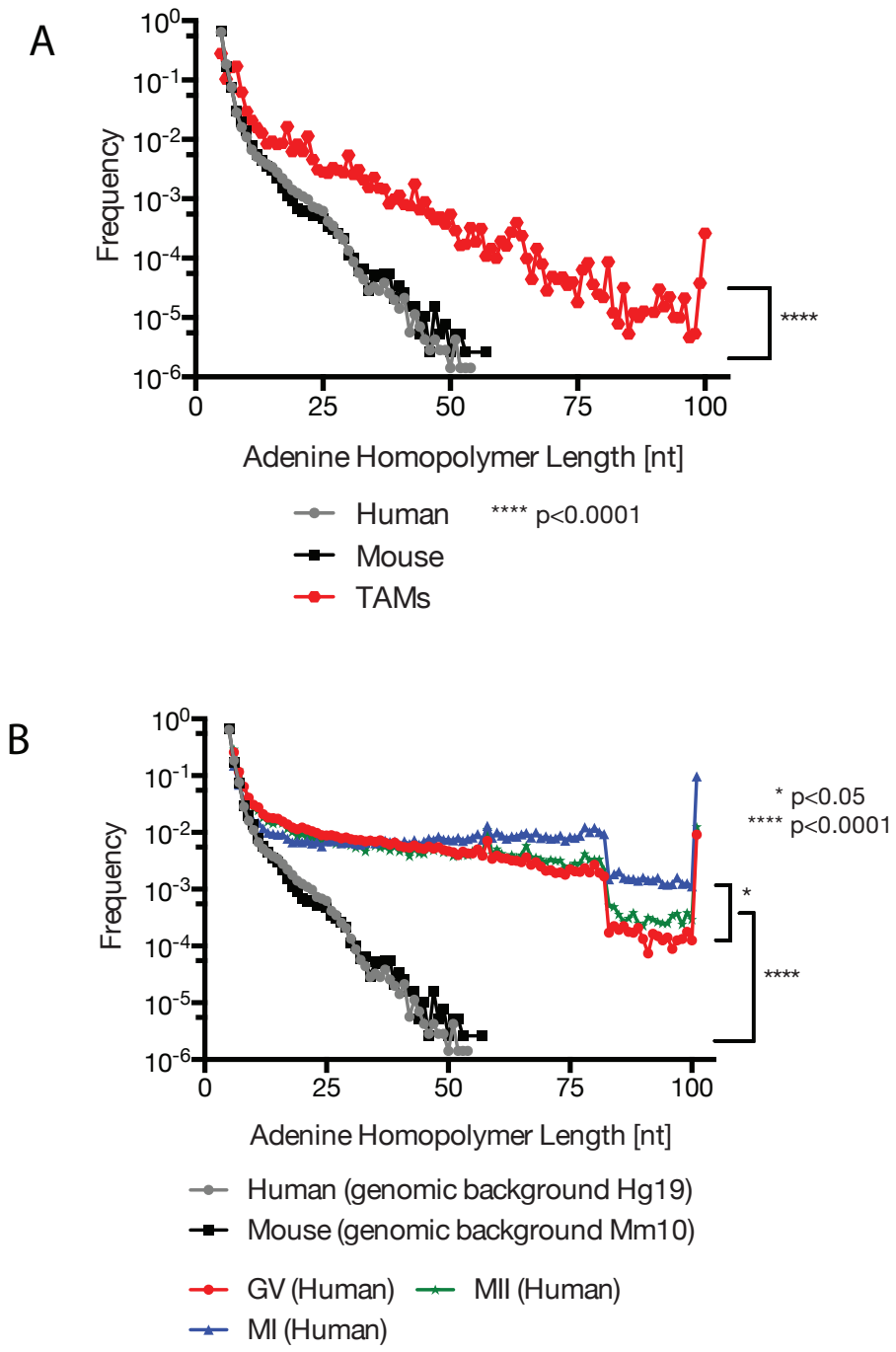


Figure 3.1 Comparison of genomic adenine homopolymer frequency to polyA length in sequenced samples. A) Tumor associated macrophages (TAMs) polyA length compared to mouse genomic adenine homopolymer frequency. Statistical test was performed using the Wilcoxon matched-pairs signed rank test ($p<0.0001$). B) Human oocytes polyA length compared to human genomic adenine homopolymer frequency. Statistical test was performed using the Wilcoxon matched-pairs signed rank test ($p<0.0001$).

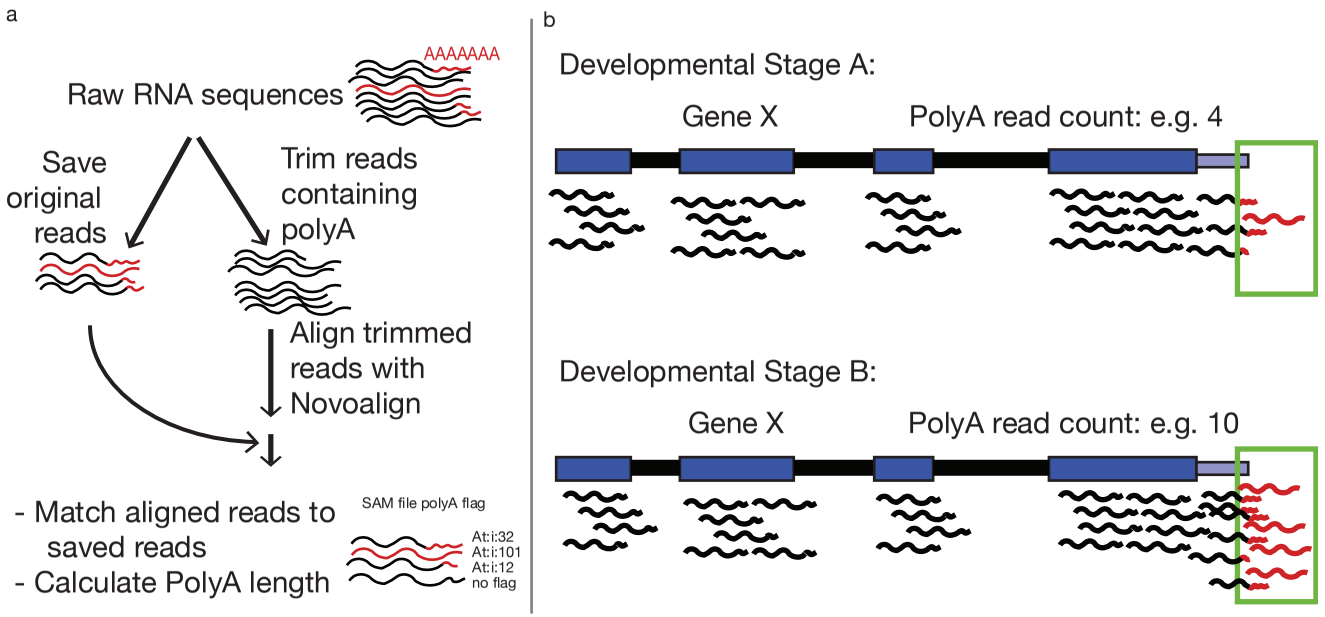


Figure 3.2 Overview of PANDA workflow. a) Raw sequencing reads get filtered for 3' polyA or 5' polyT. Matches are saved in a seperate file and reads in the original fastq file are trimmed back to allow for alignment. Subsequently, reads are checked against thesave file with the unmodified reads containing potential polyA and if the read matches and does not map to a genomic adenine-homopolymer stretch, receives a polyA At:i:length tag. b) polyA counts per exon are summed up and normalized to total exon counts. The log2 ratio of two conditions is tested with chi-squared test and Benjamini-Hochberg multiple Test corrected.

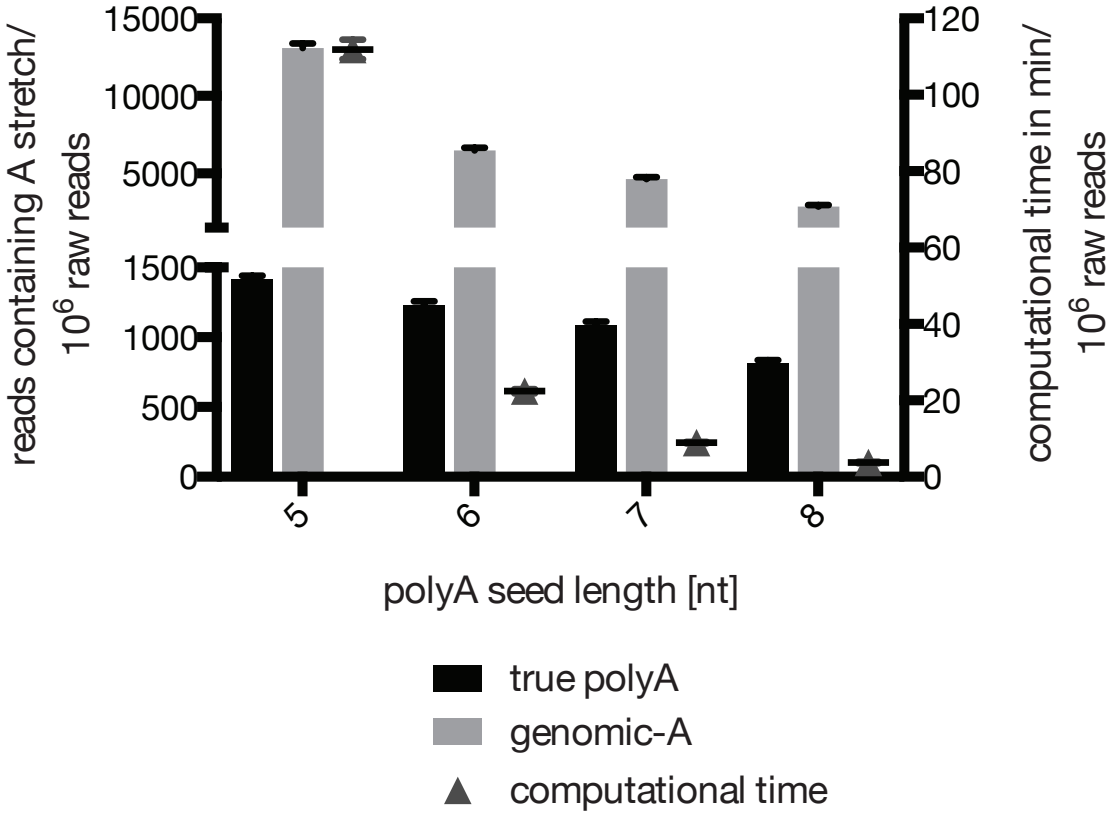


Figure 3.3 PolyA detection and computation rate. PolyA detection is dependent on polyA seed length. Shorter polyA seed length increases polyA detection but increases computational time exponentially.

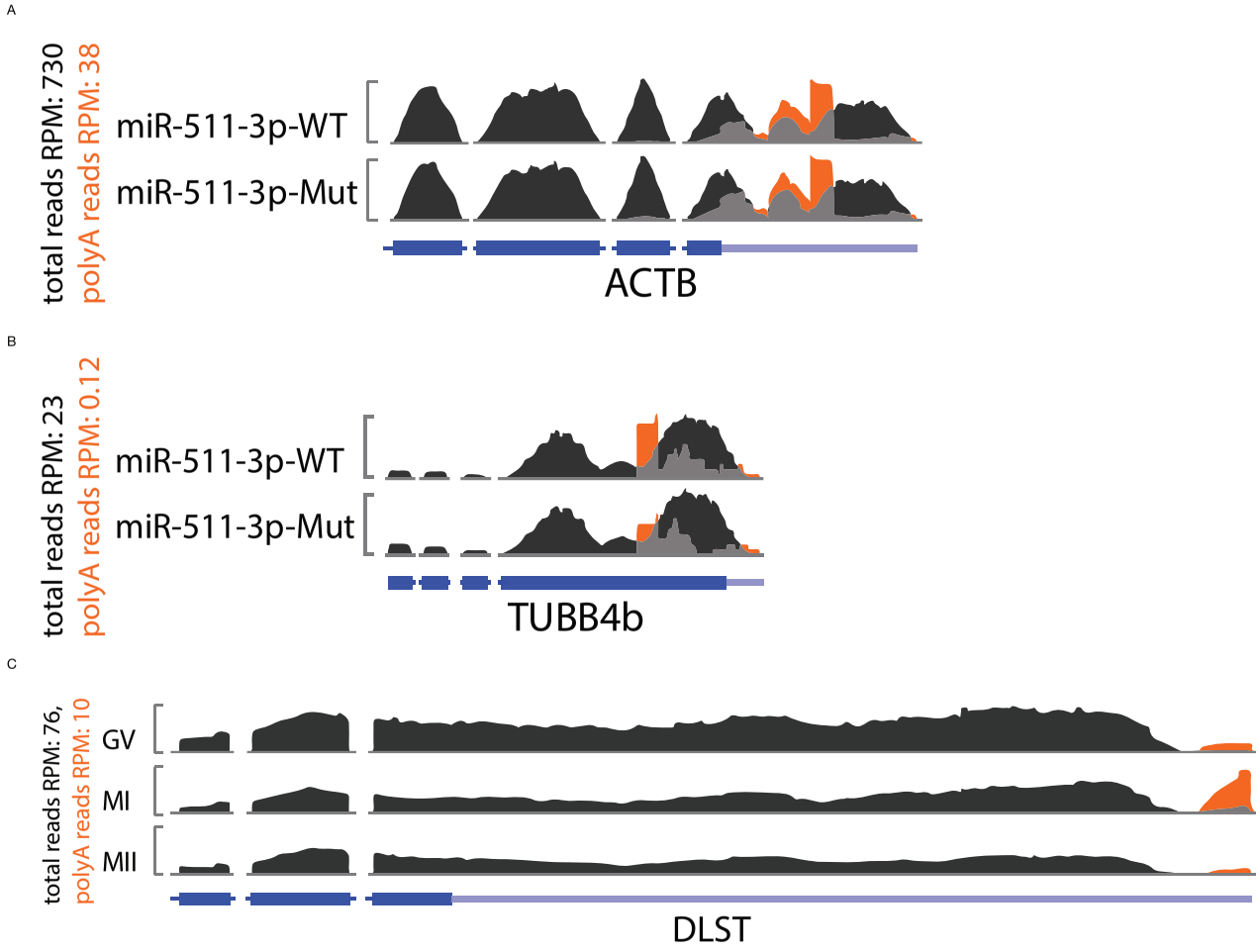


Figure 3.4 Genomic snapshot of polyA reads (orange) and total RNA reads (charcoal) for actin and tubulin in TAMs (4A and 4B). Figure 4C illustrates PolyA changes at 3'UTR, visible during human oocyte maturation (GV, MI and MII stage) at the DLST gene. Only up to the last 4 exons are shown for each transcript.

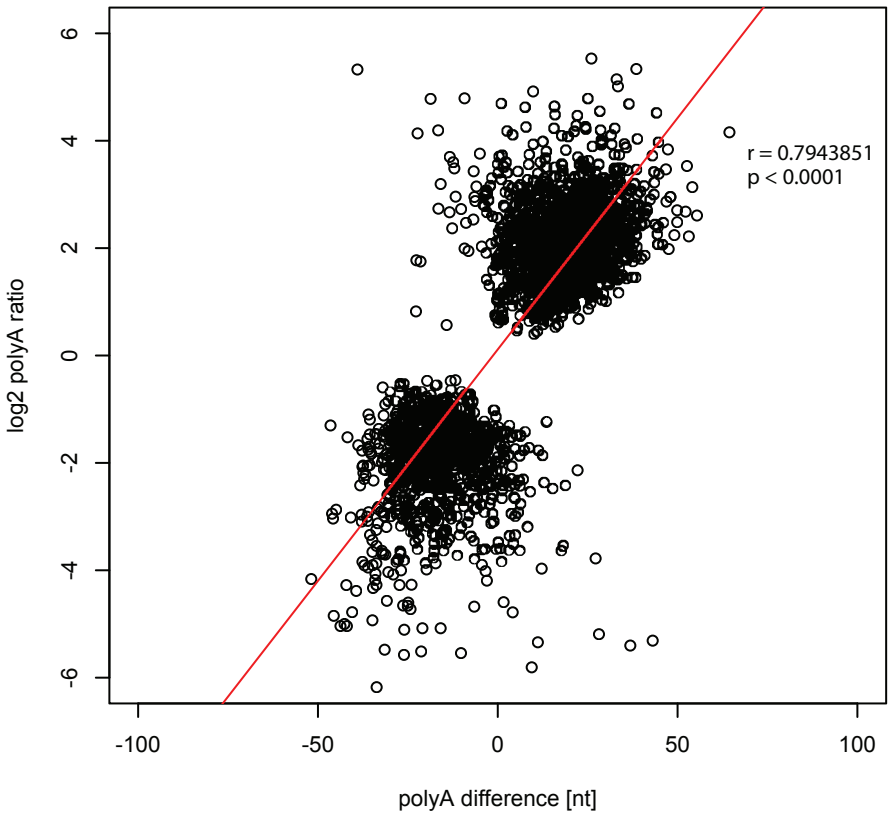


Figure 3.5 polyA length difference is strongly correlated with log2 polyA ratio change. The relative change in polyA length can also be captured by the normalized counts of polyA reads over the same region. Spearman correlation coefficient is shown (r= 0.794).

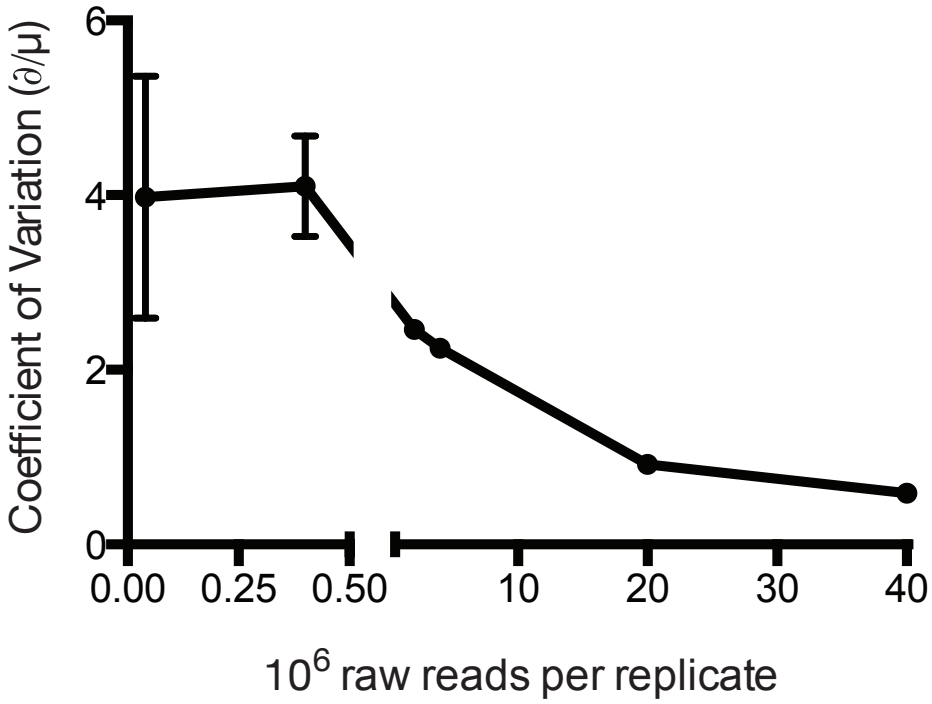


Figure 3.6 Coefficient of variation plot based on sequencing depth for the 45-55 percentile most polyadenylated transcripts.

3.5 References

1. Crick, F. Central dogma of molecular biology. Nature 227, 561-563 (1970).

2. Goldstrohm, A. C. & Wickens, M. Multifunctional deadenylase complexes diversify mRNA control. Nat Rev Mol Cell Biol 9, 337-344, doi:10.1038/nrm2370 (2008).

3. Weill, L., Belloc, E., Bava, F. A. & Mendez, R. Translational control by changes in poly(A) tail length: recycling mRNAs. Nat Struct Mol Biol 19, 577-585, doi:10.1038/nsmb.2311 (2012).

4. Guzeloglu-Kayisli, O. et al. Embryonic poly(A)-binding protein (EPAB) is required for oocyte maturation and female fertility in mice. The Biochemical journal 446, 47-58, doi:10.1042/BJ20120467 (2012).

5. Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. Gene 265, 11-23 (2001).

6. Beilharz, T. H. & Preiss, T. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. RNA 13, 982-997, doi:10.1261/rna.569407 (2007).

7. Preiss, T., Muckenthaler, M. & Hentze, M. W. Poly(A)-tail-promoted translation in yeast: implications for translational control. RNA 4, 1321-1331 (1998).

8. Bentley, D. L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol 17, 251-256, doi:10.1016/j.ceb.2005.04.006 (2005).

9. Barkoff, A., Ballantyne, S. & Wickens, M. Meiotic maturation in Xenopus requires polyadenylation of multiple mRNAs. The EMBO journal 17, 3168-3175, doi:10.1093/emboj/17.11.3168 (1998).

