DIGITALCOMMONS
—@WAYNESTATE—

Wayne State University

Wayne State University Dissertations

January 2019

Defining The Effect Of Environmental Perturbation On The Male Germline

Molly Estill
Wayne State University, mollyestill7@gmail.com

Follow this and additional works at: https://digitalcommons.wayne.edu/oa_dissertations

Part of the Molecular Biology Commons

DEFINING THE EFFECT OF ENVIRONMENTAL PERTURBATION ON THE MALE GERMLINE

by

MOLLY S. ESTILL

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: MOLECULAR BIOLOGY AND GENETICS

Approved By:

_____
Advisor   Date

_____

_____

_____

ACKNOWLEGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ART	Assisted Reproductive Technologies
CTCF	CCCTC-binding factor
DBP	Di-Butyl Phthalate
EGA	Embryonic Genome Activation
ERV	Endogenous Retrovirus
ESC	Embryonic Stem Cell
ET	Fresh Embryo Transfer
FET	Frozen Embryo Transfer
FFPE	Formalin-Fixed Paraffin-Embedded
FPKM	Fragments Per Kilobase per Million
GIFT	Game Intrafallopian Tube Transfer
GTEx	Genotype-Tissue Expression
ICSI	Intra-Cytoplasmic Sperm Injection
IUI	Intra-Uterine Insemination
IVF	*In Vitro* Fertilization
LM	Linear Model
LMEM	Linear Mixed Effects Model
LTR	Long Terminal Repeat
miRNA	microRNA
NGS	Next Generation Sequencing
piRNA	piwi-interacting RNA
Poly(A+)	Poly-A plus
PPV	Positive Predictive Value
RE	RNA element
RPM	Reads Per Million

RPKM	Reads Per Kilobase per Million
RT	Reverse Transcriptase
TAD	Topologically Associating Domain
TFBS	Transcription Factor Binding Site
TIC	Timed Intercourse
tRNA	Transfer RNA
siRNA	Small Interfering RNA
SSC	Spermatogenic Stem Cells

CHAPTER 1

"EPIGENETIC CONSEQUENCES OF PRE-CONCEPTIONAL EXPOSURES IN HUMAN"

This chapter was adapted in part from the following publication:

i. Summary

The Developmental Origins of Health and Disease (DOHaD) hypothesis proposes that periconceptional environment influences offspring phenotype, partially through epigenetic mechanisms [1-3]. Both maternal and paternal peri-conceptional environments are now believed to contribute to offspring phenotype. Assisted reproductive technologies, such as *in vitro* fertilization, alter the pre- and post-fertilization environment of the human embryo. Manipulation of the early embryo for treating human infertility is suspected of contributing to offspring abnormalities through epigenetic mechanisms. While the causal paternal components of DOHaD are insufficiently understood, particularly in human, the pre-conceptional environment of the father may influence several components of the male germline. In addition to the haploid paternal DNA, spermatozoa carry RNA, proteins and chromatin marks to the embryo. Murine and limited human studies suggest that epigenetic marks in the mature spermatozoa may be modified by the paternal environment, including common endocrine disruptor exposures. Frequent human exposures to endocrine disruptors, exogenous chemicals that can mimic or alter hormonal responses, make it imperative to fully define the affected epigenetic marks in human sperm.

ii. Background

The Developmental Origins of Health and Disease (DOHaD) theory proposes that preconceptional, prenatal and childhood exposures affect health outcomes later in life.

Maternal health is a critical component of DOHaD, as maternal stressors, diet, and chemical exposures can directly affect both the oocyte and the growing fetus across gestation. Maternal psychological stress during gestation, often measured by self-reporting, is often correlated with autism symptom severity [4] and toddler cognition and temperament [5, 6]. Maternal diet is well documented to be associated with offspring health [7, 8]. Certain chemical exposures during human gestation are well known to be detrimental to fetal development. For example, gestational thalidomide use, prescribed as an antiemetic to counter morning sickness, leads to limb and organ deformation [9]. However, more subtle effects, in the absence of a gross fetal morphological phenotype, may occur with gestational exposure to exogenous chemicals. For example, diethylstilbestrol (DES), a synthetic estrogen, was once prescribed under the assumption that DES would prevent adverse pregnancy outcomes, such as pre-term birth and miscarriages. However, gestational exposure to DES has since been associated with reproductive system cancers and reproductive tract problems in the exposed offspring [10].

With the majority of DOHaD studies focus on the maternal contribution to fetal health, the paternal aspect of DOHaD remains largely unexplored. However, there is evidence for the non-genetic transfer of paternal environmental information to the embryo and subsequent fetus (**Figure 1.1**). Murine experiments on paternal exposure to chemicals, extreme diets, exercise or adverse psychological events suggest that the paternal experience is passed along to the offspring, presenting an offspring phenotype. For example, male mice fed a Western-like diet produce offspring with metabolic dysfunction, an effect which is suggested to be mediated by spermatozoal RNAs [11]. The intergenerational effects of paternal diet also extend to simpler organisms, with a drosophila model showing that acute paternal dietary sugar reprograms offspring metabolism [12]. Stressful events in both neonatal and adult male mouse are capable of modifying phenotype, such as fear and anxiety responses, in subsequent offspring [13-15].
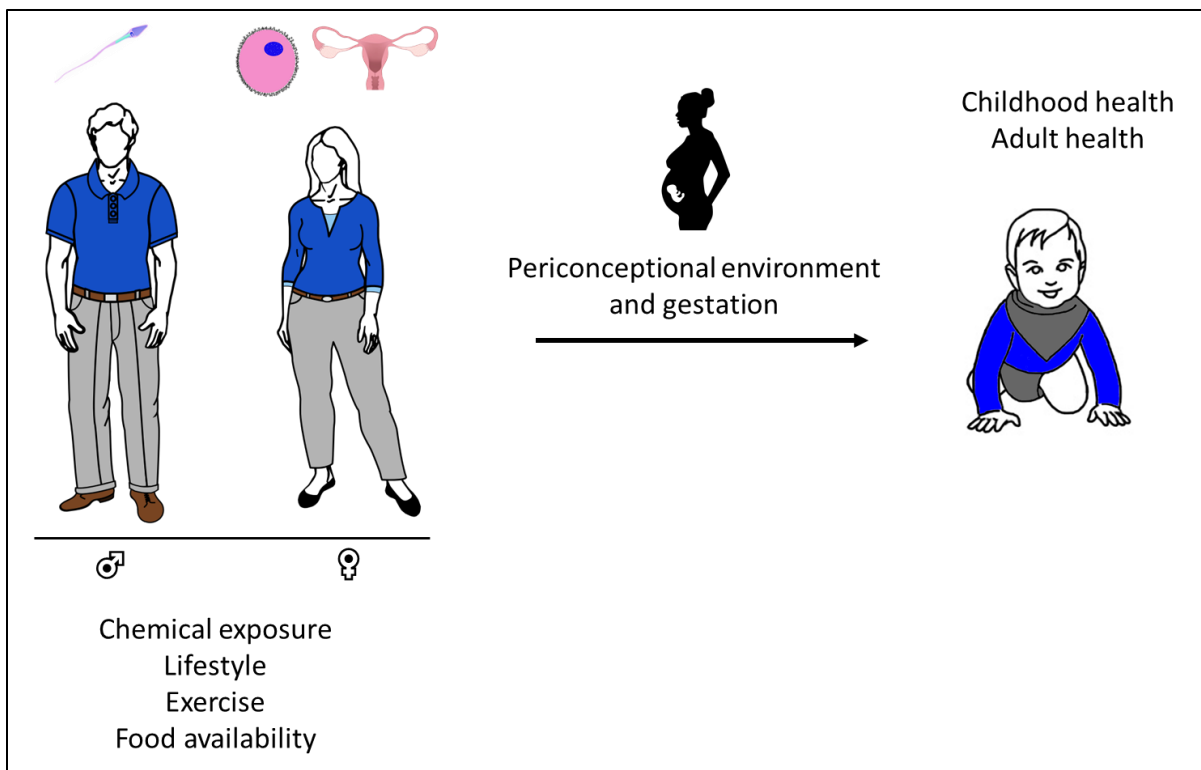
Figure 1.1. DOHaD maternal and paternal contributions. Maternal and paternal germlines may be influenced by their environments and stimuli, such as chemical exposures and food availability. On the maternal side, the uterine condition can also be affected by the maternal environment. Environmental insults to parental reproductive tissues and the developing fetus can influence the phenotype (e.g. health) of the offspring during childhood and into adulthood.

Accordingly, environmental perturbations in the human may also affect offspring. While the long-term observations are required to demonstrate the isolated/specific paternal contribution in humans to the offspring and subsequent generations, historical records of isolated communities do suggest that parental nutrition influences the disease risk of subsequent generations [16]. Observations of a Swedish parish, Överkalix, suggest that dramatic shifts in pre-pubertal food availability of paternal grandmothers increases the risk for cardiovascular mortality in female grandchildren, via the F1 sons, suggesting a sperm-mediated mechanism. Subsequent work on the Uppsala Birth Cohort Multigeneration Study further suggested that abundant access to food during the slow-growth period (pre-pubertal period) of paternal grandfathers is associated with elevated mortality in grandsons, also supporting a sperm-mediated mechanism [17]. The well-documented 1944-45 Dutch famine,

caused by a German blockade of the Western Netherlands during World War II, yielded a cohort of individuals who experienced gestational undernutrition. Maternal undernutrition during gestation of male offspring (F1 generation) resulted in the grandchildren (F2 generation) being prone to obesity [18].