10. Charlesworth, A., Cox, L. L. & MacNicol, A. M. Cytoplasmic polyadenylation element (CPE)- and CPE-binding protein (CPEB)-independent mechanisms regulate early class maternal mRNA translational activation in Xenopus oocytes. J Biol Chem 279, 17650-17659, doi:10.1074/jbc.M313837200 (2004).

11. Charlesworth, A., Ridge, J. A., King, L. A., MacNicol, M. C. & MacNicol, A. M. A novel regulatory element determines the timing of Mos mRNA translation during Xenopus oocyte maturation. The EMBO journal 21, 2798-2806, doi:10.1093/emboj/21.11.2798 (2002).

12. McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B. & Richter, J. D. Poly(A) elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. Genes Dev 3, 803-815

(1989).

13. Hodgman, R., Tay, J., Mendez, R. & Richter, J. D. CPEB phosphorylation and cytoplasmic polyadenylation are catalyzed by the kinase IAK1/Eg2 in maturing mouse oocytes. Development 128, 2815-2822 (2001).

14. Mendez, R., Barnard, D. & Richter, J. D. Differential mRNA translation and meiotic progression require Cdc2-mediated CPEB destruction. The EMBO journal 21, 1833-1844, doi:10.1093/emboj/21.7.1833 (2002).

15. Paris, J. & Richter, J. D. Maturation-specific polyadenylation and translational control: diversity of cytoplasmic polyadenylation elements, influence of poly(A) tail size, and formation of stable polyadenylation complexes. Mol Cell Biol 10, 5634-5645 (1990).

16. Groisman, I., Huang, Y. S., Mendez, R., Cao, Q. & Richter, J. D. Translational control of embryonic cell division by CPEB and maskin. Cold Spring Harb Symp Quant Biol 66, 345-351 (2001).

17. Mendez, R. & Richter, J. D. Translational control by CPEB: a means to the end. Nat Rev Mol Cell Biol 2, 521-529, doi:10.1038/35080081 (2001).

18. Richter, J. D. Cytoplasmic polyadenylation in development and beyond. Microbiol Mol Biol Rev 63, 446-456 (1999).

19. Roy, L. M. et al. The cyclin B2 component of MPF is a substrate for the c-mos(xe) proto-oncogene product. Cell 61, 825-831 (1990).

20. Song, J. et al. The type II activin receptors are essential for egg cylinder growth, gastrulation, and rostral head development in mice. Dev Biol 213, 157-169, doi:10.1006/dbio.1999.9370.S0012-1606(99)99370-3 [pii] (1999).

21. Yamashita, M. Molecular mechanisms of meiotic maturation and arrest in fish and amphibian oocytes. Semin Cell Dev Biol 9, 569-579, doi:10.1006/scdb.1998.0251 (1998).

22. Salles, F. J., Lieberfarb, M. E., Wreden, C., Gergen, J. P. & Strickland, S. Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs. Science 266, 1996-1999 (1994).

23. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66-71, doi:10.1038/nature13007 (2014).

24. Nix, D. A. et al. Next generation tools for genomic data generation, distribution, and visualization. BMC bioinformatics 11, 455, doi:10.1186/1471-2105-11-455 (2010).

25. Squadrito, M. L. et al. miR-511-3p modulates genetic programs of tumor-associated macrophages. Cell Rep 1, 141-154, doi:10.1016/j.celrep.2011.12.005 (2012).

26. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

CHAPTER 4

MAJOR CHANGES IN mRNA POLY-ADENYLATION ACCOMPANY OOGENESIS AND EARLY EMBRYO DEVELOPMENT IN HUMANS

Chapter 4 is a section part of a larger manuscript in preparation for publication. The authors of this manuscript are Pete Hendrickson, Jessie Dorais, Christian Pflüger, David Nix, Douglas Carrell and Bradley Cairns. This entire section is my contribution to the paper.

4.1 Introduction

In female mammals such as humans, meiosis can halt at prophase I, creating germinal vesicles with low transcriptionally activity. However, entry into the M1 phase is accompanied by transcriptional silencing, which has two important consequences: Firstly, all RNA transcripts that are required for germ cell specification and for processes that drive early pretranscriptional zygotic development must be inherited directly; and secondly, any regulation of inherited RNA transcripts must involve posttranscriptional mechanisms. A marquee feature of transcript stability and translational control in early development is differential poly-Adenylation (poly(A)). Recently, poly(A) dynamics were shown to play a significant role during early zebrafish development[1]. Further, studies in xenopus[2] and drosophila[3] have shown that key players such as cyclin B1, Aurora A kinase and Mos[4] are critical for the final steps in oocyte maturation as well as priming the cytoplasm with proteins critical in early embryo development. Hence, post-transcriptional regulation of mRNAs using differential polyA in late stage oocytes is paramount. However, very little is known about the polyA dynamics during early development in either mouse or human.

4.2 Methods

To address this issue, our lab obtained RNA extracted from late-stage human oocytes as well as early human embryos, obtained from IRB-consented couples. We performed total RNAseq analysis using the TotalScript (epicentre) library preparation that allows for random hexamer priming of the reverse transcript reaction under conditions of small amounts of input RNA in the low nanogram range. This in turn enabled us to examine mRNA changes as well as differential poly(A) of these transcripts. To investigate poly(A) changes within our total RNAseq datasets (which were not oligo(dT) selected and hence have no bias for poly(A)), we have

developed an innovative software tool set. Briefly, the software package, PANDA, released as part of the USeq package[5], parses raw sequencing reads for possible 3' polyA and saves these reads in a separate file as well as trims these potentially non-genomic polyA nucleotides in silico from that very read. All raw reads including the trimmed reads are then subjected to alignment against the reference genome; in this case, we used the human genome version 19 (hg19) as reference. Further, the aligned reads are then checked against the file that contains the original, non-trimmed reads. If there is a match against the file containing the original reads, the polyA length (usually between 6-101 bp) is then determined from that original read and added to the SAM alignment file as a user defined tag (At:i:length_of_ polyA, e.g. At:i:62 with a polyA length of 62 nt). Genomic polyA stretches present in certain reads are filtered out at this point. Remarkably, polyA length can reach the maximum sequencing length (in this case up to 101 nt of polyA) provided the read's mate pair can be uniquely aligned. The third and last step involves quantification and comparison of polyA levels between two different stages. Briefly, polyA reads are counted over the last exon and normalized by total reads aligned over the same exon, yielding the polyA-ratio. Then, the polyA-ratios of two different conditions are divided and log2 transformed. Statistics were performed by using a chi-square test on polyA counts of two developmental conditions, followed by multiple test correction (Benjamini-Hochberg) and negative log10 transformation.

4.3 Results and Discussion

As shown in Figure 4.1, there are two major waves of increased polyA are detectable. Firstly, there is a significant ($p<0.0001$) increase in polyA length of transcripts in the maturing oocyte (MI phase) from an average of ~13.9 nt to ~27.9 nt that, notably, are then lost or reduced in the mature MII oocyte (~10.9 nt). This is thought to allow for translation of proteins critical in maturing the oocyte as well

as in the earliest phase of embryo development. Notably, the genome of late stage human oocytes and human early embryos until the 4-cell stage is transcriptionally inactive. The activation of de novo transcription is thought to start in humans between the 4-cell and 8-cell stage, termed embryonic genome activation (EGA)[6]. Secondly, a sharp increase in polyA length ($p<0.0001$) is detectable right after EGA from 10.6 nt polyA to 26.8 nt polyA in cleavage. Clustering analysis (Cluster 3.0 and Java TreeView) with transcripts that changed polyA dynamics during any time in development with an FDR<10e-5 revealed polyA dynamics that correlate with oocyte development (Figure 4.2 - cluster 2, $p<0.0005$). Remarkably, this novel polyA analysis picked up previously published key transcripts that were shown to be critical for oocyte maturation[2,7-10].

Consequently, the first set of transcripts identified to gain polyA during oocyte maturation fall into the category of transcripts that are known to gain polyA based on previous work done in drosophila, xenopus and bos taurus[2,7-10]. Notable examples of transcripts involved in cell cycle regulation and critical in obtaining cytosolic polyadenylation are CyclinB1 (CCNB1), DAZL, PABPC1L (poly(A) binding protein, cytoplasmic 1-like) (Figure 4.2, 4.3, 4.4 and 4.5, respectively), Aurora A kinase (AURKA) (Figure 4.6) and Mos (Figure 4.7). As shown in the genome snapshots in Figures 4.6 and 4.7 (total reads are depicted in charcoal color and reads containing polyA in orange), there are distinct peaks of polyA reads in the MI oocyte phase. Remarkably, mature MII oocytes have reduced polyA peaks compared to the MI stage. This is possibly due to posttranscriptional regulation of mRNA shutting down translation of transcripts by reducing the polyA tail length when the appropriate amount of protein is made. To our knowledge, we are the first to show this resolution of polyA dynamics in mammals.

Further, we were able to identify polyA gain in the MI stage for proteins critical for oocyte functions. The majority of these transcripts have not been previously

shown to gain cytosolic polyA. However, since these transcripts are known to be essential for oocyte maturation and early embryo development, they fall into the category of predicted transcripts for cytosolic polyA regulation. Genes that exhibit an increase in PolyA during the final stage of maturation include histone H1 oocyte variant (H1FOO, Figure 4.8), growth defect factor 9 (GDF9, not shown), the sperm receptors zona pelucida 1 and 2 (ZP1/2, shown in Figure 4.9 and Figure 4.10) and the endopeptidase responsible for cleaving ZP2 into its physiologically active form, and astacin-like metallo-endopeptidase (ASTL, Figure 4.11). A western blot analysis of proteins from GV and MII stage showed an increase in cleaved ZP2 protein (Figure 4.12), suggesting a functional correlation between polyA length gain from GV to MI for ZP2 (ZP2, Figure 4.13) and an increase in ZP2 protein abundance. Since these genes are well known to be essential for oocyte maturation and fertilization, it is not surprising to see these transcripts gain polyA at this critical point in oocyte maturation. Importantly, transcripts essential at the early zygote stage prior to transcriptome activation, such as DPPA3, (PGC7/Stella, developmental pluripotency associated 3 Figure 4.14), involved in protecting the maternal genome from active DNA demethylation[11], DPPA5 (developmental pluripotency associated 5, Figure 4.15)[12,13] and OOEP (oocyte expressed protein, Figure 4.16), critical in early embryo development[14], showed a similar pattern of polyA increase at the MI stage similar to AURKA for example. Further, TRIM28 (Figure 4.17), essential in maintaining imprinting patterns and retro-transposon regulation[15,16] after EGA, also showed an increase in polyA at the cleavage stage.

Finally, this novel approach to polyA analysis enabled us to identify transcriptions factors, DNA binding proteins and other genes of interest (Table 4.1) that are notoriously understudied but may have a critical importance in zygote development. This set of transcripts are considered novel candidates for cytosolic polyA regulation and possibly play an important role in oocyte maturation as well

as early embryo development. A noteworthy example is TRIM6 (Figure 4.18), which exhibits similar polyA dynamics to c-MOS or ZP2. The intersection of the list of transcripts gaining strong polyA from GV to MI with a list of known transcription factors and DNA binding proteins (HUGO DNA binding proteins) revealed ZNF770 (Figure 4.19 and Table 4.1, cluster 5) and SOX13 as potentially interesting and important transcription factors for setting up the genome before EGA. Another notable candidate in regulating p53 at the late oocyte maturation stage as well as zygote stage is PDCD5 (programmed cell death 5, Figure 4.20). As shown in the top 25 transcripts by FDR gaining polyA from GV to MI (Figure 4.13), PDCD5 is present in that list with aforementioned transcripts critical in oocyte and zygote development. PCNA (Proliferating Cell Nuclear Antigen, Figure 4.21) is a key protein involved in DNA replication[17] and it is also significantly polyadenylated during the GV to MI transition (Figure 4b). Notably, transcripts gaining polyA from GV to MI stage are significantly ($p<0.0001$) enriched for CPE sites in their 3'UTRs vs all known 3'UTRs (Table 4.2)[18-21].

In summary, we developed a new method for analyzing polyA dynamics in early development that can also be applied to a variety of other datasets or cell types that are not oligodT selected. We were able to show an inventory of transcripts that receive polyadenylation during the early phase of egg maturation (MI) as well as loss of it during the last stage of oocyte maturation (MII). We observed a statistically significant enrichment ($p<0.05$) of transcripts critical for oocyte physiology and early embryo development (Figure 4.2, cluster 2 and Table 4.3). Since it is impossible to functionally test transcripts in human oocytes, we implicitly tested this novel polyA analysis by checking hallmark transcripts such as CCNB1, AURKA and c-MOS for polyA gain. By confirming these notable positive controls in our polyA analysis, we suggest that other transcripts identified in our analysis could also be important targets of cytosolic polyadenylation in human

oocytes. Finally, this polyA analysis suggests the importance of TFs such as Sox13 and ZNF770 for the start of embryo genome activation (EGA). Future studies will need to be performed in order to these factors for their importance during EGA.

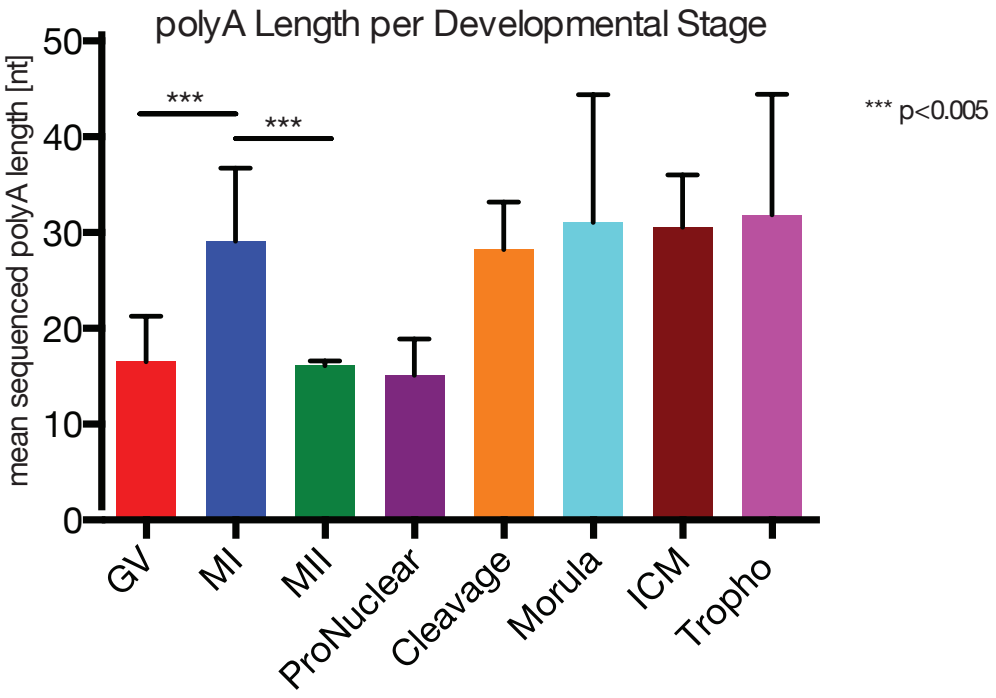


Figure 4.1 Poly differences during developmental stages. Mean PolyA length measured in 4 technical replicates per developmental stage, paired T-test between PolyA values from identical library preparations.

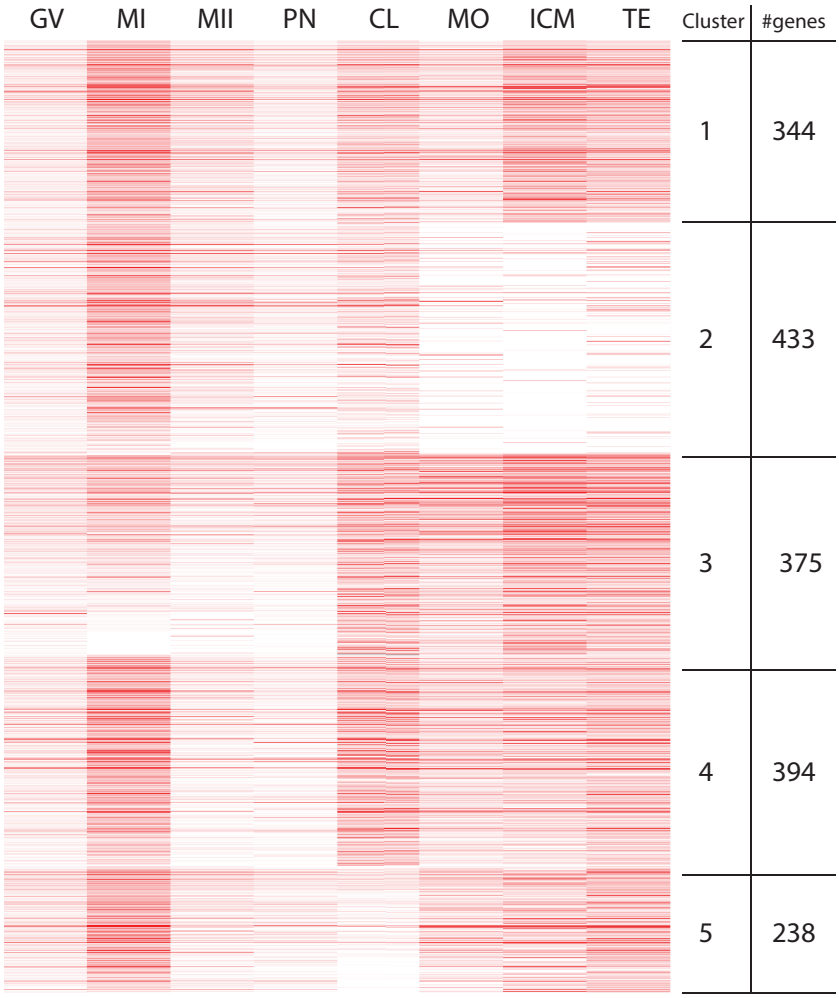


Figure 4.2 Clustering of transcripts that change in PolyA with transformed FDR > = 40

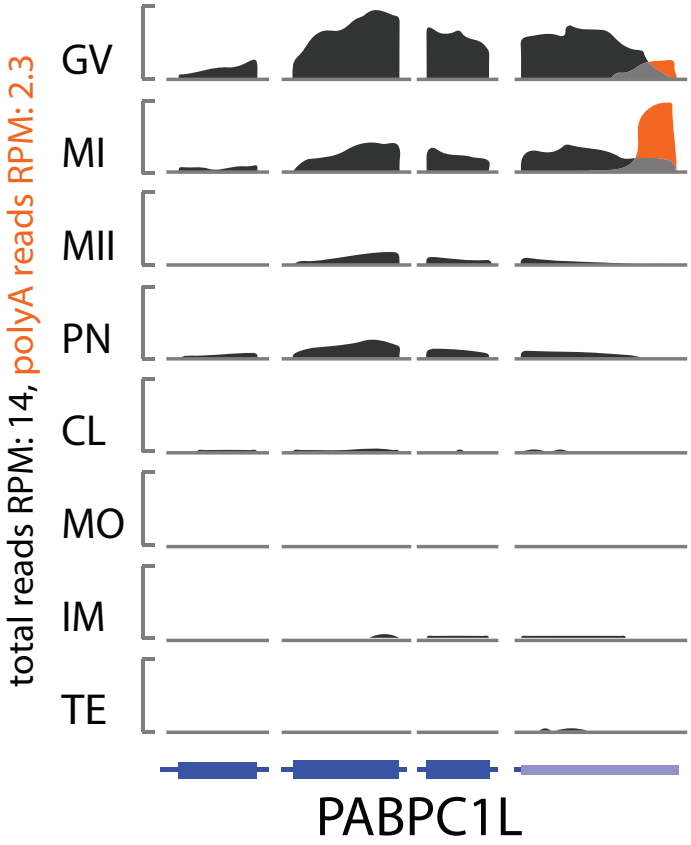


Figure 4.3 Genome snapshot of PABPC1L gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

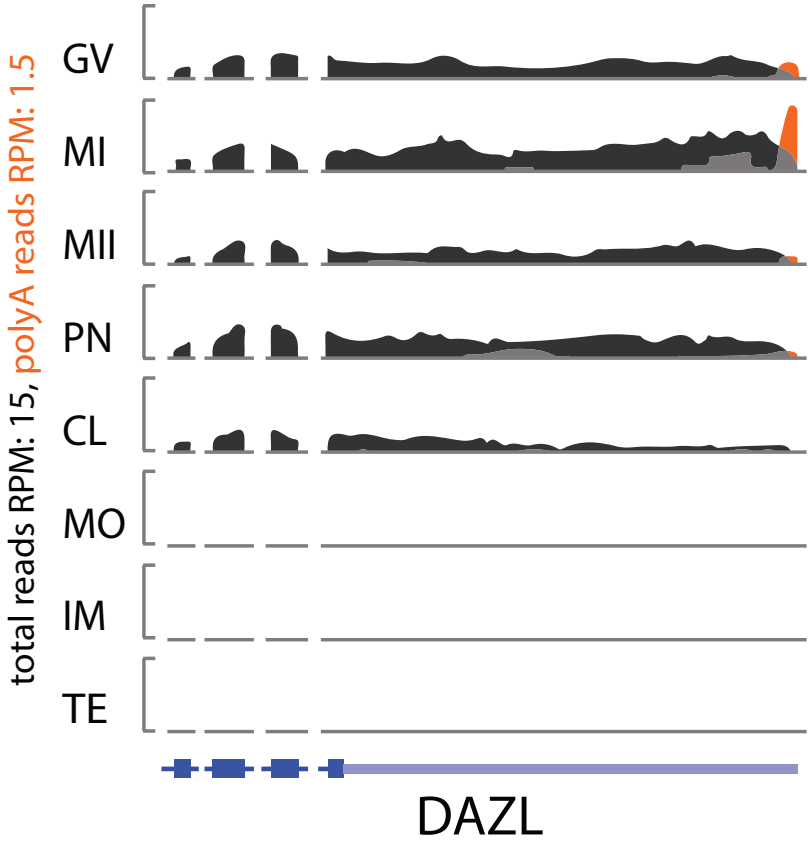


Figure 4.4 Genome snapshot of DAZL gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

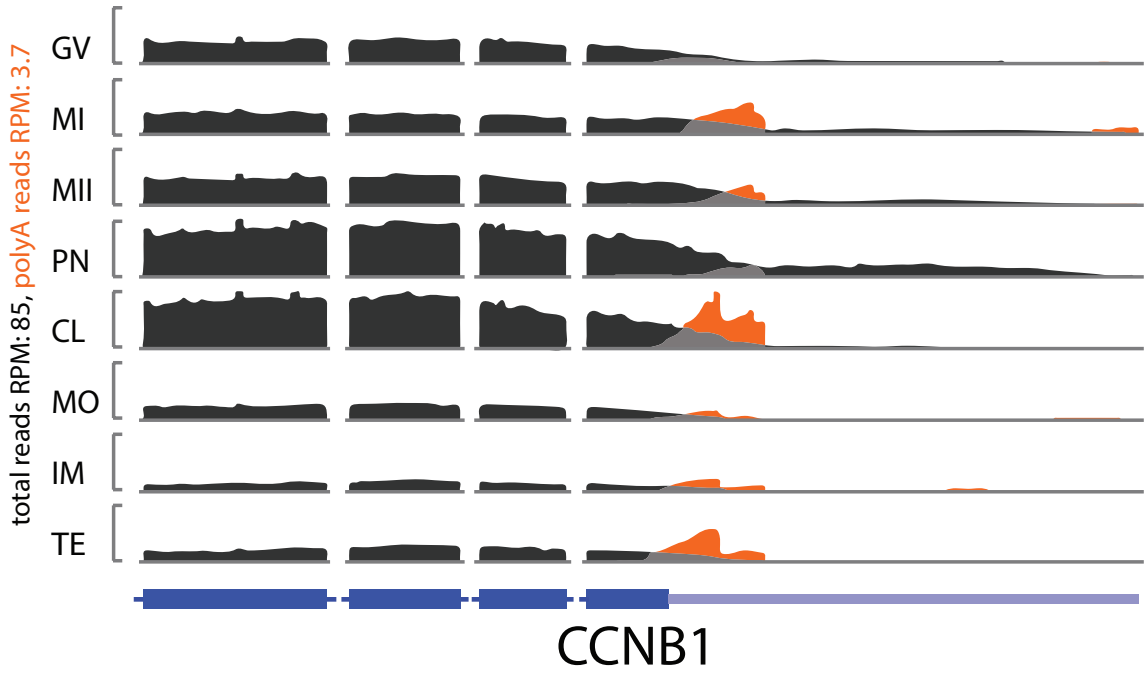


Figure 4.5 Genome snapshot of CCNB1 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation and early embryogenesis. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

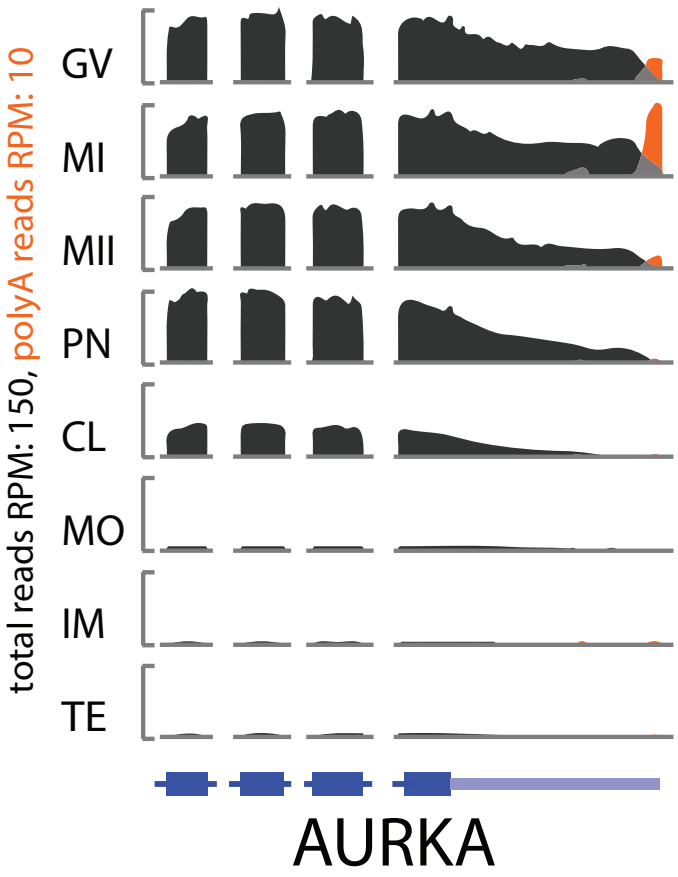


Figure 4.6 Genome snapshot of AURKA gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

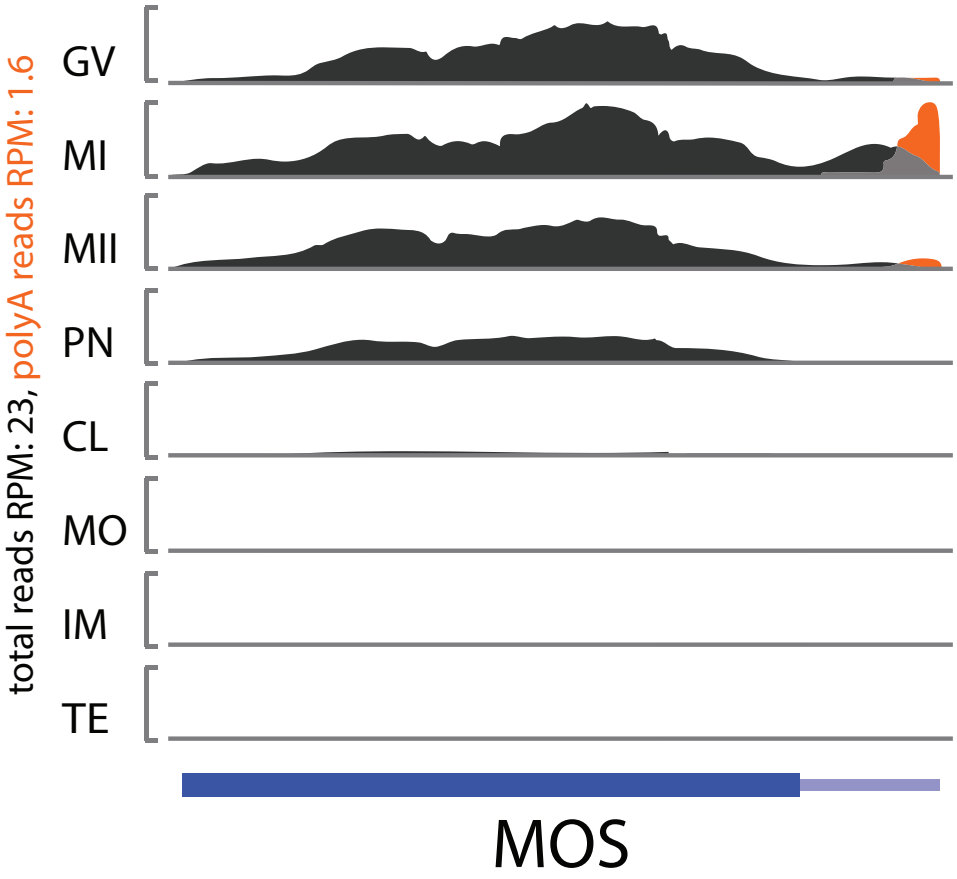


Figure 4.7 Genome snapshot of MOS gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

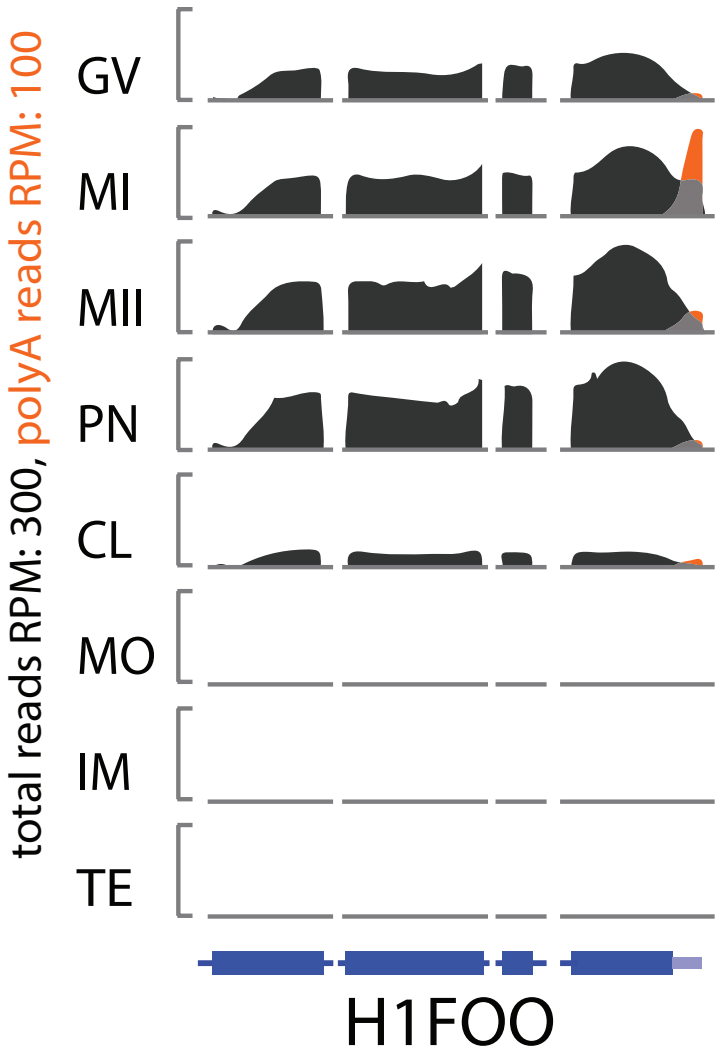


Figure 4.8 Genome snapshot of H1FOO gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

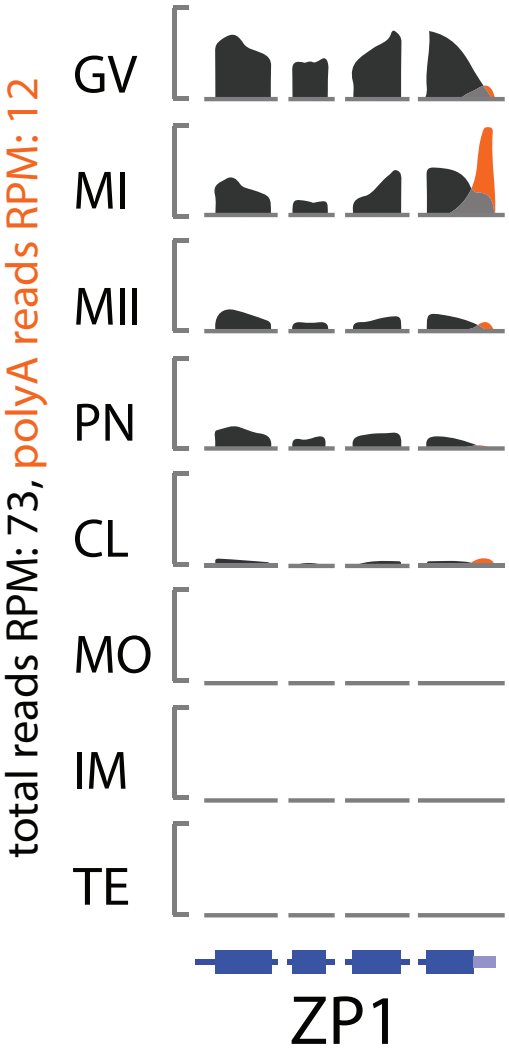


Figure 4.9 Genome snapshot of ZP1 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

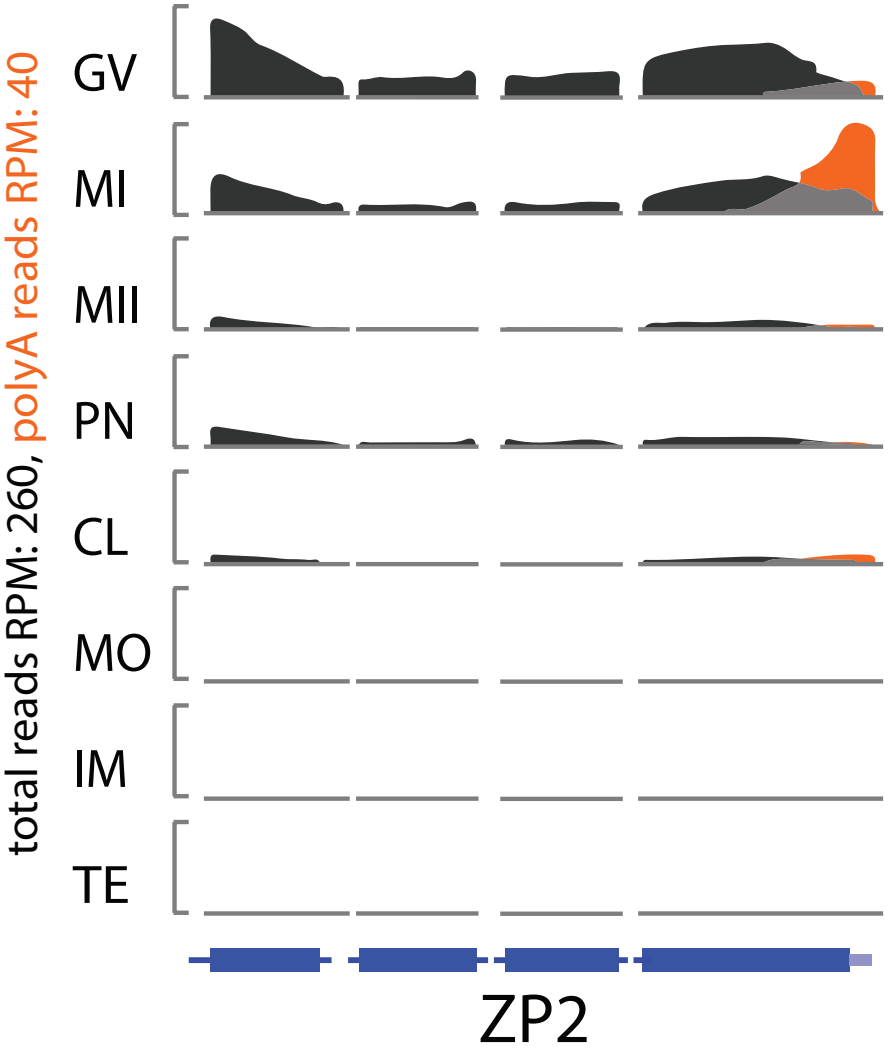


Figure 4.10 Genome snapshot of ZP2 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

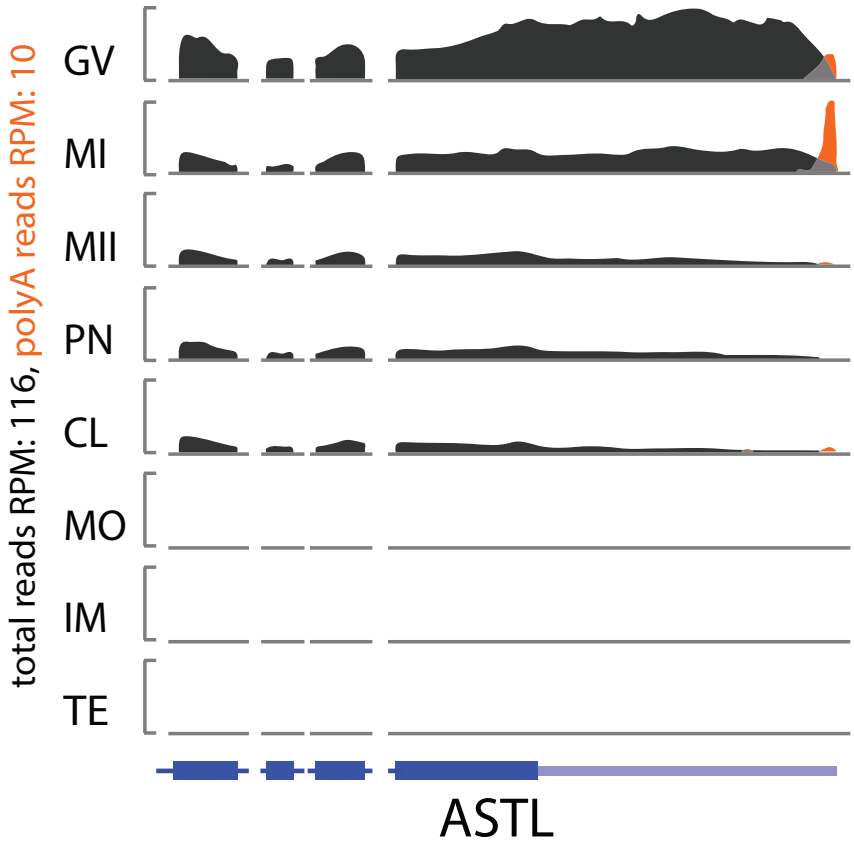


Figure 4.11 Genome snapshot of ASTL gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

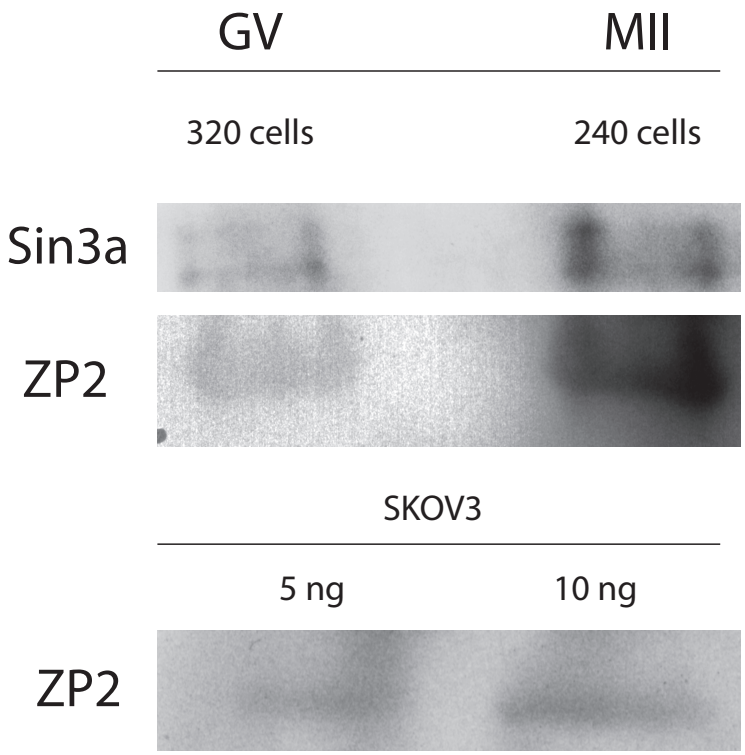


Figure 4.12 Western blot analysis showing increase of ZP2 protein levels from GV to MII stage. Sin3a protein was used as a loading control.