To understand the impact of human paternal environment, it is critical to thoroughly characterize the components of the male gamete, as well as the mechanisms by which common environmental exposures may alter spermatozoal contents. Human populations are exposed to a wide spectrum of environmental contaminants, some of which are considered reproductive toxins [19]. Certain exposures are intentional, such as the case of chemotherapeutic treatments, intended to kill fast-growing cells, including the target cancer cells. However, chemotherapy often has negative repercussions on the reproductive system, up to and including infertility. Other such reproductive toxins include certain heavy metals, such as lead, and endocrine disruptors, such as bis-phenol A (BPA), which can disrupt gametogenesis and reproductive functions [20]. Epidemiological studies have examined the effect of such toxins on the human male reproductive system. However, the basic tenant of epidemiological studies in the male human is to infer how one or more substances alter the hormonal profile, seminal characteristics, or both. Unfortunately, except in extreme cases of sperm parameter abnormalities, such as azoospermia or globozoospermia, the clinical utility of traditional seminal characteristics at predicting fertility is limited [21, 22]. The contribution of subtle epigenetic alterations, such as DNA methylation and histone marks, to reproductive phenotypes and fertility potential has, until recently, received little attention.

The mechanisms underlying phenotypic changes in DOHaD are largely proposed to be epigenetic. Although genotype and consequent gene-environment interactions certainly account for a portion of a response, controlled murine experiments indicate a long-lasting effect that perpetuates a phenotype across an individual's lifetime, and possibly to the individuals offspring [11, 23-25]. This long-lasting mechanism occurs despite the mutability of

epigenetic marks across an individuals' lifetime [26-28]. Epigenetics is broadly defined as the non-permanent alteration of DNA structure. Common epigenetic alterations are DNA methylation, histone modifications, and chromatin states (euchromatin/heterochromatin). DNA methylation, the addition of methyl groups to individual bases, occurs primarily in the context of methylated cytosines. In mammals, 5-methylcytosine (5-mC) is thought to be the primary type of methylated cytosine [29]. Other epigenetic alterations, while being more transitory in nature, also can alter DNA structure, such as regulatory RNA and TF binding. The involvement of the epigenome in Assisted Reproductive Technologies and the male germline is explored below.

iii. Epigenetics in embryonic development

Human infertility is a common condition, with approximately 12% of couples of childbearing age in the United States experiencing a prolonged time to conception or the inability to conceive [30]. Assisted reproductive technology (ART) therapies to address infertility range from non-invasive to invasive. Non-invasive therapies include Timed Intercourse (TIC) (having intercourse during the predicted window of ovulation), and Intra-Uterine Insemination (IUI) (manual placement of sperm inside a woman's uterus to facilitate fertilization). Invasive therapies include Gamete Intrafallopian Tube Transfer (GIFT) (external mixing of sperm and oocyte, followed by immediate transfer to the fallopian tube), *In Vitro* Fertilization (IVF) (*in vitro* oocyte fertilization, followed by transfer to recipient uterus), and Intra-Cytoplasmic Sperm Injection (ICSI) (IVF, with manual fertilization via direct micro-injection of sperm into the egg. Controlled ovarian hyperstimulation, commonly known as superovulation, is the administration of exogenous gonadotropins to promote the release of multiple oocytes in a single cycle. Superovulation is often used in conjunction with IVF and ICSI, in order to obtain viable embryos and increase chances of pregnancy.

With regards to offspring health, IVF is considered a safe procedure and is now commonplace. Approximately 5 million children have been born through the use of ART [31]

since the first IVF child was born approximately 36 years ago. Despite the positive impact of ART, this technology presents an atypical nutritional, biochemical, and hormonal environment to the developing embryo. Whether this has an effect on the long-term health of the conceptus and the magnitude of effect remains to be resolved in humans.

The "natural" ovulation cycle involves sequential hormone peaks, that result in follicle recruitment, maturation, and release at ovulation. Concurrently, the uterine endometrium undergoes proliferation and vascularization, to yield a thick endometrial layer capable of supporting embryo implantation and growth [32]. Intercourse (or IUI) around the time of ovulation, when successful, fertilizes the oocyte within the fallopian tubes. The resulting embryo then continues its journey to the uterus, where implantation takes place.

While each part of the "natural" fertilization process is designed for optimal oocyte selection, fertilization, and embryonic development, the focus of IVF is the production of sufficient numbers of high-quality oocytes for subsequent fertilization. Over the years, IVF procedures have been refined to optimize production of high-quality embryos [33]. However, the superovulation procedures, in addition to the psychological and physical stresses of the female patient [34], neglect the endometrial health and implantation capacity. In the context of DoHAD, the periconceptional environment of a child born through IVF is comprised of the oocyte milieu (superovulation, oocyte retrieval, and gamete storage), fertilization, and the environment surrounding early embryonic development (*in vitro* growth and uterine environment). Modification of this pre-implantation environment by assisted reproductive technologies may have unintended consequences on embryo growth and resulting health of the child. Additionally, the growing use of oocyte/embryo cryopreservation may further modify the embryonic environment. Overall, the use of ART procedures, particularly IVF and ICSI, may unduly alter the epigenome of the fetus. This concept is explored in Chapter 2.

iv. sperm RNA

The male gamete, spermatozoa, serves to deliver the paternal DNA and other cargo to the oocyte [35]. Sperm are produced in the seminiferous tubules of the testis through a series of pre-meiotic, meiotic and post-meiotic processes known as spermatogenesis (reviewed in [36]). Within the seminiferous tubule, post-meiotic processes require the transformation of a round spermatid (with typical cellular structure) into the unusual spermatozoal structure, containing a flagellum, highly condensed nucleus, mitochondrial midpiece, and very little cytoplasm. After sperm differentiation within the testis, the sperm enter the epididymis and undergo additional maturation during the approximately two weeks of epididymal transit [37]. This epididymal maturation potentiates the sperm's fertilizing ability and motility [38].

In the past, spermatozoa were perceived to merely be a vessel for the paternal DNA. However, sperm were subsequently shown to not only contain DNA, but transmit RNA, proteins and chromatin marks to the embryo [35]. Mature spermatozoa are known to be transcriptionally and translationally inert. However, labeled probes, microarrays and eventually, RNA-sequencing, have identified a wide range of RNAs present in sperm, both in mammals and other organisms [35]. The majority (~85%-96%) of sperm DNA is packaged in the specialized proteins called protamines, which provide a compact chromatin structure ~10 fold more compact than the nucleosome structure generated by histones [39-41]. However, sperm still contain some histones, which are known to preferentially localize at promoters of developmental transcription and signaling factors [39, 42]. In addition to proteins directly involved in chromatin structure, other proteins with diverse roles are also delivered to the embryo [43]. This additional cargo, which is delivered to the oocyte, may then influence fertilization and subsequent fetal health. Recent work on the mammalian epididymis has shown that epididymal exosomes (epididysomes), which contain proteins, RNAs and other molecules, are integrated into sperm during epididymal transit [44, 45]. The RNA profiles observed in the ejaculated spermatozoa thus reflect the final outcome of spermatogenesis,

which includes both RNAs generated in preparation and those acquired during epididymal maturation for transmission to the future embryo.

Recent work in humans currently indicate that sperm epigenetic marks and RNAs can be modified by paternal environment [46]. Comparison of obese males to lean controls suggested differential profiles of non-coding RNAs and DNA methylation. Additionally, in a cohort of obese men undergoing bariatric surgery, longitudinal observations of sperm DNA methylation show shifts in methylation at several genomic loci, including loci involved in appetite regulation [Donkin cell metabolism 2016]. In a human trial of endurance training, 6 weeks of endurance training was sufficient to remodel both sperm RNA levels and DNA methylation [47]. While DoHAD studies in human have primarily focused on the maternal contributions, the preconceptional environment of the male gamete may also play a role in intergenerational offspring health in human [16, 17].

v. Endocrine disruptors on male reproduction

Like the female reproductive system, the male reproductive system is under the control of the endocrine system. In humans, the hypothalamic–pituitary–gonadal axis (HPG axis) is a primary modulator of reproductive function. As shown in **Figure 1.2**, the hypothalamus produces Gonadotropin-releasing hormone (GnRH), which influences the adjacent pituitary gland to secrete follicle-stimulating hormone (FSH) and luteinizing hormone (LH) into the bloodstream. LH subsequently enters the Leydig cells and promotes testosterone production. Concurrently, FSH promotes the release of androgen-binding protein (ABP), allowing the binding of testosterone by the Sertoli cells [48]. Androgens, particularly testosterone, are critical to maintaining spermatocytes and completion of meiosis. Although FSH is not required for spermatogenesis, the targeted uptake of androgens by Sertoli cells optimizes spermatogenesis. While testosterone serves as a negative regulator of the hypothalamus and pituitary gland, thus producing a negative feedback loop in the HPG axis, the contrasting functions of activins and inhibins (**Figure 1.2**) also likely play a role regulatory role in