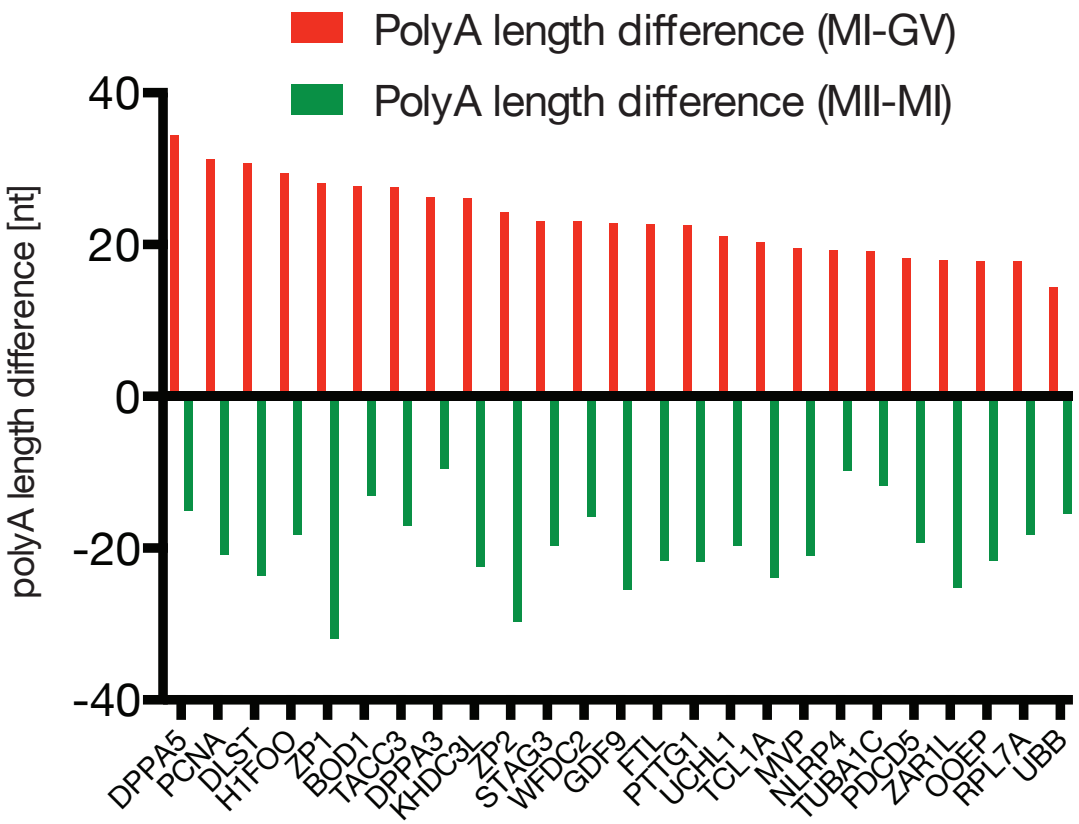


Figure 4.13 PolyA differences during developmental stages. Log2 polyA count differences for the top 25 genes (selected by smallest FDR filtering) from GV to MI, sorted in descending order by log2 polyA differences.

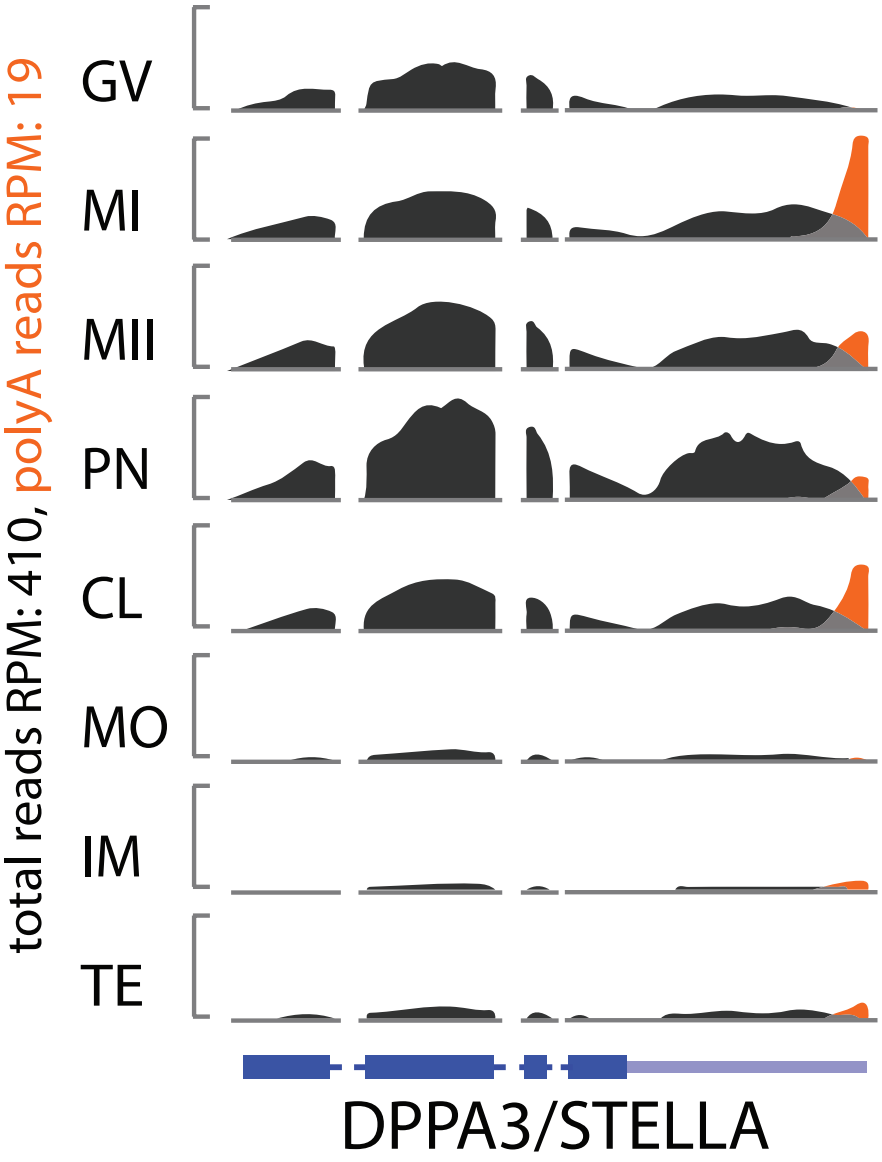


Figure 4.14 Genome snapshot of DPPA3 / STELLA gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation and early embryongenesis. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

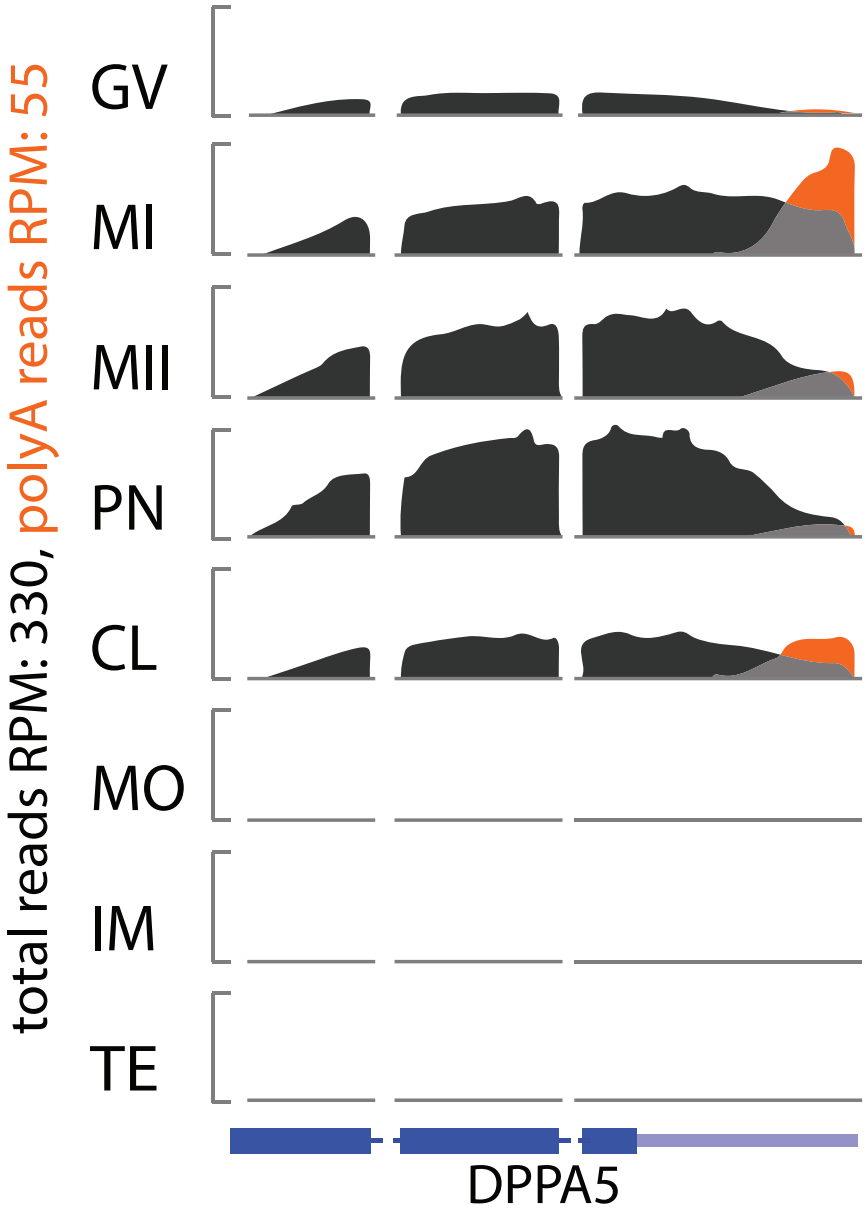


Figure 4.15 Genome snapshot of DPPA5 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation and early embryogenesis. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

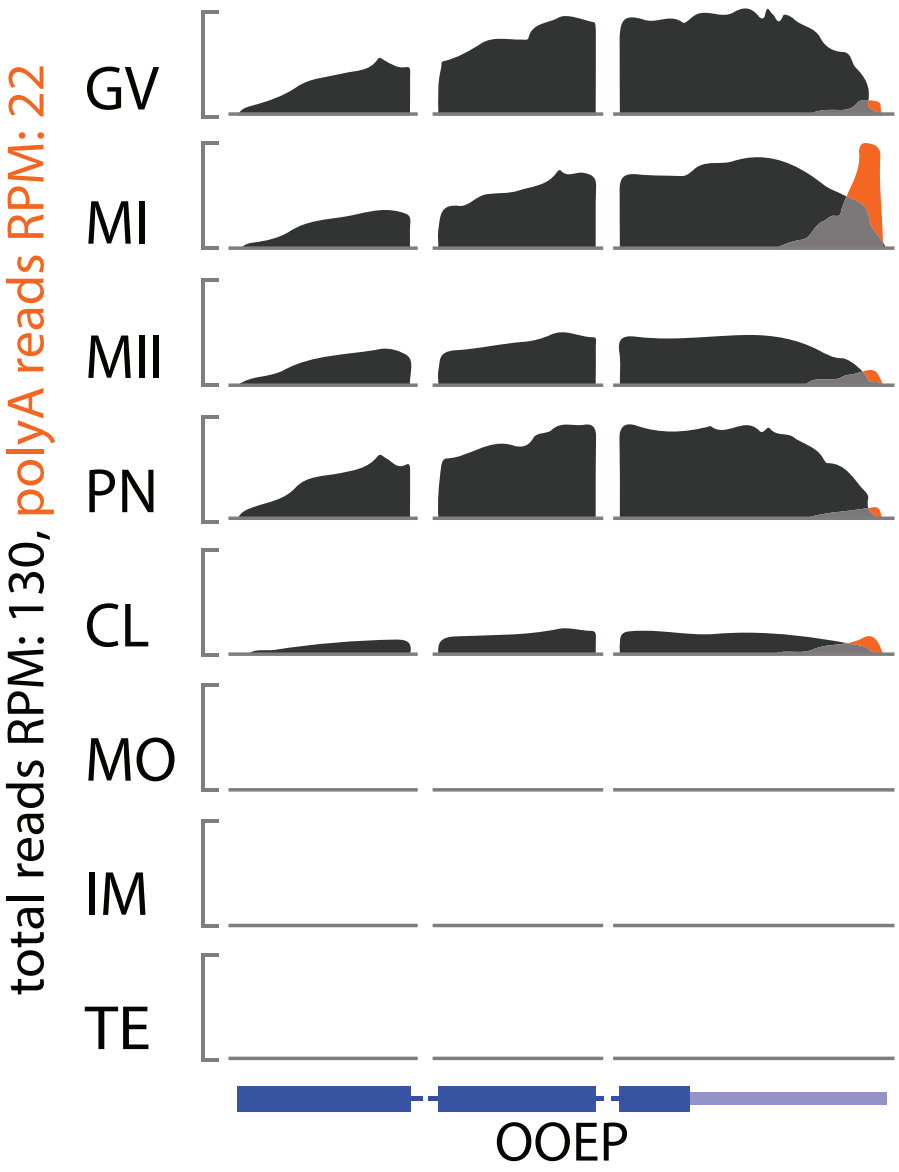


Figure 4.16 Genome snapshot of OOEP gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

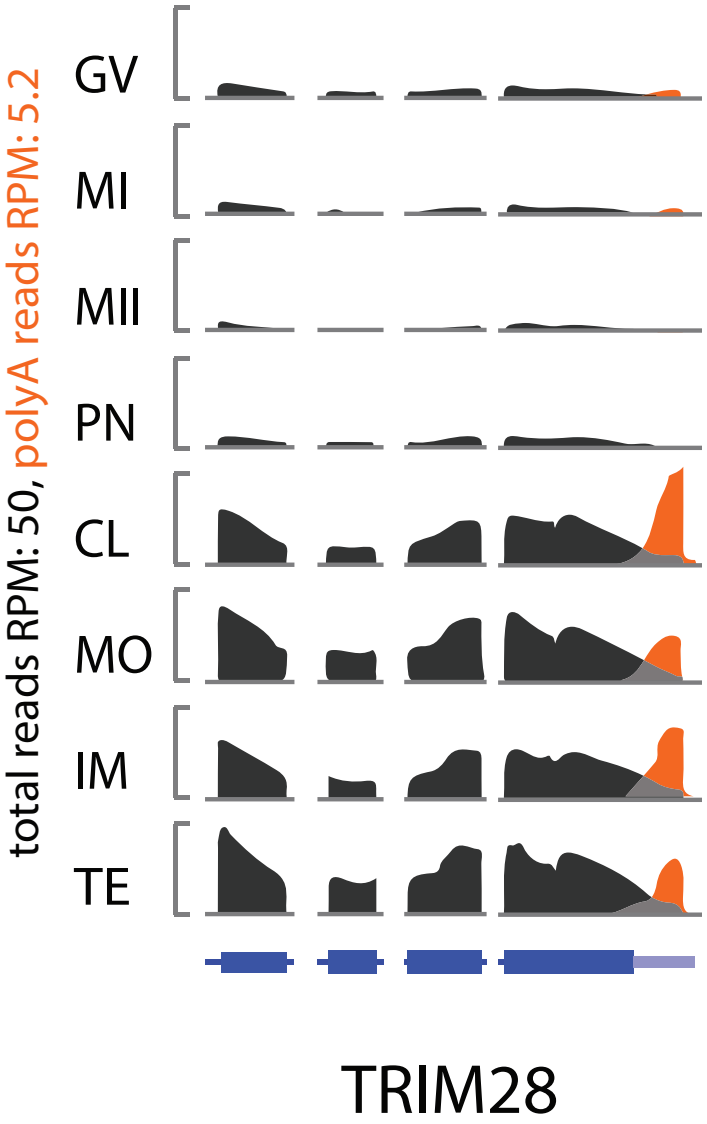


Figure 4.17 Genome snapshot of TRIM28 gene. Figure shows up to the last four exons of the gene known to alter polyA during early embryogenesis. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

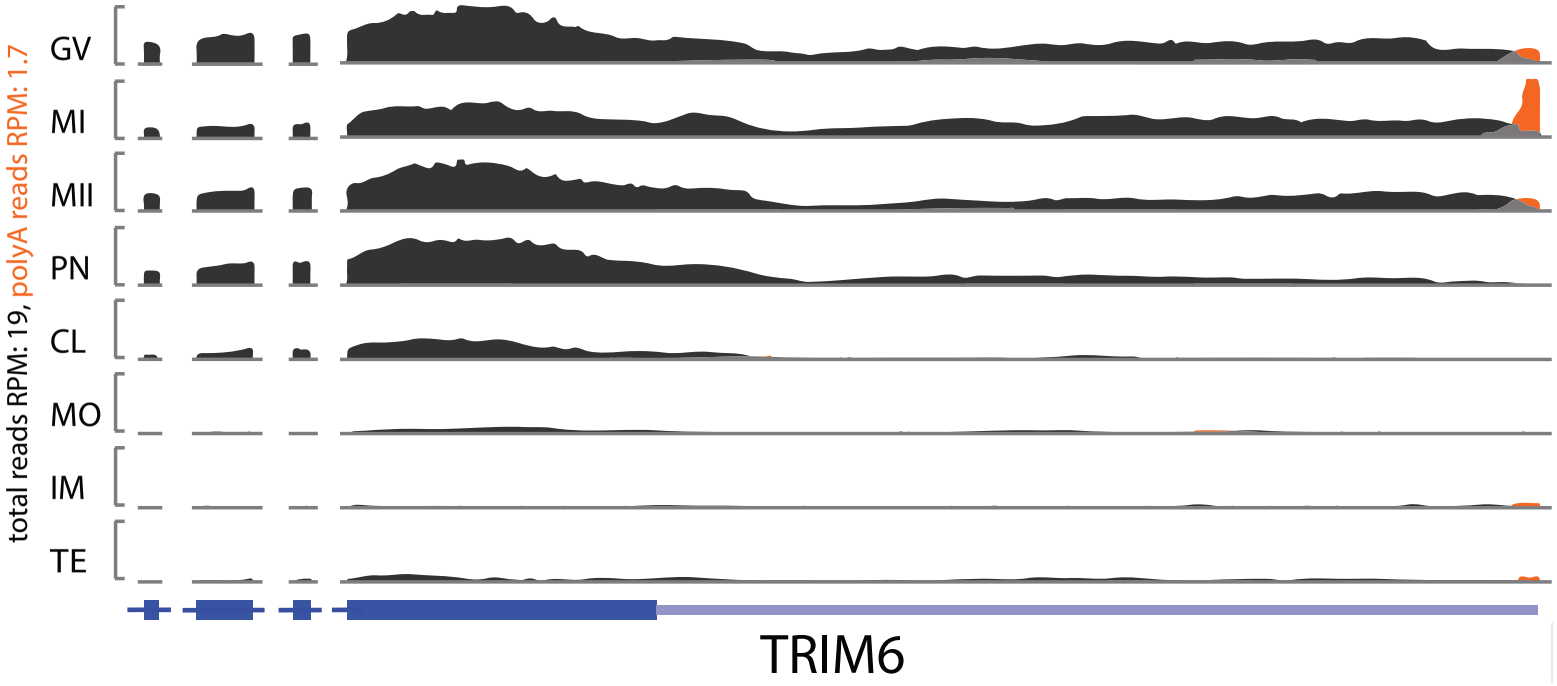


Figure 4.18 Genome snapshot of TRIM6 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

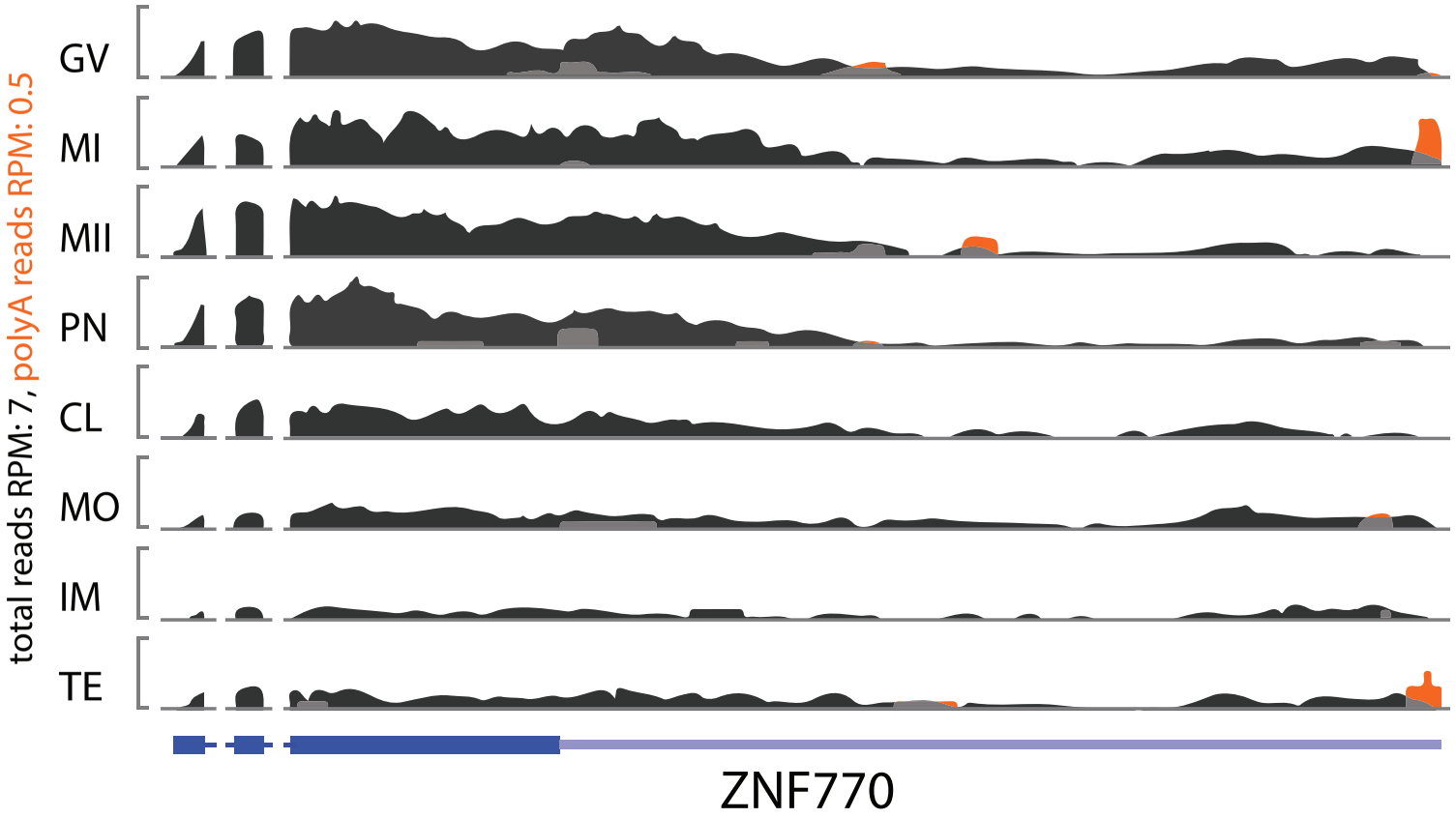


Figure 4.19 Genome snapshot of ZNF770 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation and early embryogenesis. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