maintaining spermatogenesis [49]. Among the many functions, activin is a growth factor that can stimulate the pituitary gland to secrete FSH. In contrast, inhibin, which is produced in reproductive tissues (as well as other tissues), can suppress pituitary FSH production [50, 51]. Disruption of the HPG axis in males, by environmental factors, age, or genetic mutations, can result in altered spermatogenesis or hypogonadism [52-54].
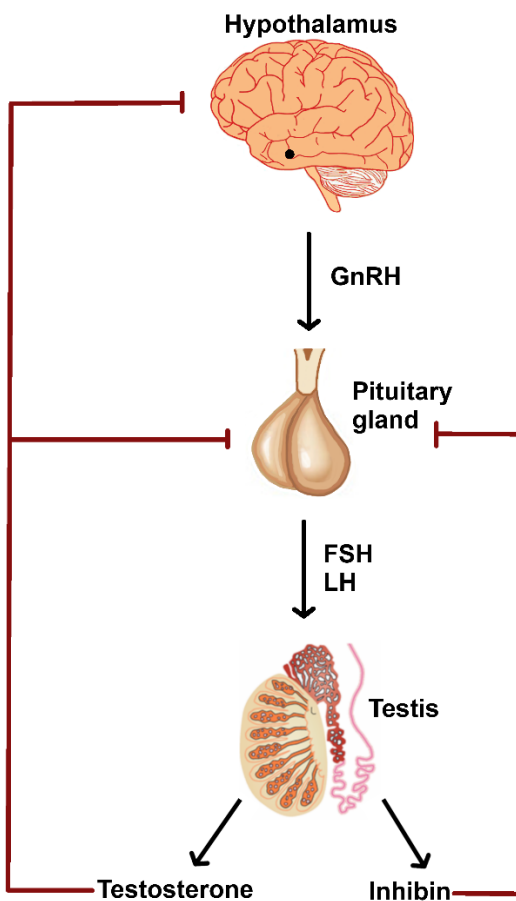
Figure 1.2. Hypothalamic–pituitary–gonadal axis in the adult human male. The hypothalamus within the brain, whose location is approximated by a black dot, secretes Gonadotropin-releasing hormone (GnRH). GnRH stimulates the adjacent pituitary gland to secrete follicle-stimulating hormone (FSH) and luteinizing hormone (LH) into the bloodstream. FSH and LH then enter the testis, promoting the production of androgens, particularly testosterone, and inhibin. Testosterone acts in a paracrine manner to create a negative feedback loop, suppressing GnRH secretion by the hypothalamus, and also suppressing FSH and LH production at the level of the pituitary gland. Concurrently, inhibin produced the Sertoli cells acts in an endocrine manner to suppress FSH production by the pituitary gland.

Endocrine disruptors, exogenous chemicals that can mimic or alter hormonal responses, are a prevalent feature in urban environments [19]. A heterogeneous collection of natural and synthetic chemicals have been identified as likely EDs, including several well-publicized pesticides, such as dichlorodiphenyltrichloroethane (DDT), and plastic components, such as bisphenol A (BPA) and phthalate esters [55]. Phthalates, suspected endocrine disruptors, are commonly used as solvents and plasticizers in consumer products, such as polyvinyl chloride. They have also been incorporated into coatings used in medications [56, 57]. Phthalates have been noted to act on peroxisome proliferator-activated receptors (PPAR) [58, 59]. Additionally, different phthalate species, including phthalate metabolites, have different capacities for modifying an endocrine response [58-60]. Although considerable literature suggests that gestational and neo-natal phthalate exposure is detrimental to reproductive function [61], the health effects of phthalates at environmentally relevant doses in adult humans is still uncertain, particularly in the adult male, although existing human studies are described below. Interestingly, chronic *in vivo* exposure of rats to the tolerable daily intake (TDI) of BPA altered protein expression and histone acetylation [62, 63]. While rats and mice cannot replace pertinent observations in humans, such model organisms do suggest that humans may also be subject to subtle reproductive changes when challenged with endocrine disruptors.

Epidemiological studies on adult phthalate exposures and semen parameters have associated elevated phthalate levels with abnormal sperm morphology [64], sperm concentration [65, 66], oxidative stress [67], and DNA damage [68]. Among IVF couples, phthalate levels in the male partner are inversely correlated with high-quality blastocysts [67]. While a controlled study of adult human males exposed to a low, chronic dose of phthalates [69, 70] indicates hypothalamo-pituitary-testis axis dysregulation and declining sperm motility, the mechanisms directly underlying the spermatozoal modifications in such individuals are yet undetermined. The intergenerational and transgenerational impact of such phthalate

exposures in the adult human male remains unknown. However, animal models, and limited data in human, suggests that paternal experiences, such as diet or stress, can have phenotypic consequences in the offspring. Such intergenerational effects are expected to be mediated through epigenetic mechanisms, such as chromatin structure or RNAs delivered by the spermatozoon at fertilization [17, 71-73].

vi. Overview

The contribution of the male gamete and pre-implantation environment to offspring health, while important, is relatively understudied. ART procedures, such as IVF and ICSI, can alter the pre-implantation environment of the oocyte, growing embryo, and uterus. Despite the manipulations involved in ART, offspring derived from ART procedures are generally healthy [74]. However, such offspring may exhibit subtle epigenetic effects, exemplified by the small, but elevated, risk of congenital defects due to imprinting disorders, such as Beckwith-Wiedemann and Prader-Willi syndromes [75]. In Chapter 2, I explore the role of IVF/ICSI protocols in the DNA methylation of infants. The study presents DNA methylation changes in infants conceived through the use of Fresh Embryo Transfer (ET), Frozen Embryo Transfer (FET), or Intrauterine Insemination (IUI). The relative similarity between FET and IUI indicates that the fetus's DNA methylation is either unchanged by, or altered in an inconsistent manner, by FET. This is likely due to improved uterine receptivity in FET and IUI conceptions compared to ET. Overall, this study supports the use of FET in IVF/ICSI procedures [76].

During fertilization, the male gamete contributes DNA, RNA, and epigenetic marks. Spermatozoal RNA can also provide an epigenetic mechanism. Chapter 3 explores the RNAs present in the male germline and early embryo. This analysis indicates that, in addition to annotated transcripts, numerous intergenic and intronic RNAs are present in human spermatozoa. These novel RNAs were identified using a pipeline for discovering intergenic RNAs from standard read alignments. Implementation of the RNA Element Discovery Algorithm (REDa) on somatic, embryonic, and germline tissues revealed presence of both

novel RNAs and exonic RNAs in each tissue [77]. Spermatozoa contain many more novel RNAs than somatic (e.g. liver) tissue, suggesting that transcription of intergenic and intronic RNAs may be important for spermatogenesis. Notably, genomic repeats exhibit a shifting transcriptional enrichment pattern across spermatogenesis and early embryogenesis.

Using the spermatozoal RNAs identified by the REDa approach, Chapter 4 examines the impact of phthalate exposure on male reproduction. The Mesalamine and Reproductive Health Study (MARS) (NCT01331551) was initiated (https://clinicaltrials.gov/) to directly address the physiological effect of *in vivo* di-butyl phthalate (DBP) exposure on male reproduction. Within the MARS study, subjects with Inflammatory Bowel Disease (IBD) were exposed to longitudinally alternating DBP exposures. Using a cross-over design and longitudinal data structure, each subject acts as their own control, thus mitigating genetic variation and environmental variation (e.g. lifestyle and exposome) that often complicates causal assessment in epidemiology. The results of longitudinal modeling suggest that exposure to, or removal of, high DBP produces transcriptomic effects that require longer than one spermatogenic cycle to resolve, if at all. While the two study arms exhibit enrichment of different biological pathways, the H_1BH_2 arm, which initiates the study on high DBP, displayed activation of oxidative stress and DNA damage response pathways. Network analysis of small RNAs and genomic repeats also suggest that transcription of small RNAs and genomic repeats contribute to spermatid development. Together, this work provided insight into both the influence of phthalate on the male germline and the RNA dynamics of human spermiogenesis.

CHAPTER 2

"THE IMPACT OF ASSISTED REPRODUCTIVE TECHNOLOGY ON THE EPIGENETIC PROFILE OF NEWBORNS"

This chapter was adapted from the following publication:

i. Summary

Little is known of the genome-wide effect of assisted reproductive technologies (ART), on the genome and epigenome of the conceptus. To address this shortfall, I have examined the DNA methylation profile of newborns conceived naturally, or through the use of intrauterine insemination (IUI), or *in vitro* fertilization (IVF) using Fresh or Cryopreserved (Frozen) embryo transfer. Newborn methylation levels of these four different conception types, stratified by gender, were compared using the HumanMethylation 450k platform. Perturbation of probe clusters within genes and enhancers suggests that the newborns born from ART possess a dramatically different methylation profile compared to those naturally conceived. Intriguingly, there was a striking similarity of the methylation profile of IUI and IVF-Frozen embryo transfer infants, but not IVF-Fresh. This suggests a possible reduction of epigenetic aberrations in the IVF conceptions using cryopreservation and implicates that a resetting mechanism is acting upon cryopreserved embryos. These results are in accord with the observed reduction in birth defects using those protocols that employ cryopreserved embryos. Periconceptual nutrition is known to alter epigenomes of offspring at specific loci termed metastable epialleles (MEs). IVF culture conditions can mimic various nutritional conditions experienced by the early embryo. With this consideration, analysis of the ART methylation changes in MEs was undertaken to test the hypothesis that ME loci were sensitive to early nutritional exposure. IVF

culture conditions and parental infertility showed consistently altered methylation at certain MEs. This is the first study to reveal an impact of ART or fertility status on MEs and suggests a lasting epigenetic effect of IVF nutrition on the developing embryo.

ii. Introduction

The DoHAD hypothesis [78] suggests that environment during the periconceptional period, as well as later stages of embryonic, fetal, and postnatal growth, can persistently impact health. Epigenetic modifications, including histone modifications and DNA methylation, are suspected to be one of the mechanisms by which prenatal environment influences offspring health. Accordingly, nutritional status at the time of conception in rural Gambian women altered the DNA methylation profile at metastable epiallele (ME) loci in the children. This effect was modulated, at least in part, by methyl donor availability in the food sources available at the time of conception [79, 80].