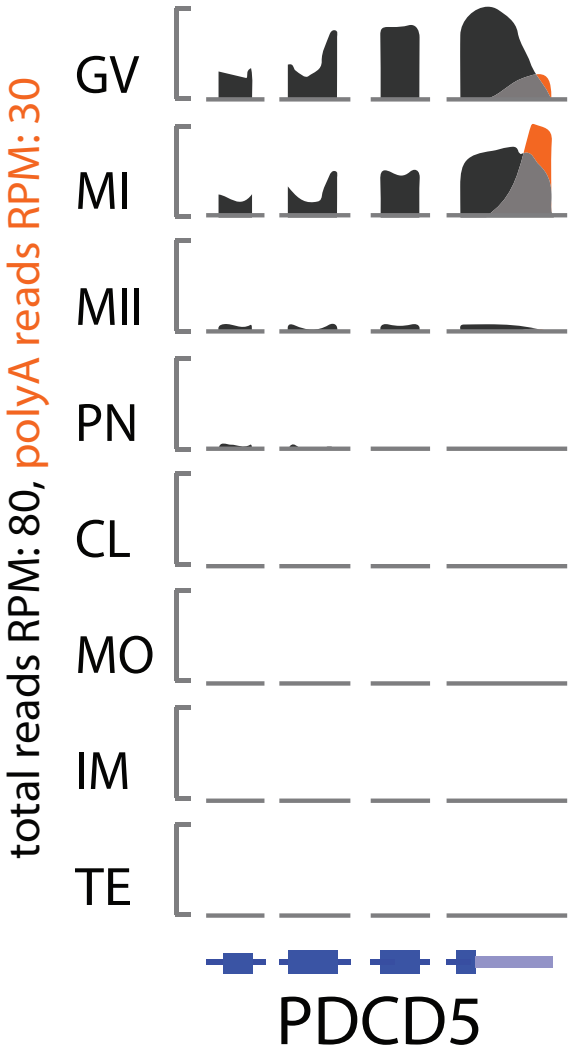


Figure 4.20 Genome snapshot of PCDC5 gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

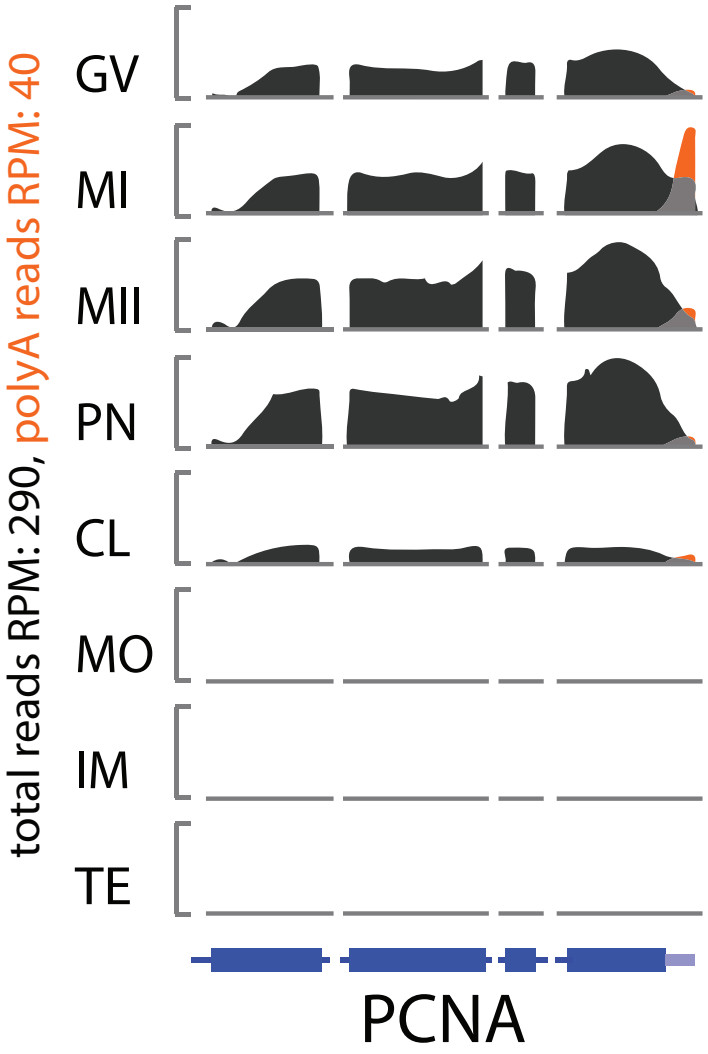


Figure 4.21 Genome snapshot of PCNA gene. Figure shows up to the last four exons of the gene known to alter polyA during oocyte maturation. Total reads and polyA reads are shown in charcoal and orange, respectively, overlapping reads in light gray.

Table 4.1

Transcription Factors and DNA binding proteins in PolyA Clusters

Cluster	Ensemble Gene ID	Gene Symbol	Description
1	ENSG00000126767	ELK1	ETS domain-containing protein Elk-1
	ENSG00000157404	KIT SCFR	Mast/stem cell growth factor receptor Kit
	ENSG00000143842	SOX13	Transcription factor SOX-13
	ENSG00000159479	MED8	Mediator of RNA polymerase II transcription subunit 8
	ENSG00000169714	CNBP RNF163 ZNF9	Cellular nucleic acid-binding protein
	ENSG00000059728	MXD1 MAD	Max dimerization protein 1
	ENSG00000145741	BTF3 NACB OK/SW-cl.8	Transcription factor BTF3
	ENSG00000198604	BAZ1A ACF1 WCRF180 HSPC317	Bromodomain adjacent to zinc finger domain protein 1A
	ENSG00000116288	PARK7	Protein DJ-1
	ENSG00000115548	KDM3A	Lysine-specific demethylase 3A
	ENSG00000100410	PHF5A	PHD finger-like domain-containing protein 5A
	ENSG00000169375	SIN3A	Paired amphipathic helix protein Sin3a
	ENSG00000143947	RPS27A UBA80 UBCEP1	Ubiquitin-40S ribosomal protein S27a
	ENSG00000137193	PIM1	Serine/threonine-protein kinase pim-1
2	ENSG00000170854	MINA	Bifunctional lysine-specific demethylase and histidyl-hydroxylase MINA
	ENSG00000221818	EBF2	Transcription factor COE2
	ENSG00000143006	DMRTB1	Doublesex- and mab-3-related transcription factor B1
	ENSG00000075142	SRI	Sorcin
	ENSG00000176399	DMRTA1 DMO	Doublesex- and mab-3-related transcription factor A1
	ENSG00000160563	MED27 CRSP34 CRSP8	Mediator of RNA polymerase II transcription subunit 27
	ENSG00000135638	EMX1	Homeobox protein EMX1
	ENSG00000187325	TAF9B TAF9L	Transcription initiation factor TFIID subunit 9B
	ENSG00000173153	ESRRA	Steroid hormone receptor ERR1
	ENSG00000262621		Uncharacterized protein
	ENSG00000163874	ZC3H12A MCPIP MCPIP1	Ribonuclease ZC3H12A
	ENSG00000090447	TFAP4	Transcription factor AP-4
	ENSG00000112333	NR2E1	Nuclear receptor subfamily 2 group E member 1
	ENSG00000174928	C3orf33 MSTP052	Protein C3orf33
	ENSG00000120738	EGR1 KROX24 ZNF225	Early growth response protein 1

Table 4.1 continued

Cluster	Ensemble Gene ID	Gene Symbol	Description
3	ENSG00000178177	LCORL MLR1 hCG_1660774	Ligand-dependent nuclear receptor corepressor-like protein
	ENSG00000179348	GATA2	Endothelial transcription factor GATA-2
	ENSG00000029363	BCLAF1	Bcl-2-associated transcription factor 1
	ENSG00000152518	ZFP36L2 BRF2 ERF2 RNF162C TIS11D	Zinc finger protein 36, C3H1 type-like 2
	ENSG00000159592	GPBP1L1 SP192	Vasculin-like protein 1
	ENSG00000245848	CEBPA	CCAAT/enhancer-binding protein alpha
	ENSG00000148840	PPRC1 KIAA0595	Peroxisome proliferator-activated receptor gamma coactivator-related protein 1
	ENSG00000148835	TAF5 TAF2D	Transcription initiation factor TFIID subunit 5
	ENSG00000141384	TAF4B hCG_38478	HCG38478, isoform CRA_a
	ENSG00000006576	PHTF2	Putative homeodomain transcription factor 2
	ENSG00000135164	DMTF1 DMP1	Cyclin-D-binding Myb-like transcription factor 1
	ENSG00000062194	GPBP1 hCG_40617	GC-rich promoter binding protein 1, isoform CRA_c
	ENSG00000143013	LMO4	LIM domain transcription factor LMO4
	ENSG00000136826	KLF4	Krueppel-like factor 4
	ENSG00000102804	TSC22D1	TSC22 domain family protein 1
	ENSG00000148308	GTF3C5	General transcription factor 3C polypeptide 5
	ENSG00000185043	CIB1 CIB KIP PRKDCIP	Calcium and integrin-binding protein 1
	ENSG00000129194	SOX15 SOX12 SOX20 SOX26 SOX27	Protein SOX-15
	ENSG00000100393	EP300 P300	Histone acetyltransferase p300
	ENSG00000084093	REST hCG_1746842	RE1-silencing transcription factor
	ENSG00000196428	TSC22D2	TSC22 domain family protein 2
	ENSG00000115207	GTF3C2 KIAA0011	General transcription factor 3C polypeptide 2
	ENSG00000101843	PSMD10	26S proteasome non-ATPase regulatory subunit 10
	ENSG00000120690	ELF1	ETS-related transcription factor Elf-1
	ENSG00000111269	CREBL2	cAMP-responsive element-binding protein-like 2

Table 4.1 continued

Cluster	Ensemble Gene ID	Gene Symbol	Description
4	ENSG00000101057	MYBL2 BMYB	Myb-related protein B
	ENSG00000089902	RCOR1 KIAA0071 RCOR	REST corepressor 1
	ENSG00000138095	LRPPRC LRP130	Leucine-rich PPR motif-containing protein, mitochondrial
	ENSG00000105887	MTPN	Myotrophin
	ENSG00000163132	MSX1 HOX7	Homeobox protein MSX-1
	ENSG00000130726	TRIM28	Transcription intermediary factor 1-beta
	ENSG00000102241	HTATSF1	HIV Tat-specific factor 1
	ENSG00000128272	ATF4 CREB2 TXREB	Cyclic AMP-dependent transcription factor ATF-4
	ENSG00000113387	SUB1	Activated RNA polymerase II transcriptional coactivator p15
	ENSG00000197063	MAFG	Transcription factor MafG
	ENSG00000167182	SP2 KIAA0048	Transcription factor Sp2
	ENSG00000136504	KAT7	Histone acetyltransferase
	ENSG00000196924	FLNA FLN FLN1	Filamin-A
	ENSG00000133398	MED10 L6 TRG17 TRG20	Mediator of RNA polymerase II transcription subunit 10
	ENSG00000189403	HMGB1 HMG1	High mobility group protein B1
	ENSG00000188243	COMMD6 MSTP076	COMM domain-containing protein 6
	ENSG00000123562	MORF4L2 KIAA0026 MRGX	Mortality factor 4-like protein 2
	ENSG00000085231	TAF9 TAF2G TAFII31	Transcription initiation factor TFIID subunit 9
	ENSG00000122034	GTF3A	Transcription factor IIIA
	ENSG00000135801	TAF5L PAF65B	TAF5-like RNA polymerase II p300/CBP-associated factor-associated factor 65 kDa
	ENSG00000134317	GRHL1 LBP32 MGR TFCP2L2	Grainyhead-like protein 1 homolog
5	ENSG00000198146	ZNF770	Zinc finger protein 770
	ENSG00000105755	ETHE1 HSCO	Persulfide dioxygenase ETHE1, mitochondrial
	ENSG00000198517	MAFK	Transcription factor MafK
	ENSG00000013810	TACC3	Transforming acidic coiled-coil-containing protein 3
	ENSG00000198911	SREBF2 BHLHD2 SREBP2	Sterol regulatory element-binding protein 2
	ENSG00000182158	CREB3L2 BBF2H7	Cyclic AMP-responsive element-binding protein 3-like protein 2
	ENSG00000183087	GAS6 AXLLG	Growth arrest-specific protein 6
	ENSG00000075426	FOSL2	Fos-related antigen 2
	ENSG00000064313	TAF2 CIF150 TAF2B	Transcription initiation factor TFIID subunit 2
	ENSG00000164916	FOXK1 MNF	Forkhead box protein K1
	ENSG00000112033	PPARD NR1C2 PPARB	Peroxisome proliferator-activated receptor delta
	ENSG00000134107	BHLHE40 BHLHB2 DEC1 SHARP2 STRA	Class E basic helix-loop-helix protein 40
	ENSG00000100207	TCF20 KIAA0292 SPBP	Transcription factor 20
	ENSG00000137575	SDCBP MDA9 SYCL	Syntenin-1

Table 4.2

CPE Sites Enrichment in 3'UTR of Transcripts gaining PolyA increase from GV to MI phase

	CPE+CPSF	no CPE+CPSF	Total
GV to MI polyA gain, FDR<10e-5	1191	567	1758
all known 3'UTRs	26795	125842	152637
Total	27986	126409	154395

Two-tailed Chi-square Test with Yate's Correction : $p < 0.0001$

Table 4.3

Oocyte Maturation Factors present in PolyA cluster 2, $p < 0.0005$

Cluster	Ensemble Gene ID	Gene Symbol	Description
2	ENSG00000188886	ASTL	Astacin-like metalloendopeptidase
	ENSG00000103310	ZP2 ZPA	Zona pellucida sperm-binding protein 2
	ENSG00000164404	GDF9	Growth/differentiation factor 9
	ENSG00000092345	DAZL DAZH DAZL1 DAZLA SP	Deleted in azoospermia-like
	ENSG00000006047	YBX2 CSDA3 MSY2	Y-box-binding protein 2
	ENSG00000149506	ZP1	Zona pellucida sperm-binding protein 1
	ENSG00000203907	OOEP C6orf156 KHDC2 OEP1	Oocyte-expressed protein homolog
	ENSG00000178804	H1FOO H1OO OSH1	Histone H1oo
	ENSG00000116996	ZP4 ZPB	Zona pellucida sperm-binding protein 4

4.4 References

1. Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H. & Bartel, D.P. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66-71 (2014).

2. Barkoff, A., Ballantyne, S. & Wickens, M. Meiotic maturation in Xenopus requires polyadenylation of multiple mRNAs. EMBO J 17, 3168-75 (1998).

3. Salles, F.J., Lieberfarb, M.E., Wreden, C., Gergen, J.P. & Strickland, S. Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs. Science 266, 1996-9 (1994).

4. Charlesworth, A., Ridge, J.A., King, L.A., MacNicol, M.C. & MacNicol, A.M. A novel regulatory element determines the timing of Mos mRNA translation during Xenopus oocyte maturation. EMBO J 21, 2798-806 (2002).

5. Nix, D.A. et al. Next generation tools for genomic data generation, distribution, and visualization. BMC Bioinformatics 11, 455 (2010).

6. Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. Nature 332, 459-61 (1988).

7. Gohin, M., Fournier, E., Dufort, I. & Sirard, M.A. Discovery, identification and sequence analysis of RNAs selected for very short or long poly A tail in immature bovine oocytes. Mol Hum Reprod 20, 127-38 (2014).

8. Guzeloglu-Kayisli, O. et al. Embryonic poly(A)-binding protein (EPAB) is required for oocyte maturation and female fertility in mice. Biochem J 446, 47-58 (2012).

9. Richter, J.D. Cytoplasmic polyadenylation in development and beyond. Microbiol Mol Biol Rev 63, 446-56 (1999).

10. Mendez, R. & Richter, J.D. Translational control by CPEB: a means to the end. Nat Rev Mol Cell Biol 2, 521-9 (2001).

11. Nakamura, T. et al. PGC7/Stella protects against DNA demethylation in early embryogenesis. Nat Cell Biol 9, 64-71 (2007).

12. Pierre, A. et al. Atypical structure and phylogenomic evolution of the new eutherian oocyte- and embryo-expressed KHDC1/DPPA5/ECAT1/OOEP gene family. Genomics 90, 583-94 (2007).

13. Kim, S.K. et al. Identification of developmental pluripotency associated 5 expression in human pluripotent stem cells. Stem Cells 23, 458-62 (2005).

14. Tashiro, F. et al. Maternal-effect gene Ces5/Ooep/Moep19/Floped is essential for oocyte cytoplasmic lattice formation and embryonic development at the maternal-zygotic stage transition. Genes Cells 15, 813-28 (2010).

15. Messerschmidt, D.M. et al. Trim28 is required for epigenetic stability during mouse oocyte to embryo transition. Science 335, 1499-502 (2012).

16. Turelli, P. et al. Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. Genome Res 24, 1260-70 (2014).

17. Moldovan, G.L., Pfander, B. & Jentsch, S. PCNA, the maestro of the replication fork. Cell 129, 665-79 (2007).

18. Roy, L.M. et al. The cyclin B2 component of MPF is a substrate for the c-mos(xe) proto-oncogene product. Cell 61, 825-31 (1990).

19. Nebreda, A.R. & Ferby, I. Regulation of the meiotic cell cycle in oocytes. Curr Opin Cell Biol 12, 666-75 (2000).

20. Yamashita, M. Molecular mechanisms of meiotic maturation and arrest in fish and amphibian oocytes. Semin Cell Dev Biol 9, 569-79 (1998).

21. Sagata, N. Introduction: meiotic maturation and arrest in animal oocytes. Semin Cell Dev Biol 9, 535-7 (1998).

CHAPTER 5

DISCUSSION

5.1 Human Sperm Methylome Changes with Age and Environmental Impacts

The genome in human sperm is tightly packaged and DNA is hypermethylated. Only developmentally poised promoters and enhancers are the notable exceptions where hypomethylation has been documented. In our study, we asked if the DNA methylome in sperm is static or if it does change over time during the physiological process of aging. Overall, our findings suggest that the vast majority of CpG methylation in sperm does not change. We noticed a small but statistically significant global increase in DNAme over time with a rate of >0.4% global DNAme increase per year ($p<0.001$). In contrast to global DNAme increase, we identified regions that get hypomethylated over time in the human sperm genome. Remarkably, some these regions are associated with genes implicated in neuropsychiatric disorders, Dopamine receptor D4 (DRD4; ENSG00000170956) and tenascin XB (TNXB; ENSG00000168477)[1].

One question that needs to be answered is what marks these regions for loss of DNAme? Preliminary data associate regions that lose methylation with regions that retain histones in sperm. One might speculate that either the histone deposition in aging sperm or the histone aberrant histone modification may play a role in hypomethylation of these regions. Another question that remains to be answered is if these regions of hypomethylation will serve as enhancers in differentiated tissue. An interesting hypothesis is that regions losing DNAme in aging sperm serve as enhancers or other regulatory elements in neuronal cells. An important experiment would be to test the regions of hypomethylation functionally in neuronal cells. Recent advancements in induced pluripotent stem cell (iPS cell) generation and differentiation into active neurons[2] might serve as a platform to manipulate the hypomethylated regions in iPS cells and test the function in subsequently differentiated neuronal cells. An alternative mechanism by which these hypomethylated regions could be affected is a change in the non-coding

RNA (ncRNA) repertoire. It is now well established that there is a link of ncRNA and DNAme deposition[3,4]. A hypothesis worth testing is that ncRNAs acting in *cis* to regions of hypomethylation in sperm are either less expressed or not fully processed. Sequencing small RNA and total RNA from human sperm comparing young and aged donor profiles with each other could test this hypothesis. If there is a change detectable in these RNAseq datasets, an experiment manipulating the expression of these ncRNAs in mice could shed some light onto the function of the ncRNAs in respect to DNAme deposition in sperm. The field of aging effects and environmental influences on the sperm methylome is still in its infancy. Recent advancements in targeted and reduced representation genome bisulfite sequencing[5-7] and DNAme array technology[8,9] have enabled us to answer some of the questions, but it can still be quite cost prohibitive to analyze sufficient samples.