Assisted reproductive technology (ART) provides infertile couples several treatment options, including ovulation induction followed by intrauterine insemination (IUI), *in vitro* fertilization (IVF), and intracytoplasmic sperm injection (ICSI). Despite the positive impact of ART on fertility outcomes, these procedures present an atypical nutritional, biochemical, and hormonal environment to the developing embryo. Murine studies suggest that the transient stresses that IVF places on the growing embryo can result in a considerable change in gene expression, metabolism, and growth trajectory [81-83], which can persist into adulthood [84-87]. Although there are likely several underlying mechanisms for these effects of IVF, one such mechanism that has been proposed is the increased incidence of random epigenetic errors, including changes in imprinted genes [83, 88].

ART procedures are generally accepted as safe for the mother and conceptus in humans. However, there is an elevated risk of birth defects, neurologic disorders, and imprinting disorders in conceptions generated through IVF and ICSI [89, 90]. In addition to the increased rate of imprinting disorders in the ART population [75, 91], several recent studies

comparing the metabolic and cardiovascular health of children conceived by ART with those conceived naturally indicate a trend toward impaired glucose metabolism and cardiovascular function [92-94]. Such health risks observed in ART-conceived children may result from underlying epigenetic aberrations presented by parental infertility or ART itself. However, it is important to note that evidence for concerted epigenetic changes in the ART population [95-100] is countered by a number of studies asserting that the epigenome of ART-conceived offspring is essentially unchanged [101-104].

Studies investigating the effect of ART on the human epigenome often observe a limited number of genomic loci, which are usually regions of known interest associated with imprinting syndromes. However, two studies recently expanded the number of genomic regions examined in ART studies. One described DNA methylation in the promoter regions of 736 select genes, which were primarily associated with imprinting and growth regulation, from 10 IVF and 13 *in vivo*–conceived children [95]. The second used an Illumina Infinium Human Methylation27 Beadchip array, identifying 733 CpG sites of differential methylation when comparing the cord blood of 8 IVF and 10 naturally conceived children [98]. However, these two studies did not address the impact of infertility, which is the primary reason why ART is prescribed, nor did they provide a genome-wide assessment. By employing the Illumina Infinium HumanMethylation450k BeadChip on newborn blood samples, we began to fill in these gaps, revealing that a larger span of genomic sites may be impacted by parental infertility and ART than previously appreciated. A targeted investigation of children born after transfer of cryopreserved or fresh ICSI embryos further delineated the epigenetic impact of these two different ART protocols, while suggesting that embryo cryopreservation may indirectly improve epigenetic outcome.

iii. Materials and Methods

Acquisition of Newborn Bloodspots

Newborn residual bloodspots, which were collected from Michigan newborns between 24 and 36 hours after birth for metabolic screening, were obtained from the Michigan Neonatal Biobank (www.mnbb.org/). After Michigan Department of Community Health (approval #037913MP4X), Wayne State University, and Biotrust institutional review board approval, 120 women (40 per group) were selected who conceived after one of three types of infertility treatments (IUI, ICSI with fresh embryo transfer [FH], or ICSI with frozen embryo transfer groups [FZ]). Sample requests were then given to the Michigan Biotrust for Health, and bloodspots were subsequently requested from the newborn of each of the women in the different categories through the Michigan Neonatal Biobank using storage codes only (all data were deidentified). For the three procedure categories, a total of 18 IUI, 38 FH, and 38 FZ bloodspots were obtained for calculation of DNA methylation, as detailed in **Figure 2.1A**. While the clinical parameters of the newborn and mother for the samples analyzed in this study were not known, the majority of bloodspot samples represent individuals born in the Detroit and metro Detroit regions. As a naturally conceived control, 16 male and 27 female bloodspots were also examined. Intensity Data (IDAT) files from the 450k assay of naturally conceived (NAT) newborn bloodspots were graciously provided by Dr. Douglas Ruden, Wayne State University [105]. The NAT samples were also previously obtained from the Michigan Neonatal Biobank. In total, 137 individuals were utilized in methylation comparisons. All methylation analyses were performed at the Wayne State University Applied Genomics Technology Center.
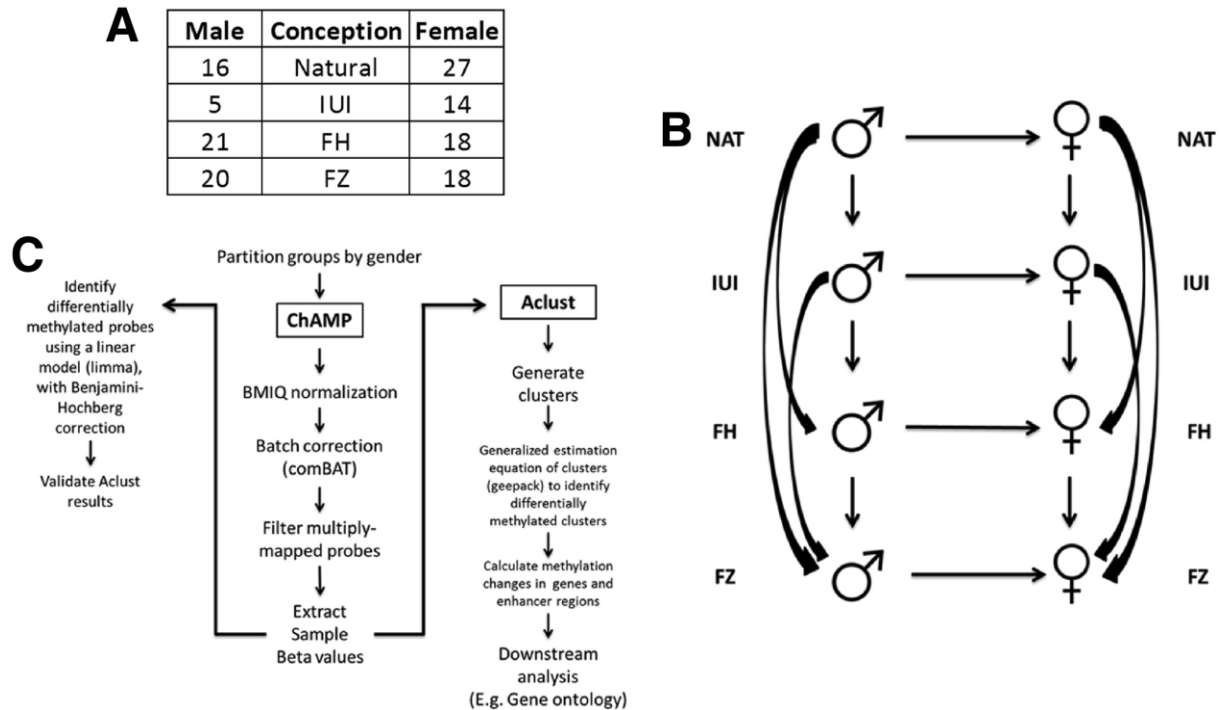
Figure 2.1. Samples and conception group comparisons subject to differential methylation analysis. (A) Number of individuals analyzed in this study, according to conception type and gender. (B) Directional arrows indicate that the conception group at the source of the arrow is the methylation dataset (control) against which that at the termination of the arrow (case) is being compared. (C) Pipeline provided by ChAMP was used to filter, normalize, and apply batch correction, to obtain corrected methylation values. Aclust was then implemented to calculate methylation changes in probe clusters, followed by downstream analyses. Concurrently, the same corrected methylation values were analyzed using a linear model to calculate differential methylation of individual probes and verify the results obtained from Aclust.

Study Design and Data Processing

To assess the effects of infertility treatments (specifically ICSI and cryopreservation) on genome-wide DNA methylation profiles of newborns, a case–control design was implemented (**Figure 2.1B**). Neonatal bloodspots from newborns conceived through unassisted (NAT) and IUI conceptions provided controls from fertile and infertile backgrounds, respectively. In this case–control design, a given conception group considered as a "case," such as ICSI fresh embryo transfer, was matched to a "control," such as IUI. The unassisted conception control, NAT, served as a control for the IUI, FH, and FZ conception groups. The DNA extracted from bloodspots was assessed for DNA quality (**Appendix A**) and assayed

using the Illumina Infinium HumanMethylation450 BeadChip array. The Chip Analysis Methylation Pipeline (ChAMP) pipeline was utilized for processing the datasets into methylation profile (as measured by β-values) (**Figure 2.1C**).

Each conception comparison, outlined in **Figure 2.1B**, addressed a specific query. Comparison of either FH or FZ with IUI was undertaken to identify ICSI-associated changes within the group of parents requiring artificial insemination to conceive. Within the ICSI (FH and FZ) group, differences between FH and FZ were assessed to discern the effects of fresh embryo transfer and cryopreserved embryo transfer. Subsequent comparison of IUI, FH, or FZ with NAT identified differences between children born to parents undergoing infertility treatment and those born to fertile parents. Methylation changes between males and females within the same conception group were undertaken to examine how the epigenome might differ between genders [106-108].

The ChAMP pipeline was used to analyze the 450k signals, normalize the methylation values to produce β-values (the proportion of methylated CpG sites) and identify the differentially methylated probes [109]. Blood cell distributions of each sample were estimated using the estimatecellcount function in minfi. The single value decomposition function of ChAMP provided the relative influence of plate, assay characteristics, and blood cell proportions on sample methylation. In an effort to minimize technical variation, all multiply-mapped probes, as well as probes containing single-nucleotide polymorphisms in either the target CpG or the 10 bases of probe closest to the target CpG site [110, 111], were removed from differential methylation analysis. This excluded 91,058 probes from analysis, leaving 394,454 probes prior to calculating differential methylation. All sample comparisons were subjected to batch correction using ComBat, in order to eliminate the effects of array batch. To eliminate the risk of identifying false positives due to gender imbalances, I implemented a strategy of comparing results from male and female groups separately. This approach

increased the confidence in assigning genes and other genomic loci as differentially methylated, while removing gender as a confounding variable.

Batch-corrected β-values from the ChAMP pipeline were used in the A-clustering algorithm [112]. The algorithm generates clusters of correlated autosomal probes, followed by general estimating equation (geepack) estimation of the differentially methylated autosomal clusters. Differentially methylated clusters were as having a minimum absolute average effect size (methylation) change of 2.5% and Benjamini-Hochberg adjusted P value of <.05.

Calculating Differential Methylation Using Limma

The β-mixture quantile dilation (BMIQ)-normalized, batch-corrected β-values of all reliable probes produced by the ChAMP pipeline were applied to limma, which calculated differential methylation of individual probes [113]. Differentially methylated probes were defined as having a minimum absolute methylation change of 2.5% and Benjamini-Hochberg adjusted P value of <.05. Limma yielded similar proportions of hypermethylated and hypomethylated probes, when contrasted to clusters of probes, and generally supported the trends in counts of differentially methylated regions between the conception types.

Calculating Differential Methylation of Promoters and Gene Bodies

The locations of all RefSeq genes were obtained from Ensembl, build hg19/GRCh37. Promoters of genes were defined as 1 kb upstream of the unified gene start site. The identification of clusters and probes within gene bodies or genomic loci was performed using a series of bedtools utilities and custom R scripts. For a given comparison, all differentially methylated clusters (filtered for a methylation change of 2.5% and Benjamini-Hochberg corrected P value of <.05) with at least 75% of the cluster intersecting a gene body or promoter were assigned to that particular gene. Because of the potential for numerous non-differential clusters to reduce average gene body methylation below the threshold of 2.5% methylation change, only differentially methylated clusters were considered in the calculations for gene body methylation. Therefore, for a given gene, the methylation change was calculated from

the average of all effect size (methylation) changes of all differentially methylated clusters assigned to that gene. Gene bodies were then further filtered for a minimum absolute average effect size (methylation) change of 2.5%. Differentially methylated genes, as calculated using Aclust clusters, are referred to as hypermethylated or hypomethylated according to an effect size (methylation) change of greater than 0.025 or less than −0.025, respectively.

Regulatory/enhancer Regions

Human regulatory regions and putative human enhancers (indicated here as "regulators"), were obtained from Andersson et al. [114]. Clusters were intersected with the set of 43,011 permissive regulators to identify the clusters overlapping the regulators. The change in regulator methylation was calculated as the average of all differentially methylated clusters located in the given regulator. To avoid spurious associations of methylation changes, all differentially methylated regulators were required to have a minimum absolute average methylation change of 2.5%.

Genome Annotation and Relative Enrichment

Annotation of significant clusters with respect to promoters, introns, exons, and intergenic regions was calculated with the "Annotation & Statistics" function of Genomatix (July 2014 build), using the December 2013 Eldorado annotation of the hg19 genome.

iv. Results

Study Sample Characteristics

As the cohort was composed entirely of newborns, the postnatal environment was expected to have minimal effect on DNA methylation. While cellular composition of blood samples can affect methylation profiles, the estimated cellular composition of each sample did not significantly impact the methylation status in the cohort. The general differences among the samples were small, as determined by pairwise comparison of all autosomal probe β-values (minimal correlation of at least 0.90). This is concordant with previous studies, which

suggest that the global pattern of methylation of prepubertal children is not affected by ART [101, 103]. Supporting the previous observation, I found that principle component analysis (unsupervised clustering) did not effectively cluster the NAT, IUI, and ICSI groups. This recapitulated previous observations that methylation profiles of ART and control individuals do not segregate in clustering analysis [95].

Differential Methylation of Probe Clusters

As outlined in **Figure 2.1B**, the conception groups were compared to identify differentially methylated probes and clusters. Aclust was used to identify each cluster of two or more correlated probes and calculate differential methylation at each cluster [112]. While numerous previous studies have listed differential methylation at individual CpG sites, the biological impact of a single CpG is questionable. Therefore, the Aclust approach was used to identify methylation changes across multiple adjacent CpG sites, given that methylation differences that extend over multiple CpG sites can be viewed as confirmatory and are likely to have a biological impact. To avoid spurious associations of methylation change, regions of interest were required to contain multiple differentially methylated probes, as well as exhibit a minimum absolute average methylation change of 2.5% [112]. As summarized in **Figure 2.2A,B and Figure 2.3A**, several interesting differential methylation patterns emerged. The naturally conceived group exhibited considerable differential methylation when compared to all three assisted conception groups. Hypomethylated clusters were observed more frequently within the ART groups (**Figure 2.2B**), concordant with a similar result in a non-ICSI sample group [98]. Within each conception group, differentially methylated clusters in females (compared with the male control) were more frequently hypermethylated than hypomethylated (**Figure 2.3A**). Comparisons between three assisted conception groups revealed considerable differences between FH and IUI or FH and FZ, while, in contrast, there were comparatively fewer differences between IUI and FZ. It should be noted that the A-clustering algorithm used for calculating differential methylation is a comparatively recent technique

[112]. Therefore, the possibility that the A-clustering algorithm was responsible for the unexpected patterns in methylation changes between assisted conception types was considered. In order to verify the A-clustering results with a different method, a commonly used linear modeling algorithm, limma, was also used to identify differentially methylated probes [113]. Based on the trends observed with A-clust and limma approaches, the similarity between the IUI and FZ was considered reliable. This similarity was independent of BMIQ normalization and was not due to the batch correction or the A-clustering algorithm or inadvertent sampling bias, as shown by random subsampling.
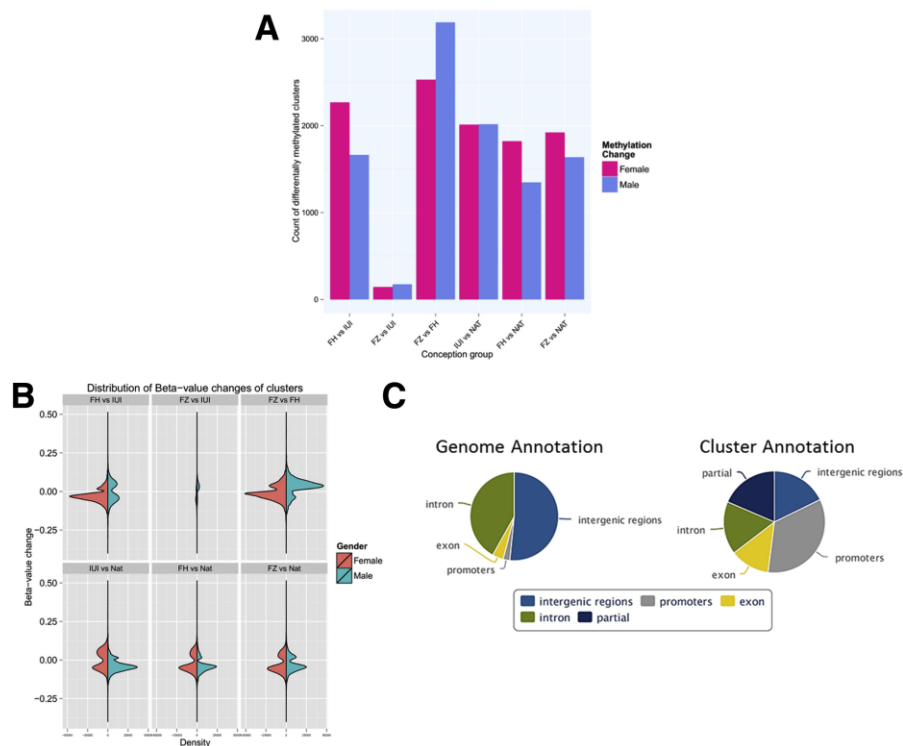
Figure 2.2. Intracytoplasmic sperm injection and IUI compared with NAT show differential methylation of clusters and gene bodies. (A) Total counts of differentially methylated clusters between conception groups for males and females, shown in blue and pink, respectively. Red bracket indicates comparisons of FH and FZ with IUI, presenting the greater degree of differential methylation in the FH vs. IUI comparison than that of FZ vs. IUI. (B) Distribution of the change in the β-value for statistically significant clusters, as a function of conception comparison. Changes in the female and male comparisons are shown in light red and blue, respectively. (C) Pie chart labeled "Genome Annotation" indicates the proportion of the human genome that lies within exons, introns, promoters, and intergenic regions. The pie chart labeled as "Cluster Annotation" provides the proportion of all statistically significant clusters (regardless of methylation change) that overlap exons, introns, promoters, and intergenic regions. Clusters that overlay one or more features are denoted as "partial."
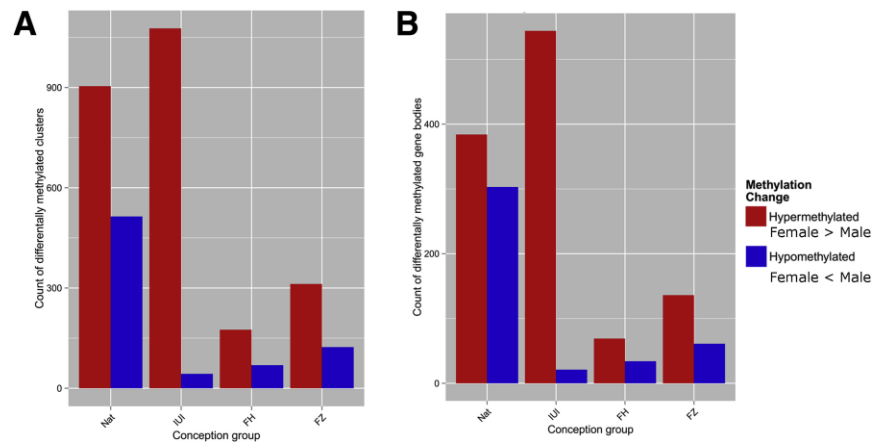
Figure 2.3. Trends in differentially methylated clusters between males and females of identical conception groups. Hypermethylation in the female group (compared with a male control) is shown in red, whereas hypomethylation is presented in blue. (A) Counts of clusters differentially methylated between females and males of the same conception group. (B) Counts of gene bodies differentially methylated between females and males of the same conception group.