Finally, we are planning to test the effects of DNAme altering agents on sperm and the offspring's sperm. A drug used for chemotherapy, 5-aza-cytidine (5azaC), works by irreversibly binding to DNMTs, thus killing their enzyme activity[10,11]. Consequently, cells lose DNA methylation in a passive process, since the reduced availability of DNTMs after the cell division leads to loss of DNAme. This mechanism of DNAme inhibition has been successfully used to treat patients with leukemia or colon cancer[12,13]. However, the important question of 5azaC effects on sperm and the offspring remains to be answered. We are currently in the process of testing this concern by treating male mice with low, medium and high dose of 5azaC. We will test the effect on the sperm of the F0 generation for DNAme aberrations. Importantly, to track parent of origin effects in the offspring, we plan to cross the treated fathers to highly polymorphic females. This enables us to track the DNAme changes in an allele specific manner, allowing us to pinpoint DNAme changes inherited from the father's allele. If changes persist in the sperm of the offspring, then we will continue to track the transgenerational changes into

the F2 and possibly the F3 generation.

This question is of particular interest since the sperm epigenome may be poised to guide early embryo development[14,15]. Consequently, elucidating the role of the very defined DNA methylation profile in sperm in context of early embryo development is critical in furthering our understanding. Equally important is the setup of DNA methylation in oocytes. Recent advancements in single cell bisulfite sequencing[16,17] will help us understand the role of DNA methylation in oocytes. Similarly to alterations in sperm DNAme with 5azaC, an interesting question would be to test oocytes for resistance to 5azaC. It is paramount to understand if DNAme changes related to 5azaC exposure are permanent or if they will be completely repaired after recovery from 5azaC.

These experiments will help educate and advice couples in respect to family planning, where at least one partner had to undergo 5azaC chemotherapy during cancer treatment. While bulk levels of 5mC will most likely recover after 5azaC treatment, it is important to understand if regions in the genome are exempt from this recovery and to what extend DNAme changes will manifest. Testing 5azaC exposure in female mice followed by recovery will give clues of the robustness of DNAme deposition in oocytes. If changes in DNAme persist after recovery from 5azaC, then the next step would involve testing the transgenerational inheritance of said DNAme changes. Similarly to the 5azaC study in male mice, breeding with polymorphic mice will enable us to track alleles in a parent of origin specific manner.

In summary, DNAme changes in gametes and the influences on offspring is a very important question to answer. Previous epidemiology studies suggested a major influence of diet and malnutrition on the offspring from parents exposed to these abnormalities[18,19]. Consequently, it will be interesting to test drugs known to modify epigenetic factors in context of gametes and further, on offspring from

parents exposed to these drugs. Mouse models seem to be the clear choice for answering these important questions as well as patient samples before and after treatment with 5azaC.

5.2 Polyadenylation Changes in Oocytes and Early Embryos Can be Monitored and Tested using PANDA

In Chapters 3 and 4, we detailed a novel approach for investigating polyA changes in total RNAseq datasets. Previous studies in Xenopus and Drosophila established a detail mechanism by which polyA regulation occurs in the cytosol after the genome transitioned from active to an inactive state[20-28]. Cytosolic polyadenylation involves a sequence element in the 3'UTR of the transcript that is recognized by CPEB (cytosolic polyadenylation element binding protein)[29-32]. The challenge for the cell is to maintain and regulate RNAs during a time when no new transcripts are made. Further, the decision and timing for translating the loaded messages into protein is also critical for oocyte maturation as well as setting up early embryo development, with a high priority for instructing genome reactivation.

There is currently no study detailing transcriptome wide polyA changes in maturing oocytes in mammals. One important reason is the limiting material available to study polyA changes in oocytes for example. Recent advancements in single cell RNA sequencing enabled researches to investigate transcript levels and compare them between single cells[33,34]. However, this method relies on oligodT selection which in turn biases against polyA presence at the RNA. Further, single cell sequencing involves polyA trailing on the 5' end of the transcript, changing the naturally occurring polyadenylation sites completely. Fortuitously, we decided to prepare RNA library with a protocol that is independent of oligodT selection and instead uses random hexamer priming for the reverse transcription step in

the protocol. This library preparation method from Epicentre called TotalScript™ bases the ribosomal RNA depletion on a proprietary buffer, which skews the reverse transcription polymerase towards transcribing RNA with comparatively less secondary structure. For example, ribosomal RNAs, which represent vast majority of RNA transcripts (>90%) in the cell, contain extensive secondary structure and are hence less likely reverse transcribed under the aforementioned reverse transcription protocol. Further, the TotalScript™ kit uses a transposases (Tn5) enzyme that will fragment the library and add adapters using much less input DNA than any other commonly used method. Notably, the input amount of total RNA per sample can be as low as 5 ng. This is in stark contrast to total RNAseq with prior RiboZero™ treatment, which requires a minimum input of 100 ng total RNA. Taken together, the TotalScript™ library preparation method enabled us to create and sequence RNA from as low as 17 cells (MI phase) without biasing for the polyadenylation status of transcripts.

The second part of our analysis involved a novel approach at identifying the polyA status of transcripts present at the oocytes stages as well as in the early embryo. Since the polyadenylation of transcripts is a posttranscriptional process, the reference genome will not harbor any information of these added adenines. While alignment programs (software that assigns genomic coordinates to sequenced transcripts) are able to ignore certain homopolymer occurrences, usually reads containing polyA will be thrown out of the analysis and classified as impossible to align. Hence, reads that could potentially be informative for the polyA status are usually discarded. To tackle this issue, we developed a software package called PANDA that will parse the raw sequencing reads for potential polyA. It saves these reads in a separate file, creates a copy and trims the copy of polyA. Subsequently, the trimmed reads are then subjected to the alignment program to obtain genomic coordinates for all raw reads. The next step involves

the parsing of the aligned reads for polyA. As mentioned before, the original, untrimmed reads are saved in a separate file, allowing for the assignment of the aligned reads to polyA, if the aligned read has a matching “original” read. The original read carries the information about the length of the polyA captured. This information is then saved and encoded into the SAM alignment file as a dedicated polyA tag (At:i:lengthOfPolyA). The last step involves the comparison of each transcript’s polyA status between different conditions. Here we investigated different developmental time points and asked if any of the transcripts gained or lost polyA. The PANDA program ‘PolyADifferantialSeq’ works with polyA counts but also outputs the relative polyA length change. As shown and discussed in Chapter 3, polyA length and normalized polyA counts are very well correlated (r= 0.794). This property also enables visualization of normalized polyA counts in the genome browser (see Chapter 4).

We applied PANDA to the human oocyte and early embryo development RNAseq data. Remarkably, we were able to identify >1800 transcripts that gained polyA in the oocytes from the GV to the MI stage with a false discovery rate of less than 10^{-5}. When we analyzed transcripts that gained polyA with an FDR < 10^{-5}, we noticed that previously identified transcripts critical in cytosolic polyadenylation were in the top 100 list of transcripts gaining polyA. Namely, Aurora A kinase (AURKA), MOS and CEPB were all part of mediating polyA gain and were previously shown to receive polyA gain themselves in Xenopus and Drosophila oocytes[29,31,32]. Moreover, proteins known to be essential for oocyte maturation and early embryo development exhibited the same polyA gain from GV to MI as seen for AURKA, MOS and CEPB. For example, OOEP, GDF9, ZP1, ZP2, DPPA3, DAZL and H1FOO are all proteins previously identified in oocyte maturation and early embryo development. H1FOO and DAZL were shown to gain polyA in maturing bovine oocytes but neither of the other proteins was previously reported in gaining

polyA at that stage. Since these proteins are essential for oocyte maturation and early embryo development, it is not surprising to see them gaining polyA.

Here we showed for the first time a transcript wide polyA analysis in maturing oocytes of mammals. Moreover, we also identified a decrease in polyA from MI to MII stage for most of the same transcripts that gained polyA from GV to MI. This transcript wide resolution of polyA dynamics in oocytes has also not been reported before. We postulate that transcripts gaining polyA are subjected to translation but must be turned off afterwards, since resources are limited in the egg and new nutrients will only be replenished after the placenta has formed. Thus, it is essential for the oocyte to turn off translation after the appropriate amounts of proteins are produced. Notably, transcripts only important in oocyte maturation, such as the zona pelucida sperm receptor transcripts ZP1 and ZP2, are not only stripped of their polyA but are also degraded afterwards. In contrast, a protein known to protect the maternal genome from active DNA demethylation, Stella/DPPA3/PGC7, receives polyA from GV to MI, gets deadenylated from MI to MII, but the transcript levels remain stable all the way to the cleavage stage. We speculate that genes falling in the same class as Stella are also required for early embryo development and hence are kept around to conserve precious resources. We identified a third class of transcripts that exhibit the same polyA pattern as transcripts crucial for oocyte maturation and early embryo development. For example, TRIM6 and ZNF770 are part of this class, which also gain polyA from GV to MI and get deadenylated from MI to MII. This class of transcripts is completely understudied and nothing is known in context of oocyte maturation and early embryo development. Analogues to transcripts essential at these developmental time points (AURKA, DPPA3, etc.), we speculate that they are also important factors necessary for oocyte maturation and early embryo development. Future research into polyA changes during mouse late oogenesis may reveal paralogues genes that can then be tested for their

importance in oocyte maturation and early embryo development.

Lastly, we have observed a shortening of 3'UTR usage for transcripts expressed in oocytes. The shortening of the 3'UTRs is also accompanied by a shift in polyA site usage. Previous studies identified alternative polyA site usage or 3'UTR shortening in context of U1 limitation. U1 is part of the splicing machinery and low cellular U1 levels are implicated in driving 3'UTR shortening during the splicing process. Interestingly, we observed low levels of U1 in human oocytes and an overall shortening of 3'UTRs. The Dreyfus group described this dependency previously in neuronal cells and coined the term telescripting[35,36]. A hallmark protein for neuronal activity is HOMER1, whose function is to modulate the depolarization of neurons. The short isoform only contains exons 1 to 5 as well as part of intron 5, where it gets cleaved and polyadenylated. The short HOMER1 isoform acts as a dominant negative, blocking recruitment of the long HOMER1 isoform to the cell membrane[36] The switch between the long and the short isoform is modulated by the presence of U1snRNP in neuronal cells. Remarkably, we identified the same isoform switch of HOMER1 and low U1 levels in human oocytes. After transcription is turned on again in the cleavage stage, U1 levels are increased and HOMER1 is made predominantly full length starting at the morula stage. While no function of HOMER1 is reported in oocytes, the effect of low U1 levels and shortened 3'UTR is also observed in human oocytes. We postulate that oocytes are optimized to deal with transcription and splicing under low U1 levels to produce transcripts with shortened 3'UTRs. A possible explanation would be that the shortened 3'UTRs lack miRNA and piRNA binding sites, thus stabilizing transcripts by protecting them from DICER or Argonaute dependent degradation until the genome becomes active again. Future research will need to test this hypothesis in mice.

5.3 Cancer and Neurobiology Could Benefit from PANDA Analysis

As mentioned before, polyA site usage and 3'UTR length is modulated in development as well as in neuronal cells. Recently, reports of alternative polyA usage and differential polyA were shown in different cancer settings[37,38]. The field is still in the process of understanding why it is beneficial for cancer cells to use polyA modulation similarly to oocytes and neuronal cells. One hypothesis is that the stability and the usage of tumorigenic transcripts are driven by 3'UTR shortening. It is important to note that a hallmark of cancer is to reuse developmental programs[39,40], thus turning on pathways that were meant for early development but not for differentiated cells. Similarly, the use of shortened 3'UTRs and alternative polyA sites observed in oocytes might also adhere to this theme.

Interestingly, neuronal cells also use the toolkit of cytosolic polyA regulation and 3'UTR shortening under limiting U1 levels[41]. It is thought that these features are necessary to regulate and adapt synapses, which can be up to 2 meters apart from the nucleus. Hence, spatial and temporal transcript regulation independent of novel transcription is critical when neuronal cells can span these enormous distances. Consequently, it is of great interest to the field of neurobiology to understand the transcriptome wide use of cytosolic polyA changes.

With the development of PANDA, we demonstrated how this application could be used to investigate polyA changes in human oocyte maturation and early embryo development. In our opinion PANDA could be used to investigate transcriptome wide polyA changes in cancer and neurobiology. The only prerequisite is that RNA from either sample is sequenced for total RNA and is not subjected to oligodT selection. We showed that low amounts of RNA could be used for total RNAseq in conjunction with the TotalScript™ (Epicentre) library preparation. Transcriptional changes as well as relative polyA changes can then be analyzed using the USeq progams DRDS and PANDA, respectively.

5.4 Perspectives and Future Direction

We showed that PANDA analysis could be used to identify transcripts changing their polyA status during human oocyte maturation and early embryo development. However, it is impossible to test these changes in human oocytes. Moreover, manipulations of proteins involved in cytosolic polyadenylation or transcript's 3'UTR are also not feasible options in humans. Consequently, it would be interesting to test these findings in mice. Firstly, a total RNAseq dataset comparable to our human total RNAseq dataset would be necessary to test similarities of polyA regulation between humans and mice. Once that is established, transcripts 3'UTR could be altered by means of genome editing using CRISPR to test for the importance of the CPE for example. Also, proteins and kinases involved in modulating cytosolic polyadenylation could be tested for their necessity and sufficiency in mice. Further, timing and signals for the deadenylation process of transcripts from MI to MII could be tested using PANDA. In summary, applying PANDA to understand oocyte maturation and early embryo development in mice could provide key insights into a fascinating point in development when the material to work with is extremely limited. Alternatively, zebrafish oocytes are much easier to obtain compared to mouse oocytes. Once zebrafish oocytes are staged, they can also be subjected to the PANDA workflow. Recent advancements in genome editing prove also very useful for the work in zebrafish. Similarly to mouse, transcripts and proteins involved in cytosolic polyadenylation could be altered to check their importance in oocyte maturation and early embryo development. An important question in the field is the role of AURKA and CAMK2α phosphorylation of CPEB at Tyr^{147}. Using PANDA, a transcriptome wide effect of polyA depending on each kinase could be teased apart.

In summary, PANDA has great potential in studying transcriptome wide polyA changes where the material is extremely limited. Further, since the workflow

does not involve a specialized sequencing protocol, it dramatically simplifies and thus lowers the bar for transcriptome wide polyA analysis.

5.5 References

1 Jenkins, T. G., Aston, K. I., Pflueger, C., Cairns, B. R. & Carrell, D. T. Age-associated sperm DNA methylation alterations: possible implications in offspring disease susceptibility. PLoS Genet 10, e1004458, doi:10.1371/journal.pgen.1004458 (2014).

2 Denham, M. & Dottori, M. Neural differentiation of induced pluripotent stem cells. Methods Mol Biol 793, 99-110, doi:10.1007/978-1-61779-328-8_7 (2011).

3 Rinn, J. L. et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

4 Lai, F. & Shiekhattar, R. Where long noncoding RNAs meet DNA methylation. Cell Res 24, 263-264, doi:10.1038/cr.2014.13 (2014).

5 Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. Methods 48, 226-232, doi:10.1016/j.ymeth.2009.05.003 (2009).

6 Meissner, A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 33, 5868-5877, doi:10.1093/nar/gki901 (2005).

7 Hahn, M. A., Li, A. X., Wu, X. & Pfeifer, G. P. Single base resolution analysis of 5-methylcytosine and 5-hydroxymethylcytosine by RRBS and TAB-RRBS. Methods Mol Biol 1238, 273-287, doi:10.1007/978-1-4939-1804-1_14 (2015).

8 Sandoval, J. et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6, 692-702 (2011).

9 Dedeurwaerder, S. et al. Evaluation of the Infinium Methylation 450K technology. Epigenomics 3, 771-784, doi:10.2217/epi.11.105 (2011).

10 Jones, P. A. & Taylor, S. M. Cellular differentiation, cytidine analogs and DNA methylation. Cell 20, 85-93 (1980).

11 Robert, M. F. et al. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nat Genet 33, 61-65, doi:10.1038/ng1068 (2003).

12 Cohen, M. B. & Glazer, R. I. Cytotoxicity and the inhibition of ribosomal RNA processing in human colon carcinoma cells. Mol Pharmacol 27, 308-313 (1985).

13 Chiuten, D. F., Muggia, F. M. & Johnson, R. K. Antitumor activity of pyrazofurin in combination with 5-azacytidine against murine P388 and L1210 leukemias and

colon carcinoma 26. Cancer Treat Rep 63, 1857-1862 (1979).

14 Jenkins, T. G. & Carrell, D. T. The paternal epigenome and embryogenesis: poising mechanisms for development. Asian J Androl, doi:aja201061 [pii]

10.1038/aja.2010.61.

15 Hammoud, S. S. et al. Distinctive chromatin in human sperm packages genes for embryo development. Nature 460, 473-478, doi:nature08162 [pii] 10.1038/nature08162 (2009).

16 Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods 11, 817-820, doi:10.1038/nmeth.3035 (2014).

17 Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res 23, 2126-2135, doi:10.1101/gr.161679.113 (2013).

18 Pembrey, M. E. et al. Sex-specific, male-line transgenerational responses in humans. Eur J Hum Genet 14, 159-166, doi:5201538 [pii] 10.1038/sj.ejhg.5201538 (2006).

19 Kaati, G., Bygren, L. O., Pembrey, M. & Sjostrom, M. Transgenerational response to nutrition, early life circumstances and longevity. Eur J Hum Genet 15, 784-790, doi:5201832 [pii] 10.1038/sj.ejhg.5201832 (2007).

20 Barkoff, A., Ballantyne, S. & Wickens, M. Meiotic maturation in Xenopus requires polyadenylation of multiple mRNAs. The EMBO journal 17, 3168-3175, doi:10.1093/emboj/17.11.3168 (1998).

21 McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B. & Richter, J. D. Poly(A) elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. Genes Dev 3, 803-815 (1989).

22 Mendez, R., Barnard, D. & Richter, J. D. Differential mRNA translation and meiotic progression require Cdc2-mediated CPEB destruction. The EMBO journal 21, 1833-1844, doi:10.1093/emboj/21.7.1833 (2002).

23 Paris, J. & Richter, J. D. Maturation-specific polyadenylation and translational control: diversity of cytoplasmic polyadenylation elements, influence of poly(A) tail size, and formation of stable polyadenylation complexes. Mol Cell Biol 10, 5634-5645 (1990).

24 Richter, J. D. Cytoplasmic polyadenylation in development and beyond. Microbiol Mol Biol Rev 63, 446-456 (1999).

25 Roy, L. M. et al. The cyclin B2 component of MPF is a substrate for the c-mos(xe) proto-oncogene product. Cell 61, 825-831 (1990).

26 Song, J. et al. The type II activin receptors are essential for egg cylinder growth, gastrulation, and rostral head development in mice. Dev Biol 213, 157-169, doi:10.1006/dbio.1999.9370 S0012-1606(99)99370-3 [pii] (1999).

27 Yamashita, M. Molecular mechanisms of meiotic maturation and arrest in fish and amphibian oocytes. Semin Cell Dev Biol 9, 569-579, doi:10.1006/scdb.1998.0251 (1998).

28 Salles, F. J., Lieberfarb, M. E., Wreden, C., Gergen, J. P. & Strickland, S. Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs. Science 266, 1996-1999 (1994).

29 Mendez, R. et al. Phosphorylation of CPE binding factor by Eg2 regulates translation of c-mos mRNA. Nature 404, 302-307, doi:10.1038/35005126 (2000).

30 Groisman, I., Huang, Y. S., Mendez, R., Cao, Q. & Richter, J. D. Translational control of embryonic cell division by CPEB and maskin. Cold Spring Harb Symp Quant Biol 66, 345-351 (2001).

31 Hodgman, R., Tay, J., Mendez, R. & Richter, J. D. CPEB phosphorylation and cytoplasmic polyadenylation are catalyzed by the kinase IAK1/Eg2 in maturing mouse oocytes. Development 128, 2815-2822 (2001).

32 Mendez, R. & Richter, J. D. Translational control by CPEB: a means to the end. Nat Rev Mol Cell Biol 2, 521-529, doi:10.1038/35080081 (2001).

33 Tang, F. et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468-478, doi:10.1016/j.stem.2010.03.015 (2010).

34 Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377-382, doi:10.1038/nmeth.1315 (2009).

35 Kaida, D. et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 468, 664-668, doi:10.1038/nature09479 (2010).

36 Berg, M. G. et al. U1 snRNP determines mRNA length and regulates isoform expression. Cell 150, 53-64, doi:10.1016/j.cell.2012.05.029 (2012).