Differentially methylated clusters tended to be associated with promoters and exons of protein-encoding loci (**Figure 2.2C**) with β-value changes averaging approximately 5% (**Figure 2.2B**). Several clusters exhibited changes larger than 10%, including a hypomethylation of the Speriolin-like protein (SPATC1L) promoter in IUI and ICSI (FH and FZ) when compared to NAT (with the hypomethylation frequently exceeding 10%).

Characteristics of Differential Methylation

Methylation changes between the individual conception methods were first examined with respect to the promoters and gene bodies of all unified RefSeq genes. Differential methylation in gene bodies was observed more frequently than differential promoter methylation. **Figure 2.4** summarizes the counts of differentially methylated gene bodies, which reflected the overall number and direction of methylation changes of differentially methylated clusters for each given comparison. Collectively, these gene-specific methylation differences may reflect phenotypic alterations [115] that may occur in the various conception groups.
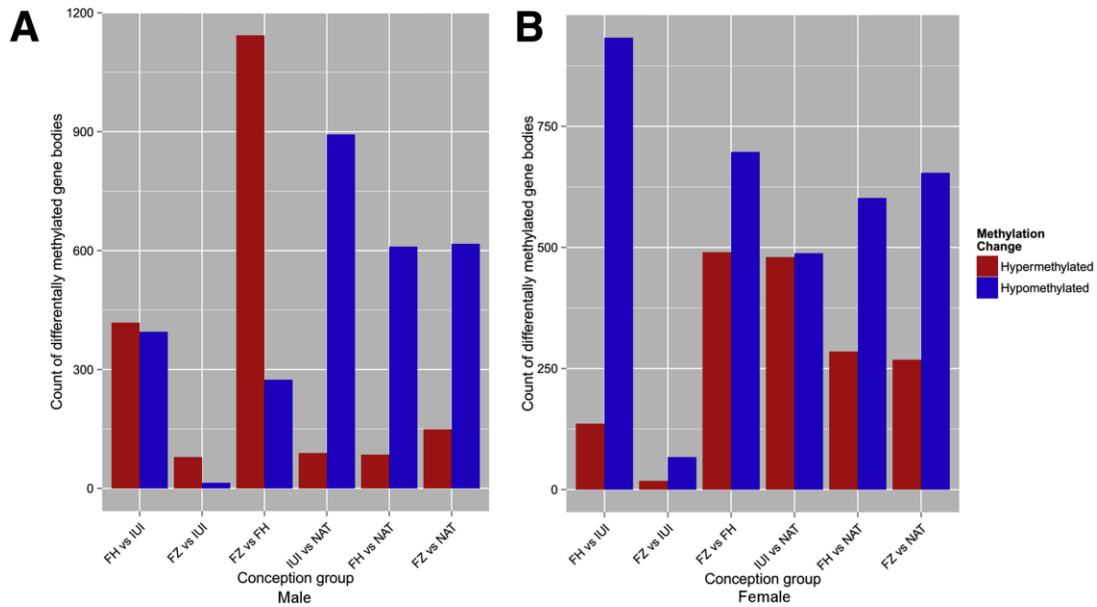
Figure 2.4. Intracytoplasmic sperm injection and IUI show considerable differential methylation in gene bodies compared with NAT. Bars indicate the counts of differentially methylated gene bodies between conception groups. Counts of differentially methylated gene bodies generated for (A) male- and (B) female-specific comparisons. Hypermethylated and hypomethylated gene bodies are represented in red and blue, respectively.

In addition to providing coverage of RefSeq genes, the Infinium HumanMethylation450 BeadChip contains probes designed for various intergenic regions. These regions can encompass regulatory loci (e.g., enhancers and silencers/regulators) residing outside of promoters and gene bodies. Such regulatory loci can act independently or in concert with proximal regulatory regions, and consequently alter chromatin organization and tissue expression [116, 117]. Using putative regulators identified from the FANTOM5 cap analysis of gene expression atlas, changes in the methylation status of regulators [114, 118] between conception types was examined.

Numerous regulators were differentially methylated when the three assisted and NAT conception groups were compared (**Appendix B**). Fewer regulators were differentially methylated when FZ was compared with IUI, as opposed to when FH was compared with IUI or FZ, reflecting the relative similarity of the FZ and IUI groups (**Appendix B**). As shown in **Table 2.1**, certain regulators were also altered between males and females of the same

conception group, suggesting that certain enhancers may exhibit gender-specific methylation patterns. A total of 21 regulators exhibited altered patterns of methylation between the genders in two or more conception groups (**Appendix C**). This common group of 21 regulators may play therefore a role in sexual dimorphism of the human fetus [86, 119].

Table 2.1. Counts of enhancers altered between males and females of identical conception groups.

	Female vs Male, NAT	Female vs Male, IUI	Female vs Male, FH	Female vs Male, FZ
Hypermethylated	28	44	5	12
Hypomethylated	15	4	5	10
Total	43	48	10	22

Comparing ICSI (FH and FZ) and IUI with a NAT control showed that regulators were most frequently hypomethylated (**Figure 2.5 and Appendix D**), which reflected the general trends towards hypomethylation in the complete cluster set for the given comparisons. Interestingly, three hypomethylated regulators were consistently differentially methylated among all three assisted conception groups and NAT control. In addition, 15 regulators were consistently altered between the ICSI (FH and FZ) and NAT control, as well as the subset of 5 regulators (**Appendix D**) altered between all three assisted conception groups and NAT control.

FH vs NAT   **FZ vs NAT**

2 Hyper
22 Hypo   0 Hyper
15 Hypo   4 Hyper
28 Hypo

0 Hyper
5 Hypo

1 Hyper
6 Hypo   0 Hyper
6 Hypo

3 Hyper
12 Hypo

IUI vs NAT
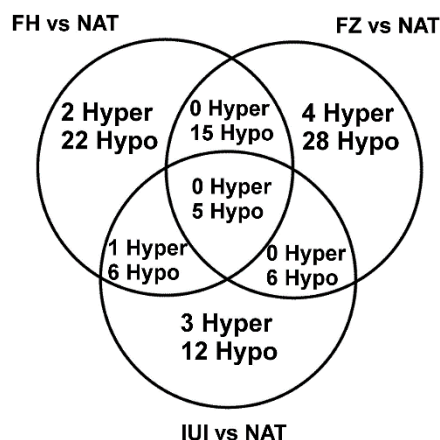
Figure 2.5. Enhancers consistently altered in ICSI groups compared with NAT. Populations of hypermethylated and hypomethylated enhancers altered in the IUI, FH, or FZ vs. NAT comparison were identified and enumerated. Quantities of hyper- or hypomethylated enhancers are denoted as "Hyper" and "Hypo," respectively. The quantity of enhancers found in common between two or more comparisons and exhibiting identical methylation trends (e.g., increased or decreased methylation in both comparisons) are indicated in the intersections of the Venn diagram.

Imprinted Genes are Differentially Methylated in IUI and ICSI Newborns

Imprinted genes play a critical role in fetal growth and are epigenetically repressed in a parent-of-origin-specific manner [120]. Due to this epigenetic repression by cis-acting differentially methylated regions, an imprinted gene will exhibit preferential expression from either the paternal or maternal allele [121]. In humans, the combination of the 450k platform and bisulphite sequencing has allowed for the identification of cis-acting differentially methylated regions, which may act as imprinting control regions [121].

Previous mouse and human studies have investigated changes in the methylation status at imprinted loci, including loci associated with metabolism and fetal growth (e.g., H19 and IGF2) after ART [75, 83, 88, 90]. Recent work has also shown that high-quality human embryos generated from ART exhibit a high degree of DNA methylation errors in imprinted loci [122]. Within the collections of differentially methylated genes, an enrichment of imprinted genes was found when comparing ICSI (FH and FZ) with the IUI control (**Table 2.2**), with hypomethylated genes contributing strongly to this enrichment. This enrichment trend was also observed in the comparison of ICSI (FH and FZ) with NAT. Taken together, imprinted genes were differentially methylated more often than expected by chance when comparing any assisted conception group with NAT (P value: 1.47×10^{-3} to 2.04×10^{-11}). The biological implications of the observed methylation changes in the imprinted genes remains to be determined.

Table 2.2. Imprinted genes are differentially methylated between different conception types. Enrichment of differentially methylated imprinted genes between conception groups, segregated according to gender. Comparisons yielding P-values less than 0.05 and 0.01 are highlighted in red and blue, respectively.

	Female vs Male, NAT	Female vs Male, IUI	Female vs Male, FH	Female vs Male, FZ	FH vs IUI, Male	FH vs IUI, Female	FZ vs IUI, Male	FZ vs IUI, Female	FZ vs FH, Male	FZ vs FH, Female	IUI vs NAT, Male	IUI vs NAT, Female	FH vs NAT, Male	FH vs NAT, Female	FZ vs NAT, Male	FZ vs NAT, Female
Differentially methylated genes	687	565	103	197	813	1069	93	85	1417	1187	982	968	695	887	766	922
Imprinted genes	18	18	4	4	23	23	5	10	27	27	32	25	18	25	27	33
Maternally imprinted genes	7	9	1	2	8	15	3	1	14	16	16	13	8	13	13	14
Paternally imprinted genes	11	9	3	2	15	8	2	9	13	11	16	12	10	12	14	19
Significance of imprinted genes	2.17E-05	1.71E-06	9.48E-03	5.92E-02	5.39E-07	3.66E-05	1.03E-03	2.85E-09	6.17E-05	3.41E-06	1.51E-10	9.05E-07	2.50E-05	1.96E-07	7.32E-10	7.34E-12
Significance of maternally imprinted genes	2.18E-02	7.91E-04	2.89E-01	1.58E-01	1.82E-02	7.83E-05	7.37E-03	2.58E-01	2.89E-03	7.31E-05	8.48E-06	3.24E-04	8.66E-03	1.47E-04	3.68E-05	6.00E-05
Significance of paternally imprinted genes	1.03E-04	4.25E-04	7.56E-03	1.41E-01	1.30E-06	3.83E-02	4.74E-02	7.89E-11	3.38E-03	5.89E-03	2.79E-06	4.76E-04	4.54E-04	2.29E-04	3.18E-06	1.06E-08

The effect of IVF/ICSI conception on long-term health and metabolism is still actively debated. Therefore, our collection of imprinted genes were assessed for known associations with human metabolism. Differentially methylated clusters in the promoter and gene body of the maternally expressed long noncoding RNA H19 were observed when IUI and ICSI (FH and FZ) were compared with NAT, but not when ICSI (FH and FZ) was compared with IUI (**Figure 2.6**) [123]. These observed cluster changes occurred in the paternally methylated H19 Differentially Methylated Region (DMR) [121]. IGF2, a paternally expressed growth-promoting hormone, was also hypomethylated in both promoter and gene body clusters when female IUI and ICSI (FH and FZ) were compared with NAT (**Figure 2.6**). The paternally methylated IGF2 DMR2 region [121], located in the last two exons of the IGF2 gene, was relatively hypomethylated in female IUI and ICSI (FH and FZ) compared with NAT.
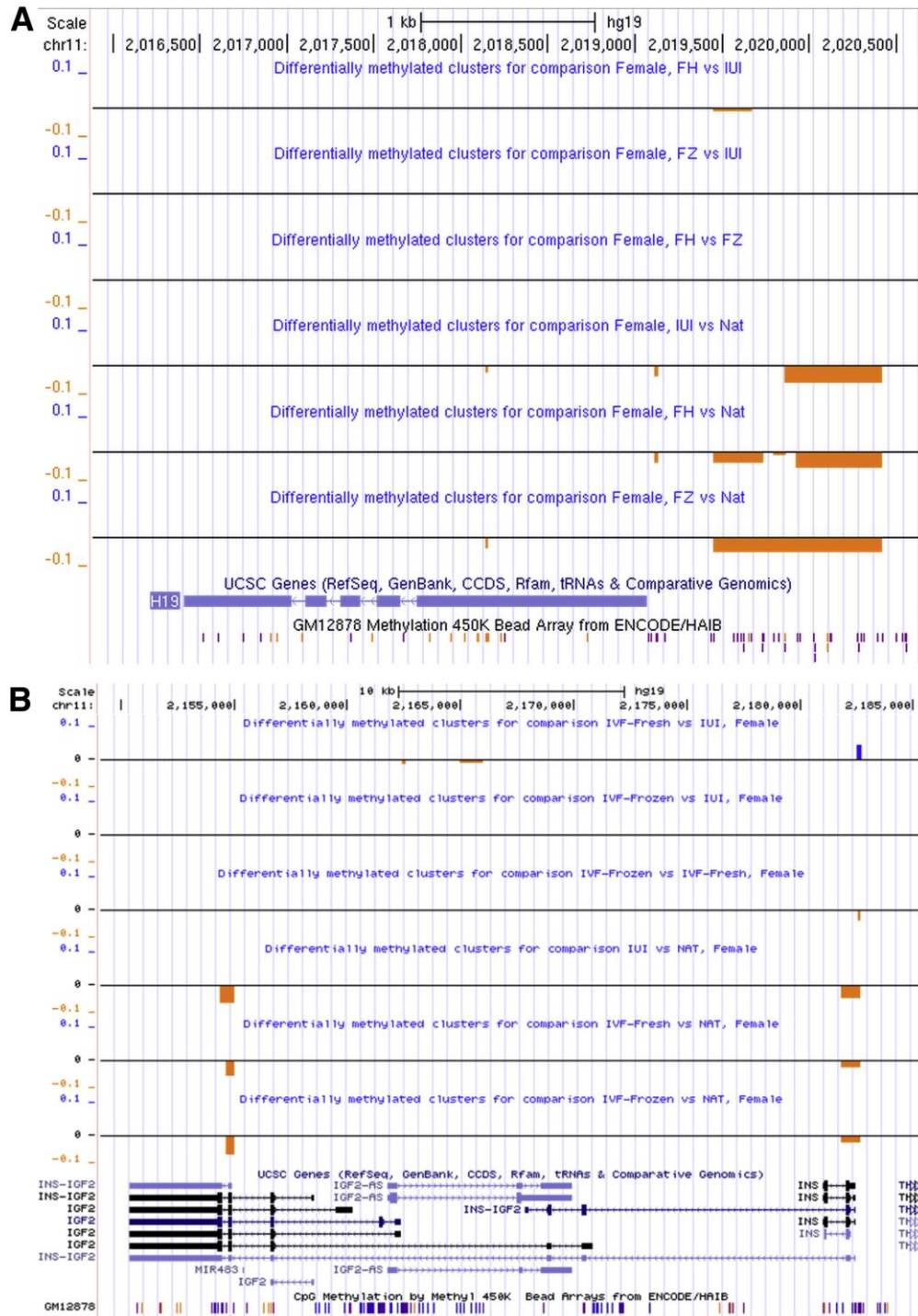
Figure 2.6. Certain imprinted genes associated with metabolism and cancer exhibit differential methylation. UCSC genome browser images of promoter and gene bodies of imprinted genes (A) H19, (B) IGF2. Hypermethylated clusters and hypomethylated clusters are colored in blue and orange, respectively. Cluster heights represent the magnitude of methylation change, in scale with the y axis.

MEs are Altered in Both the Infertile Control and ICSI Newborns

The nutritional environment encountered by the early embryo is among the factors that likely differ between unassisted conception and IUI or ICSI (FH and FZ). Correspondingly, a modulation of metabolism-associated genes suggests that a nutritional component, such as the culture conditions of IVF/ICSI protocols, may affect the epigenome. Therefore, genomic loci known to be impacted by periconceptional nutrition (i.e., metastable epialleles), were assessed.

Metastable epialleles (MEs) were first characterized in mouse models [124] and have recently been identified in humans [79]. DNA methylation at these loci is stochastically established within the blastocyst and is also driven by periconceptional nutritional status [79, 125, 126]. In order to assess effects of the early embryonic nutritional environment associated with ART, the clusters identified as differentially methylated in the comparison of NAT with the three assisted conception groups were overlapped with MEs. Of 109 high-confidence MEs identified in a recent genome-wide screen [127], 22 were informative in the 450k-based analysis. Remarkably, 19 (86%) of these, plus two MEs that were not directly assayed but existed within a cluster, were differentially methylated in at least one of all possible conception group comparisons (**Table 2.3**). Hypomethlation predominated the MEs in all three assisted vs. NAT comparisons. Taken together, these clusters suggest that certain epialleles (e.g., those associated with DUSP22 and SPATC1L) are indeed susceptible to methylation changes in both genders upon IUI or ICSI (FH and FZ) as compared with NAT conception (**Figure 2.7, Table 2.3**). Interestingly, DUSP22, a dual specificity phosphatase, has been implicated as a negative regulator of estrogen receptor-α–mediated signaling and STAT3-mediated signaling, and hypermethylation of the DUSP22 promoter is correlated with Alzheimer's disease risk [128-130]. The function of the protein encoded by SPATC1L is unknown, but it has been shown to be expressed in testis as well as certain cell lines. The paralog of SPATC1L, SPATC1, encodes a novel sperm centrosome protein (speriolin) that binds CDC20 [131], and