37 Mayr, C. & Bartel, D. P. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138, 673-684 (2009).

38 Fu, Y. et al. Differential genome-wide profiling of tandem 3′ UTRs among

human breast cancer and normal cells by high-throughput sequencing. Genome research 21, 741-747 (2011).

39 Yang, J. & Weinberg, R. A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. Developmental cell 14, 818-829 (2008).

40 Clarke, M. F. & Fuller, M. Stem cells and cancer: two faces of eve. Cell 124, 1111-1115 (2006).

41 Atkins, C. M., Nozaki, N., Shigeri, Y. & Soderling, T. R. Cytoplasmic polyadenylation element binding protein-dependent protein synthesis is regulated by calcium/calmodulin-dependent protein kinase II. J Neurosci 24, 5193-5201, doi:10.1523/JNEUROSCI.0854-04.2004 (2004).

APPENDIX A

ALTERATIONS IN SPERM DNA METHYLATION PATTERNS AT IMPRINTED LOCI IN TWO CLASSES OF INFERTILITY

Reprinted with permission from Fertility and Sterility. Hammoud SS, Purwar J, Pflueger C, Cairns BR, Carrell DT (2010) Alterations in sperm DNA methylation patterns at imprinted loci in two classes of infertility. Fertility and Sterility 94: 1728-1733.

Chapter 3 is a published article. Both Saher Sue Hammoud and Jahnvi Purwar contributed to this work equally. My contribution to this work involved developing a tool to analyze the DNA methylation data.

Alterations in sperm DNA methylation patterns at imprinted loci in two classes of infertility

Saher Sue Hammoud, B.S.,[a,b] *Jahnvi Purwar, B.S.,*[d] *Christian Pflueger, M.S.,*[d] *Bradley R. Cairns, Ph.D.,*[d,e] *and Douglas T. Carrell, Ph.D.*[a,b,c]

[a]Andrology and IVF Laboratories, Division of Urology, Department of Surgery, [b]Department of Physiology, and [c]Department of Obstetrics and Gynecology, School of Medicine, and [d]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah; and [e]Howard Hughes Medical Institute, University of Utah School of Medicine, Salt Lake City, Utah

Objective: To evaluate the associations between proper protamine incorporation and DNA methylation at imprinted loci.
Design: Experimental research study.
Setting: Research laboratory.
Patient(s): Three populations were tested—abnormal protamine patients, oligozoospermic patients, and fertile donors.
Intervention(s): The CpG methylation patterns were examined at seven imprinted loci sequenced: *LIT1, MEST, SNRPN, PLAGL1, PEG3, H19,* and *IGF2.*
Main Outcome Measure(s): The DNA methylation patterns were analyzed using bisulfite sequencing. The percentage of methylation was compared between fertile and infertile patients displaying abnormal protamination.
Result(s): At six of the seven imprinted genes, the overall DNA methylation patterns at their respective differentially methylated regions were significantly altered in both infertile patient populations. When comparing the severity of methylation alterations among infertile patients, the oligozoospermic patients were significantly affected at mesoderm-specific transcript (*MEST),* whereas abnormal protamine patients were affected at *KCNQ1,* overlapping transcript 1 *(LIT1),* and at small nuclear ribonucleoprotein polypeptide N *(SNRPN).*
Conclusion(s): Patients with male factor infertility had significantly increased methylation alteration at six of seven imprinted loci tested, with differences in significance observed between oligozoospermic and abnormal protamine patients. This could suggest that risk of transmission of epigenetic alterations may be different with diagnoses. However, this study does not provide a causal link for epigenetic inheritance of imprinting diseases, but does show significant association between male factor infertility and alterations in sperm DNA methylation at imprinted loci. (Fertil Steril® 2010;94:1728–33. ©2010 by American Society for Reproductive Medicine.)

Key Words: Imprinting, Beckwith-Wiedemann syndrome and epigenetic alterations, Angelman syndrome, chromatin, assisted reproductive technology, IVF, ICSI, oligozoospermic, protamines

Genomic imprinting is established and inherited during gametogenesis and preimplantation to ensure parent-of-origin monoallelic gene expression (1, 2). The mechanism by which either one of the two alleles are differentially expressed is not completely understood; however, it is known that the majority of imprinted genes are clustered and are predominately regulated by imprinting control regions (ICRs) (3, 4). At present, approximately 80 imprinted genes have been identified, many of which are implicated in tumorigenesis, fetal growth regulation, and embryonic development (5–8). Pathological perturbation in the methylation imprints during gametogenesis or development can give rise to growth-related syndromes and is frequently observed in cancer (9–20).

After fertilization, both parental genomes are globally demethylated through active or passive demethylation mechanisms, whereas the methylation patterns at imprinted genes are maintained and only erased and re-established in the primordial germ cell. The presence of abnormal methylation patterns residing in gametes raises concerns, as these may be inherited and maintained in the embryo. Meta-analysis showed that children born from assisted reproductive technology (ART) have a fourfold increased incidence of Beckwith-Weidemann syndrome compared with children conceived naturally (21–24). In addition, imprinting syndromes such as Angelman, Prader-Willi, and Silver-Russell have been associated with ART, although no strong correlations were established. Currently, it is unclear whether imprinting abnormalities arise from the ART procedure itself or from pre-existing methylation aberrations in the gametes of infertile patients (25–27).

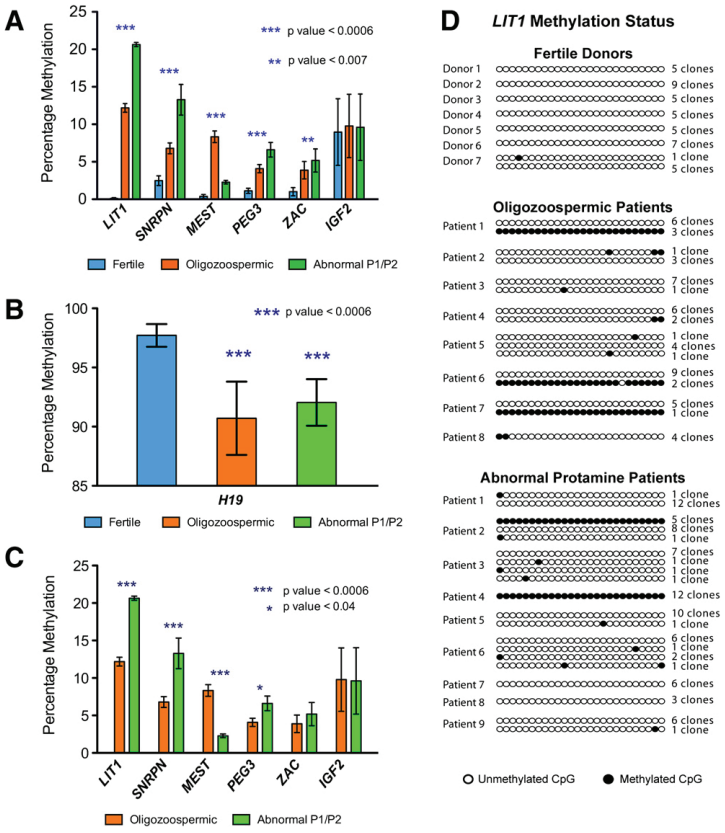
Recent studies have shown that epigenetic abnormalities are common in the sperm of severely oligozoospermic patients, favoring the latter hypothesis (26, 27). Whether epigenetic alterations at imprinted loci of infertile men are limited to oligozoospermic patients or whether epigenetic alterations extend beyond oligozoospermic patients is unknown. In this study we examine methylation changes in patients with an alternative cause for their male factor infertility—patients with abnormal sperm protamine replacement of histones. Protamines 1 and 2 are sperm-specific nuclear proteins that are incorporated into the DNA in a 1:1 ratio and ensure chromatin

Received April 8, 2009; revised August 5, 2009; accepted September 8, 2009; published online November 1, 2009.
S.S.H. has nothing to disclose. J.P. has nothing to disclose. C.P. has nothing to disclose. B.R.C. has nothing to disclose. D.T.C. has nothing to disclose.
Saher Sue Hammoud and Jahnvi Purwar contributed equally.
Reprint requests: Douglas T. Carrell, Ph.D., University of Utah IVF and Andrology Laboratories, 675 Arapeen Drive, Suite 205, Salt Lake City, UT 84108 (FAX: 801-581-6127; E-mail: douglas.carrell@hsc.utah.edu).

0015-0282/$36.00
doi:10.1016/j.fertnstert.2009.09.010

FIGURE 1

The overall methylation patterns at both paternally and maternally imprinted genes were altered in the sperm of infertile patients. (**A,B,C**) The mean percentage of methylation with standard error. $P < .05$ is significant. (**A**) The percentage of methylated CpGs at normally paternally demethylated loci. (**B**) The percentage of demethylation at a paternally methylated DMR of *H19*. (**C**) Comparing methylation changes between the two infertile patient populations. (**D**) Methylation status at the differentially methylated region of *LIT1* for fertile donors, oligozoospermic patients, and abnormal protamine patients.



Hammoud. Imprinting abnormalities in infertile men. Fertil Steril 2010.

condensation. The average P1:P2 ratio in fertile men is ~1, whereas in some infertile patients this ratio is significantly altered (28, 29) and consequently associated with severe sperm defects that can usually be addressed through ART (30, 31). It has been proposed that chromatin packaging may have a role in properly establishing and maintaining methylation patterns, hence, hypothetically, patients with abnormal protamine ratios may be at an increased risk of conceiving an ART offspring with imprinting disease (32, 33). This study evaluates the relationship between protamine ratios and methylation patterns at seven imprinted loci in the sperm of abnormal protamine patients or oligozoospermic patients. We reveal significant changes in the overall DNA methylation patterns at six of these loci, with varying impact on methylation patterns within each class of infertility: oligozoospermic or abnormal protamine levels (p-value < 0.05, Figure 1). These data suggest that aberrant imprinting patterns are observed in patients with abnormal protamine ratios, and that the abnormal patterns may vary among different pathologies, providing a spectrum of risks for transmitting epigenetic abnormalities to the embryo.

MATERIALS AND METHODS

Patient Population

Of the seven tested imprinted loci, six are paternally demethylated and expressed: KCNQ1 overlapping transcript 1 (*LIT1*), insulin-like growth factor 2 (*IGF2*), paternally expressed gene 3 (*PEG3*), pleiomorphic adenoma gene-like 1 (*PLAGL1* also known as *ZAC*), small nuclear ribonucleoprotein

polypeptide N *(SNRPN),* and mesoderm-specific trancript (*MEST*), and one is maternally expressed and is normally DNA methylated in sperm (*H19*). For each locus 10 oligozoospermic (sperm count $\leq 10 \times 10^6$/mL), 10 abnormal protamine replacement patients (average sperm count of $73 \times 10^6 \pm 60$ SD/mL), and 5 known fertile donors were evaluated. For *LIT1* only, eight oligozoospermic patients and nine abnormal protamine patients were evaluated.

Sample Collection and Bisulfite Treatment

Institutional Review Board (IRB) approval was obtained before initiation of this study. Frozen sperm DNA samples were treated with sodium bisulfite to convert unmethylated cytosines to uracil and leaving methylated cytosines unchanged, as previously described by Clark et al. (34). DNA was purified using Qiagen DNeasy clean up kit (Qiagen, Valencia, CA) and eluted twice, each time with 100 μL of elution buffer. The purified DNA was desulfonated by the addition of 20 μL NaOH and incubated at 37°C for 15 minutes. After incubation, 22 μL of 4 M NaOAC, glycogen, and two volumes of ethanol were added to precipitate the DNA overnight at -20°C. Precipitated DNA was washed twice with 70% ethanol and eluted in 30 μL of elution buffer.

PCR Amplification of Bisulfite Converted DNA

Primer sequences and temperatures for *SNRPN, PEG3, ZAC, MEST, LIT1, H19 ICR*, and *IGF2* are available upon request (35, 36). The polymerase chain reaction (PCR) reactions were performed in 50-μL volume reactions containing 5 μL of 10 × PCR buffer–$MgCl_2$ (Invitrogen, Carlsbad, CA), 5 μL of 10 × Enhancer Buffer (Invitrogen), 1.5 μL of $MgCl_2$, 1 μL of 10 mM dNTPs, 0.5 μL of *Taq* (Invitrogen), 2.5 μL of each forward and reverse primer (10 μM stock), and 30 μL of water. The PCR results were analyzed on a 1% agarose gel, and gel purified if multiple products were detected.

TOPO TA Cloning and Sequencing

The PCR products were cloned into a TOPO 2.1 pCR vectors (Invitrogen) and plated onto KAN-X-GAL plates for blue-white screening. Positive colonies were reinoculated into LB-KAN (50 μg/mL), cultured overnight, and plasmids were purified using the Qiagen 96-well clean-up kit. To address sperm sample heterogeneity five or more clones/alleles were sequenced per patient for each of the imprinted loci (sequencing done at Genewiz San Diego Laboratory).

Data Visualization and Analysis

The CG/TG-analyzer, a Perl program, was used to examine the methylation status of a bisulfite-converted sequence and provides an output in the form of 1s and 0s, where 1s represent methylated cytosines and 0s represent unmethylated cytosines (thymine). The CpG positions were defined in a multifasta file, text-based file containing multiple DNA or protein sequences, which includes the CpG position number flanked by four nucleotides on each side. The output was used to calculate the percentage of CpG methylation (program is be available upon request). To compare the overall methylation profile in infertile patients versus fertile donors (Fig. 1), the Wilcoxon-Mann-Whitney test was used. This test is a nonparametric significance test for assessing whether two independent samples of observations came from the same distribution. To determine significance between fertile donors and oligozoospermic patients or fertile and abnormal protamine patients the percentage of methylated CpGs represented in columns 2 and 3 (in Tables 1, 2, and 3) were compared as independent sample populations. A P value $< .05$ was considered significant. The χ^2 analysis was used to compare the percentage of methylated CpGs in the abnormal protamine or oligozoospermic patients with known fertile donors.

RESULTS

Six imprinted genes, that are normally paternally demethylated, were examined: *LIT1, SNRPN, MEST, ZAC, PEG3,* and *IGF2*. Here, all except *IGF2,* showed significant hypermethylation in oligozoospermic and abnormal protamine patients compared with fertile donors (Fig. 1A). Furthermore, the differentially methylated region (DMR) of *H19* (a paternally methylated locus) was

TABLE 1

The percentage of methylated CpGs in the DMR of *LIT1* of oligozoospermic and abnormal protamine patients.

CpG	Abnormal P1/P2 (n = 9)	Oligozoospermic (n = 8)	Fertile donors (n = 7)	Fertile vs. abnormal	Fertile vs. oligozoospermic
CpG 1	25.882	18.181	0	0.0003	0.0035
CpG 2	20	18.181	0	0.0021	0.0035
CpG 3	20	10.909	0	0.0021	0.0271
CpG 4	20	10.909	2.38	0.0066	0.17
CpG 5	21.176	10.909	0	0.0015	0.0271
CpG 6	20	10.909	0	0.0021	0.0271
CpG 7	21.176	10.909	0	0.0015	0.0271
CpG 8	20	10.909	0	0.0021	0.0271
CpG 9	20	10.909	0	0.0021	0.0271
CpG 10	20	10.909	0	0.0021	0.0271
CpG 11	21.176	12.277	0	0.0015	0.0186
CpG 12	20	10.909	0	0.0021	0.0271
CpG 13	21.176	10.909	0	0.0015	0.0271
CpG 14	20	14.454	0	0.0021	0.0101
CpG 15	20	10.909	0	0.0021	0.0271
CpG 16	20	7.272	0	0.0021	0.0742
CpG 17	20	10.909	0	0.0021	0.0271
CpG 18	21.176	12.272	0	0.0015	0.0093
CpG 19	20	10.909	0	0.0021	0.0271
CpG 20	20	10.909	0	0.0021	0.0271
CpG 21	21.176	16.363	0	0.0015	0.0059
CpG 22	21.176	16.363	0	0.0015	0.0059

Note: DMR = differentially methylated region.

Hammoud. Imprinting abnormalities in infertile men. Fertil Steril 2010.

TABLE 2

The percentage of methylated CpG in the DMR of *SNRPN*.

CpGs	Abnormal P1/P2 (n = 11)	Oligozoospermic (n = 13)	Fertile donors (n = 5)	Fertile vs. abnormal	Fertile vs. oligozoospermic
CpG 1	4.3	4.0	0	0.152	0.169
CpG 2	5.8	5.0	0	0.09	0.123
CpG 3	5.8	5.0	0	0.09	0.123
CpG 4	5.8	5.0	0	0.09	0.123
CpG 5	5.8	5.0	0	0.09	0.123
CpG 6	10.6	5.0	0	0.026	0.123
CpG 7	14.8	8.0	4.3	0.08	0.413
CpG 8	8.7	5.0	0	0.04	0.123
CpG 9	13.0	5.0	4.3	0.10	0.864
CpG 10	23.1	16	6.5	0.05	0.114
CpG 11	10.1	6.0	0	0.026	0.09
CpG 12	11.6	6.1	8.7	0.618	0.526
CpG 13	15.9	8.0	6.5	0.1	0.753
CpG 14	47.8	10	2.2	0.0001	0.09
CpG 15	11.6	6.1	0	0.017	0.08
CpG 16	5.8	4.0	2.2	0.351	0.566
CpG 17	11.6	13	2.2	0.065	0.039
CpG 18	15.9	12.2	6.5	0.130	0.295
CpG 19	15.9	5.1	6.7	0.140	0.705
CpG 20	17.4	5.1	0	0.003	0.119
CpG 21	17.6	4.1	2.2	0.011	0.560

Note: DMR = differentially methylated region.

Hammoud. Imprinting abnormalities in infertile men. Fertil Steril 2010.

significantly hypomethylated in both infertile classes ($P<.006$ for all except *ZAC*, $P<.002$) (Fig. 1B). Thus, these infertile patients show methylation alterations at six of seven loci tested. However, when comparing overall methylation changes between the two infertile populations, abnormal protamine patients show more extensive hypermethylation at the DMRs of *LIT1* and *SNRPN* in comparison with oligozoospermic patients. In contrast, hypermethylation at *MEST* is significantly higher in oligozoospermic patients (p-value < 0.006, Fig. 1C).

Notably, in both patient populations, the locus that displays the highest number of affected CpGs is *LIT1*. In the DMR of *LIT1*, the percentage of methylated CpGs ranged from 7%–18% or 20%–25% for oligozoospermic or abnormal protamine patients, respectively (Table 1). In contrast, for fertile donors, virtually all CpGs were demethylated. The percentages of methylated CpGs in oligozoospermic and abnormal protamine patients were statistically significant when compared with fertile donors (p-value < 0.05, Table 1). To address the uniformity of methylation changes at *LIT1* in individual sperm from a single patient, we sequenced multiple alleles (5–12) from each patient, and found striking heterogeneity. In three of the eight oligozoospermic patients, *LIT1* was completely methylated in 20%–30% of the alleles, whereas in the other five patients, only sporadic increases were observed (Fig. 1D). Similarly, in the abnormal protamine category one patient always displayed complete methylation, a second displayed methylation on 50% of his alleles, and the remainder (seven) displayed little or no increase.

Consistent with the findings reported previously, the DMR of *SNRPN* was also susceptible to acquiring methylation in infertile men. Abnormal protamine patients had a significant increase in CpG methylation (methylation at individual CpGs typically ranged from 4%–20%) (p-value < 0.05 Table 2). Alterations were also observed in oligozoospermic patients (range of methylation, 4%–8%), but the increase lacked statistical significance (Table 2). At *SNRPN*, alterations in methylation were common (observed at a majority of the alleles) but typically involved only a moderate number of CpGs acquiring methylation. However, in both patient categories, a small number of patients displayed complete methylation at 10% of the alleles tested.

Methylation levels in the DMR of *MEST* (for each CpG) ranged from 7%–19% or 1%–3% in oligozoospermic or abnormal protamine patients, respectively (Table 3). The changes in methylation at many of the CpGs in oligozoospermic patients were near the range of statistical significance ($P=.07$; Table 3). In addition, 3 of 10 oligozoospermic patients had 12%–33% of their alleles completely methylated, whereas the remaining 7 patients displayed very little change. Likewise, in the abnormal protamine class, one patient had 14% of his alleles completely methylated and in the remaining nine patients, there was virtually no change observed. In contrast, very few individual CpGs were significantly ($P<.05$) affected in *PEG3, ZAC*, *IGF2* promoter 3, and *H19* in infertile patients (data not shown).

DISCUSSION

In this study we evaluated the methylation status of seven imprinted loci in two patient populations: oligozoospermic and abnormal protamine ratio patients. The overall methylation patterns in sperm of infertile patients were significantly altered at all imprinted loci (except *IGF2*) when compared with fertile donors. However, when comparing the two infertile patient populations, oligozoospermic patients were hypermethylated at *MEST*, an imprinted gene associated with Silver-Russell syndrome, whereas abnormal protamine patients had significant changes at *LIT1* and *SNRPN* (Figure 1), genes that may be associated with cases of transient neonatal diabetes mellitus and Angelman syndrome. These data suggest that risk of transmission of epigenetic alterations may be different with diagnoses.

TABLE 3

The percentage of methylated CpGs at the DMR of *MEST* in oligozoospermic and abnormal protamine patients.

CpG	Abnormal P1/P2 (n = 10)	Oligozoospermic (n = 10)	Fertile donors (n = 5)	Fertile vs. abnormal	Fertile vs. oligozoospermic
CpG 1	1.785	14.28	0	0.2346	0.0167
CpG 2	1.785	19.04	0	0.2346	0.0063
CpG 3	3.571	7.1428	0	0.1515	0.070
CpG 4	3.571	7.142	3.4	0.483	0.250
CpG 5	1.785	7.1428	0	0.2346	0.070
CpG 6	1.785	9.5238	0	0.2346	0.436
CpG 7	3.571	7.1428	0	0.1515	0.070
CpG 8	1.785	7.1428	0	0.2346	0.070
CpG 9	1.785	7.1428	0	0.2346	0.070
CpG 10	1.785	7.1428	0	0.2346	0.070
CpG 11	1.785	7.1428	0	0.2346	0.070
CpG 12	3.571	7.1428	0	0.1515	0.070
CpG 13	1.785	9.523	3.4	0.642	0.1604
CpG 14	3.571	4.7619	0	0.1515	0.1167
CpG 15	1.785	7.1428	0	0.2346	0.070
CpG 16	3.571	7.1428	0	0.1515	0.070
CpG 17	1.785	7.1428	0	0.2346	0.070
CpG 18	0	7.1428	0	NA	0.070

Note: DMR = differentially methylated region.

Hammoud. Imprinting abnormalities in infertile men. Fertil Steril 2010.

Our data evaluate and demonstrate abnormal imprinting in a different class of abnormal spermatogenesis, abnormal replacement of nuclear proteins by protamine 1 and protamine 2. It was our hypothesis that abnormal chromatin packaging may be associated with methylation defects, which is supported by the data presented from this study. These data, along with previously published data from oligozoospermic patients, reveal that alteration in DNA methylation patterns are common at a handful of imprinted loci tested, suggesting that imprinting abnormalities may reside in the sperm of infertile patients (25–27), but whether these alterations can be inherited is uncertain. Remarkably, when examining normally demethylated DMRs, the alleles of infertile patients are often either unaffected or entirely methylated, suggesting a bistable status, and a susceptibility to complete methylation. Clearly, complete methylation of a normally unmethylated locus may lead to an imprinting disorder in the embryo if proper imprint reestablishment mechanisms are not implemented. Also of note are the small differences in the degree of methylation within some genes and alleles. It is important to determine whether this abnormal methylation has reached a threshold level that might lead to complete methylation in the embryo (at a certain unknown probability) and confer disease, or whether there is a gradual continuum with a threshold for disease.

Whether imprinting diseases in ART offspring arise as a result of abnormal methylation of gametes, or acquire methylation changes during in vitro culture, or both, is still unknown. Current human data suggest that methylation alteration at imprinted loci may reside in gametes and may be inherited by the embryo. Supporting evidence comes from two reports showing that a gain in methylation on the paternal alleles of *LIT1* or *MEST* in sperm is maintained in the baby and associated with transient neonatal diabetes (37) or Silver-Russell syndrome (38). The findings suggest that paternal imprints in sperm may be needed for a healthy and uncomplicated pregnancy. The need to study sperm from fathers of children with imprinting diseases is imperative.

This study does not report a causal link between abnormal methylation of imprinted genes and disease. The relative risk of the defects reported in our study to patients is unknown. However, we demonstrate a link between abnormal spermatogenesis and abnormal methylation of genes associated with rare imprinting diseases previously reported to have elevated incidences in ART offspring (21–24). This suggests that such a link may be strengthened in infertile men with known abnormalities in chromatin packaging. Characterizing these epigenetic alterations in the sperm of infertile men may help predict the likelihood of IVF success rate.

REFERENCES

1. Morison IM, Paton CJ, Cleverley SD. The imprinted gene and parent-of-origin effect database. Nucleic Acids Res 2001;29:275–6.
2. Ferguson-Smith AC, Surani MA. Imprinting and the epigenetic asymmetry between parental genomes. Science 2001;293:1086–9.
3. Verona RI, Mann MR, Bartolomei MS. Genomic imprinting: intricacies of epigenetic regulation in clusters. Annu Rev Cell Dev Biol 2003;19:237–59.
4. Spahn L, Barlow DP. An ICE pattern crystallizes. Nat Genet 2003;35:11–2.
5. Morgan HD, Santos F, Green K, Dean W, Reik W. Epigenetic reprogramming in mammals. Hum Mol Genet 2005. 14 Spec No 1: R47–R58.
6. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 2003;33. Suppl: 245–54.
7. Royo H, Bortolin ML, Seitz H, Cavaille J. Small noncoding RNAs and genomic imprinting. Cytogenet Genome Res 2006;113:99–108.
8. Seitz H, Royo H, Bortolin ML, Lin SP, Ferguson-Smith AC, Cavaille J. A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. Genome Res 2004;14:1741–8.
9. Bjornsson HT, Brown LJ, Fallin MD, Rongione MA, Bibikova M, Wickham E, et al. Epigenetic specificity of loss of imprinting of the IGF2 gene in Wilms tumors. J Natl Cancer Inst 2007;99:1270–3.
10. Chao W, D'Amore PA. IGF2: epigenetic regulation and role in development and disease. Cytokine Growth Factor Rev 2008;19:111–20.

11. Cui H. Loss of imprinting of IGF2 as an epigenetic marker for the risk of human cancer. Dis Markers 2007;23:105–12.
12. Hemberger M. Epigenetic landscape required for placental development. Cell Mol Life Sci 2007;64: 2422–36.
13. Henckel A, Feil R. Differential epigenetic marking on imprinted genes and consequences in human diseases. Med Sci (Paris) 2008;24:747–52.
14. Lemeta S, Jarmalaite S, Pylkkanen L, Bohling T, Husgafvel-Pursiainen K. Preferential loss of the nonimprinted allele for the ZAC1 tumor suppressor gene in human capillary hemangioblastoma. J Neuropathol Exp Neurol 2007;66:860–7.
15. Murrell A, Ito Y, Verde G, Huddleston J, Woodfine K, Silengo MC, et al. Distinct methylation changes at the IGF2-H19 locus in congenital growth disorders and cancer. PLoS ONE 2008;3:e1849.
16. Nakano S, Murakami K, Meguro M, Soejima H, Higashimoto K, Urano T, et al. Expression profile of LIT1/KCNQ1OT1 and epigenetic status at the KvDMR1 in colorectal cancers. Cancer Sci 2006;97:1147–54.
17. Nowaczyk MJ, Carter MT, Xu J, Huggins M, Raca G, Das S, et al. Paternal deletion 6q24.3: a new congenital anomaly syndrome associated with intrauterine growth failure, early developmental delay and characteristic facial appearance. Am J Med Genet A 2008;146:354–60.
18. Shin JY, Fitzpatrick GV, Higgins MJ. Two distinct mechanisms of silencing by the KvDMR1 imprinting control region. EMBO J 2008;27:168–78.
19. Valleley EM, Cordery SF, Bonthron DT. Tissue-specific imprinting of the ZAC/PLAGL1 tumour suppressor gene results from variable utilization of monoallelic and biallelic promoters. Hum Mol Genet 2007;16:972–81.
20. Yoshimizu T, Miroglio A, Ripoche MA, Gabory A, Vernucci M, Riccio A, et al. The H19 locus acts in vivo as a tumor suppressor. Proc Natl Acad Sci U S A 2008;105:12417–22.
21. Cox GF, Burger J, Lip V, Mau UA, Sperling K, Wu BL, et al. Intracytoplasmic sperm injection may increase the risk of imprinting defects. Am J Hum Genet 2002;71:162–4.
22. Niemitz EL, DeBaun MR, Fallon J, Murakami K, Kugoh H, Oshimura M, et al. Microdeletion of LIT1 in familial Beckwith-Wiedemann syndrome. Am J Hum Genet 2004;75:844–9.
23. DeBaun MR, Niemitz EL, Feinberg AP. Association of in vitro fertilization with Beckwith-Wiedemann syndrome and epigenetic alterations of LIT1 and H19. Am J Hum Genet 2003;72:156–60.
24. Gosden R, Trasler J, Lucifero D, Faddy M. Rare congenital disorders, imprinted genes, and assisted reproductive technology. Lancet 2003;361: 1975–7.
25. Kobayashi H, Sato A, Otsu E, Hiura H, Tomatsu C, Utsunomiya T, et al. Aberrant DNA methylation of imprinted loci in sperm from oligospermic patients. Hum Mol Genet 2007;16:2542–51.
26. Marques CJ, Carvalho F, Sousa M, Barros A. Genomic imprinting in disruptive spermatogenesis. Lancet 2004;363:1700–2.
27. Marques CJ, Costa P, Vaz B, Carvalho F, Fernandes S, Barros A, et al. Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia. Mol Hum Reprod 2008;14:67–74.
28. Aoki VW, Emery BR, Liu L, Carrell DT. Protamine levels vary between individual sperm cells of infertile human males and correlate with viability and DNA integrity. J Androl 2006;27:890–8.
29. Balhorn R, Reed S, Tanphaichitr N. Aberrant protamine 1/protamine 2 ratios in sperm of infertile human males. Experientia 1988;44:52–5.
30. Aoki VW, Moskovtsev SI, Willis J, Liu L, Mullen JB, Carrell DT. DNA integrity is compromised in protamine-deficient human sperm. J Androl 2005;26: 741–8.
31. Aoki VW, Liu L, Jones KP, Hatasaka HH, Gibson M, Peterson CM, et al. Sperm protamine 1/protamine 2 ratios are related to in vitro fertilization pregnancy rates and predictive of fertilization ability. Fertil Steril 2006;86:1408–15.
32. Paldi A. Genomic imprinting: could the chromatin structure be the driving force? Curr Top Dev Biol 2003;53:115–38.
33. Rousseaux S, Caron C, Govin J, Lestrat C, Faure AK, Khochbin S. Establishment of male-specific epigenetic information. Gene 2005;345:139–53.
34. Clark SJ, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. Nucleic Acids Res 1994;22:2990–7.
35. El-Maarri O, Seoud M, Coullin P, Herbiniaux U, Oldenburg J, Rouleau G, et al. Maternal alleles acquiring paternal methylation patterns in biparental complete hydatidiform moles. Hum Mol Genet 2003;12:1405–13.
36. Kamikihara T, Arima T, Kato K, Matsuda T, Kato H, Douchi T, et al. Epigenetic silencing of the imprinted gene ZAC by DNA methylation is an early event in the progression of human ovarian cancer. Int J Cancer 2005;115:690–700.
37. Arima T, Kamikihara T, Hayashida T, Kato K, Inoue T, Shirayoshi Y, et al. ZAC, LIT1 (KCNQ1OT1) and p57KIP2 (CDKN1C) are in an imprinted gene network that may play a role in Beckwith-Wiedemann syndrome. Nucleic Acids Res 2005;33:2650–60.
38. Kagami M, Nagai T, Fukami M, Yamazawa K, Ogata T. Silver-Russell syndrome in a girl born after in vitro fertilization: partial hypermethylation at the differentially methylated region of PEG1/MEST. J Assist Reprod Genet 2007;24: 131–6.

APPENDIX B

TO ELIMINATE SOMATIC CELL CONTAMINATION FROM HUMAN SPERM PREPARATIONS

B.1 Experimental Design

Goal: To test if somatic cell contamination in sperm preparation can change DNA methylation levels at a selected imprinted locus and to establish a protocol to eliminate any such somatic cell contamination.

1) Establish a stringent somatic cell lysis protocol that yields high sperm recovery but lacks somatic cell contamination:

2) To test for effectiveness of this somatic cell lysis (SCL) protocol and compare it to a previously published SCL protocol from the Carrell lab:

Test for DNAme levels at the imprinted, DNA hypomethylated DLK1 promoter:

a) Human sperm

b) Human white blood cells

c) Known contamination levels of white blood cells to human sperm

a. Test performance of Carrell lab SCL protocol

b. Test performance of Christian Pflueger's SCL protocol

B.1.1 Sperm Incubation Buffer

MEM media (Minimum Essential Medium Eagle media) is used at room temperature and 0.4% w/w BSA (Bovine Serum Albumin) is added. To ensure proper buffering, 10x PBS (Phosphate Buffered Saline, pH 7.4) is added to 1x concentration. For each sample, approximately 5 ml buffer is required. For ten samples, 45 mL of MEM are mixed with 2 mg BSA and 5 ml 10x PBS are added afterwards. After complete resuspension of the BSA, the solution is ready to use.

B.1.2 Somatic Cell Lysis Buffer (2x concentrated)

For each sample, 4 ml 2x concentrated SCL is required. To obtain 2x SCL,

resuspend 0.2 % w/v SDS (Sodium Dodecyl Sulfate) and 1 % Triton X-100 in RNAse and DNAse free water.

B.1.3 Cell Strainer

Cell strainers for 50 ml conical tubes are required to filter out bigger chunks of somatic cell contamination. 40 µm Nylon cell strainers from Fisher Scientific, #08-771-1, were used.

B.1.4 Stainless Steel Beads for Qiagen Tissue Lyzer

Qiagen Stainless Steel Beads, 5 mm (200), # 69989, were used to lyze the sperm cells.

B.1.5 Protocol

Male mice were sacrificed according to IACUC regulations and the epididymis with the attached vas deferens was extracted and transferred to a small petri dish (20 mm by 5 mm) and a volume of 1 ml sperm incubation buffer is added to the sample. Further, the sperm is then squeezed out of the vas deferens with a small forceps and the epididymis is poked with the same forceps to create lesions for the sperm to swim out. The sample is then incubated for 1h at 37ºC. After the incubation, the sperm is pipetted up and down to finalize separation of the sperm cells from each other. The sperm cells are then pipetted through the nylon 40 µm cell strainer into a 50 ml conical tube, followed by a 10 mL 1x PBS rinse. Then, the cell strainers are removed and the tubes capped, followed by a 10 min 2,000xg spin to pellet the sperm cells. The supernatant is subsequently removed, the sperm cells are resuspended in 40 ml 1x PBS and the samples are spun again for 5 min at 4780xg. This procedure is repeated once more. Then, the sperm samples are

resuspended in 40 ml water (RNAse and DNAse free) and followed by a 10 min spin at 4780xg. The subsequent removal of the water needs be done very carefully since the pellet will be very loose at that point. Ideally, about 0.8 to 1 ml of water was left to resuspend the pellet and maximize sperm recovery. The sperm samples are then transferred to a 15 ml conical tube and water is added to 4 ml. At this point, the 2x SCL buffer is added with 4 ml per sample for a total volume of 8 ml. The samples are then inverted roughly 5 times and incubated on ice for at least 1h to complete somatic cell lysis process. Then, the samples are spun down for 10 min spin at 4780xg. The supernatant is removed but 100 to 200 µl of buffer is left back in order to resuspend the samples. This is to ensure maximum recovery of the sperm samples and to maximize yield of DNA and RNA subsequently. Samples are then transferred to a 2 ml safe lock tube and samples can be frozen and stored at -80ºC at this point.

B.2 Results and Discussion

Determining accurate DNA methylation levels for sperm samples is critical in understanding epigenetic alterations based on environmental influences or disease relevance. Notably, small amounts of DNA from somatic cells can contaminate the DNAme analysis and lead to wrong interpretations. This issue was the motivation to develop a robust and stringent protocol to remove cells and thus DNA from somatic cells while saving the majority of sperm DNA for analysis. In order to measure DNAme alterations at a known DNA hypomethylated locus, we chose to investigate the DLK1 promoter by bisulfite pyro sequencing. The DLK1 promoter is known to be hypomethylated in sperm and hypermethylated in somatic cells. The experimental setup was to test two different SCL protocols and compare them with each other under conditions of known white cell contamination. As seen in Figure B.1, clean sperm only has roughly 10% DNA methylation at the DLK1 promoter

as opposed to white blood cells (WBC), which almost reach 80% DNA methylation levels. Each sample has four replicates, obtained from four independent human donors.

The first question was to address if DNA sticking to the outside of sperm cells or incomplete lysis of WBCs is contributing to an altered DNAme profile. As seen in Figure B.1, adding either 10% WBCs or 50% WBCs to sperm cells increases DNAme levels. However, adding just DNA from WBCs to sperm cells does not change the hypomethylated property of the DLK1 promoter. Importantly, these samples have not undergone any SCL procedure. Hence, we concluded that intact somatic cells could be the root of changed DNA methylation levels if not removed thoroughly.

The second question was to address and compare the established SLC procedure from the Carrell Lab (shown in red bars, figure 1) to the SCL method detailed above (show in green bars, Figure B.1). Remarkably, the Carrell Lab's SCL failed to remove somatic cell contaminations and showed almost a 3-4 fold increase in DNA methylation at 50% WBC contamination. This is in stark contrast to the SCL method we have developed where neither of the WBC contaminations altered the DNAme levels significantly.

Going forward, this more stringent SCL protocol will be implemented as the current standard to prepare sperm samples for analysis from either mouse or human. It is worth mentioning that mouse sperm samples can only be efficiently obtained by sacrificing the male mouse. It involves a surgical procedure in order to extract the sperm samples from the mouse. This method inherently has many more somatic cells contaminating the sperm sample due to that surgery as opposed to sperm donations from humans. However, patients or donors with low sperm counts may also have a skewed ratio of sperm to somatic cells. Consequently, the implementation of this SCL protocol for any kind of sperm samples seems a very

important step in ensuring no somatic cells will contaminate and thus alter the DNA methylation analysis of select loci.

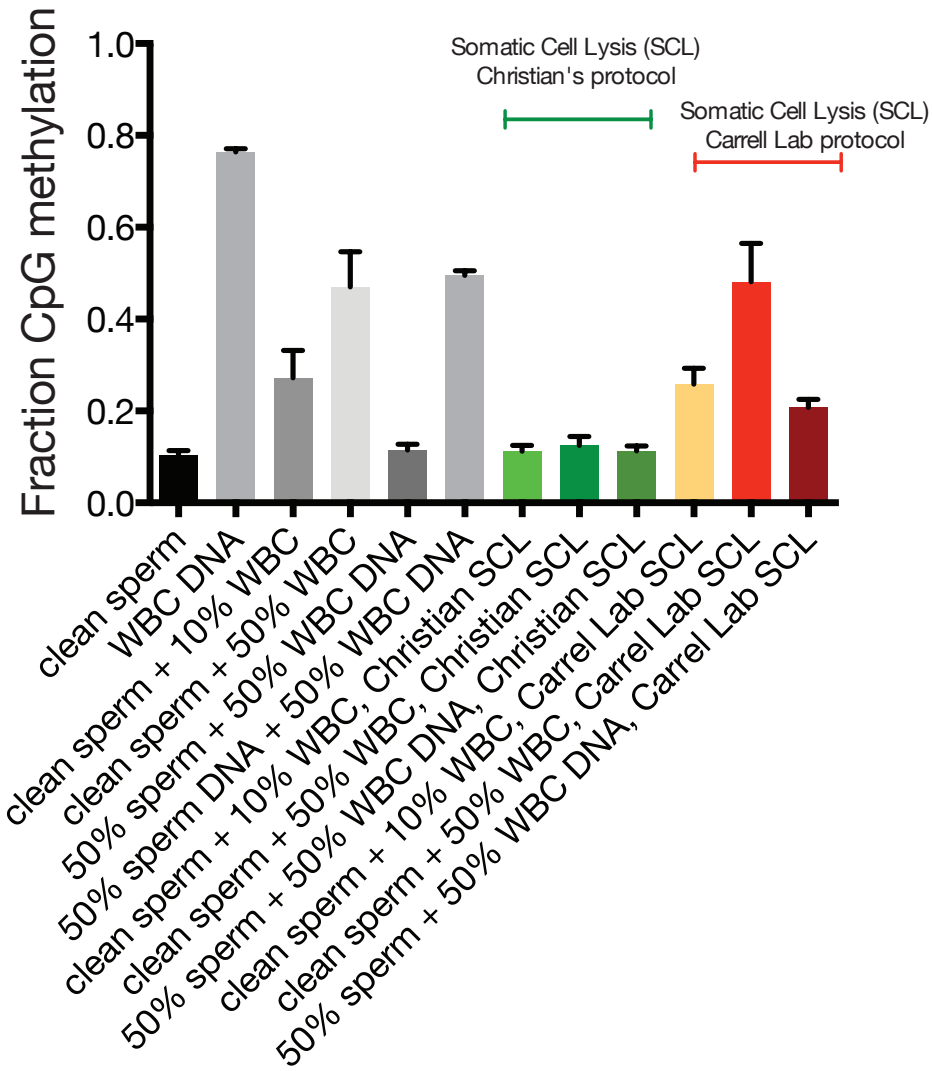


Figure B.1 Comparison of somatic cell contamination levels using different purification protocols in human sperm cells at DLK1 promoter.