perhaps SPATC1L plays a yet undetermined broader role in reproduction. The long-term health implications of altering methylation at these sites has yet to be determined.
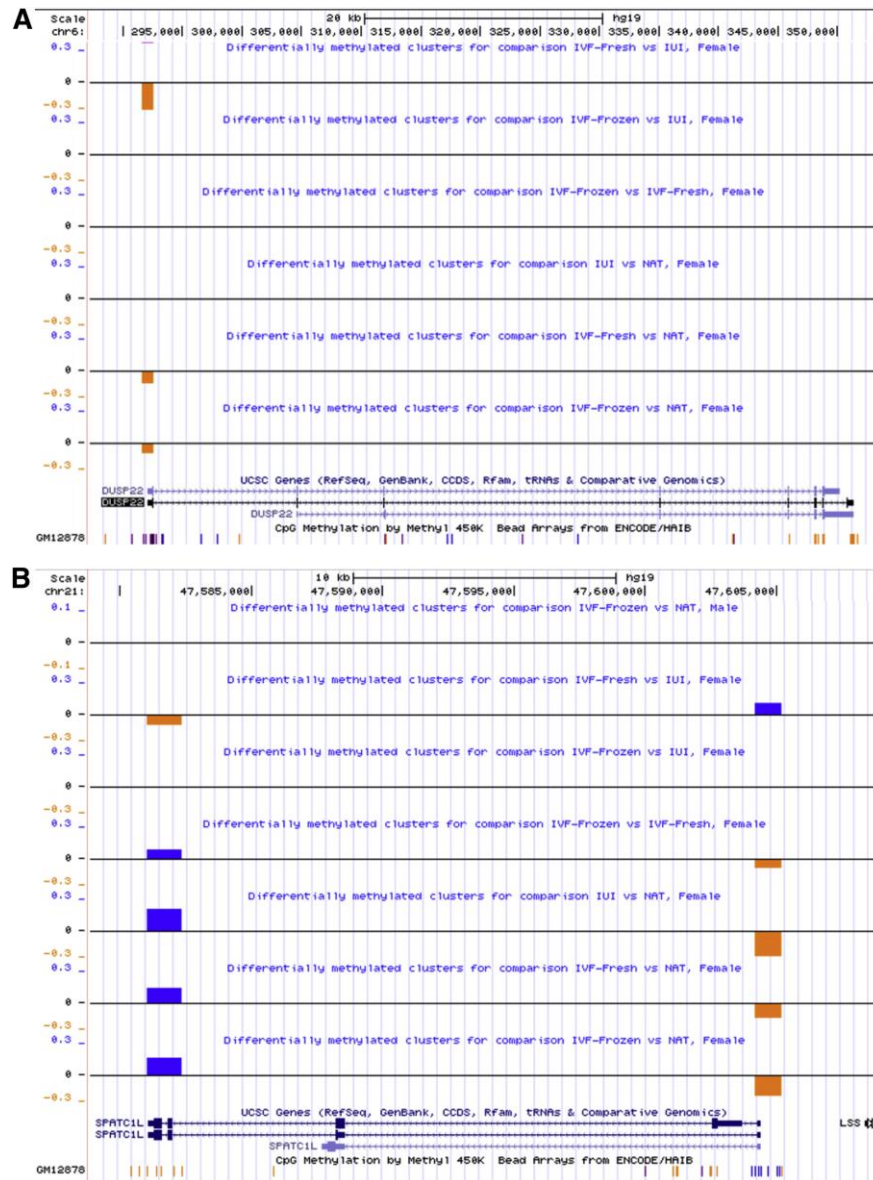
Figure 2.7. Metastable epialleles at DUSP22 and SPATC1L show considerable and concerted differential methylation. University of California Santa Cruz (UCSC) genome browser representation of female comparisons between conception groups for MEs located at (A) DUSP22 and (B) SPATC1L. Hypermethylated clusters and hypomethylated clusters are colored in blue and orange, respectively. Cluster heights represent the magnitude of methylation change, in scale with the y axis.

Table 2.3. Metastable epialleles are differentially methylated between conception types. Differentially methylated metastable epialleles were segregated into hypermethylated and hypomethylated loci. Genes associated with the respective epialleles are then indicated. The magnitude of the methylation change, in units of β-values, are shown in italics for each epiallele. FH = fresh embryo transfer; FZ = frozen embryo transfer groups; IUI = intrauterine insemination; ME = metastable epiallele; NAT = naturally conceived.

Conception comparisons	Hypermethylated MEs	Associated genes	Hypomethylated MEs	Associated genes
FH vs IUI, male	chr21:47604201-47604400 (*0.16*); chr21:47604801-47605000 (*0.16*); chr21:47605001-47605200 (*0.16*); chr7:4305001-4305200 (*0.04*);	*SPATC1L; SDK1*	chr2:113992801-113993000 (*−0.15*); chr2:113993001-113993200 (*−0.15*);	*PAX8; PAX8-AS1*
FH vs IUI, female	chr21:47604201-47604400 (*0.11*); chr21:47604801-47605000 (*0.11*); chr21:47605001-47605200 (*0.11*);	*SPATC1L*	chr14:54816001-54816200 (*−0.09*); chr19:29218001-29218200 (*−0.19*); chr2:113992801-113993000 (*−0.08*); chr2:113993001-113993200 (*−0.08*); chr5:135415601-135415800 (*−0.14*); chr5:135415801-135416000 (*−0.14*); chr5:135416201-135416400 (*−0.14*); chr6:291801-292000 (*−0.29*); chr6:292001-292200 (*−0.29*); chr6:292201-292400 (*−0.29*); chr6:292401-292600 (*−0.29*); chr7:4305001-4305200 (*−0.05*);	*LOC100420587; PAX8; PAX8-AS1; VTRNA2-1; DUSP22; SDK1*
FZ vs IUI, male	chr21:47604201-47604400 (*0.14*); chr21:47604801-47605000 (*0.14*); chr21:47605001-47605200 (*0.14*);	*SPATC1L*		
FZ vs IUI, female				
FZ vs FH, male	chr18:74514001-74514200 (*0.07*); chr2:113992801-113993000 (*0.18*); chr2:113993001-113993200 (*0.18*);	*LOC100131655; PAX8; PAX8-AS1*	chr1:19110801-19111000 (*−0.18*); chr2:128453201-128453400 (*−0.27*); chr2:128453401-128453600 (*−0.27*); chr6:291801-292000 (*−0.40*); chr6:292001-292200 (*−0.40*); chr6:292201-292400 (*−0.40*); chr6:292401-292600 (*−0.40*); chr7:4305001-4305200 (*−0.04*);	*SFT2D3; WDR33; DUSP22; SDK1*
FZ vs FH, female			chr2:128453201-128453400 (*−0.25*); chr2:128453401-128453600 (*−0.25*); chr21:47604201-47604400 (*−0.10*); chr21:47604801-47605000 (*−0.10*); chr21:47605001-47605200 (*−0.10*); chr4:1523001-1523200 (*−0.10*); chr5:135415601-135415800 (*−0.12*); chr5:135415801-135416000 (*−0.12*); chr5:135416201-135416400 (*−0.12*);	*SPATC1L; SFT2D3; WDR33; VTRNA2-1*
IUI vs NAT, male			chr21:47604201-47604400 (*−0.23*); chr21:47604801-47605000 (*−0.23*); chr21:47605001-47605200 (*−0.23*); chr7:4305001-4305200 (*−0.08*); chr2:128453201-128453400 (*−0.21*); chr2:128453401-128453600 (*−0.21*)	*SPATC1L; SDK1; SFT2D3; WDR33*
IUI vs NAT, female	chr4:1523001-1523200 (*0.10*)		chr21:47604201-47604400 (*−0.27*); chr21:47604801-47605000 (*−0.27*); chr21:47605001-47605200 (*−0.27*);	*SPATC1L*
FH vs NAT, male	chr4:1523001-1523200 (*0.11*)		chr2:128453201-128453400 (*−0.16*); chr2:128453401-128453600 (*−0.16*); chr6:291801-292000 (*−0.11*); chr6:292001-292200 (*−0.11*); chr6:292201-292400 (*−0.11*); chr6:292401-292600 (*−0.11*);	*SFT2D3; WDR33; DUSP22*
FH vs NAT, female	chr4:1523001-1523200 (*0.06*)		chr21:47604201-47604400 (*−0.16*); chr21:47604801-47605000 (*−0.16*); chr21:47605001-47605200 (*−0.16*); chr6:291801-292000 (*−0.14*); chr6:292001-292200 (*−0.14*); chr6:292201-292400 (*−0.14*); chr6:292401-292600 (*−0.14*);	*SPATC1L; DUSP22*
FZ vs NAT, male	chr4:1523001-1523200 (*0.10*)		chr10:135341601-135341800 (*−0.12*); chr1:19110801-19111000 (*−0.13*); chr2:128453201-128453400 (*−0.22*); chr2:128453401-128453600 (*−0.22*); chr6:291801-292000 (*−0.13*); chr6:292001-292200 (*−0.13*); chr6:292201-292400 (*−0.13*); chr6:292401-292600 (*−0.13*); chr7:4305001-4305200 (*−0.04*)	*CYP2E1; SPRN; SFT2D3; WDR33; DUSP22; SDK1*

Conception comparisons	Hypermethylated MEs	Associated genes	Hypomethylated MEs	Associated genes
FZ vs NAT, female			chr10:135341601-135341800 (*−0.10*); chr21:47604201-47604400 (*−0.22*); chr21:47604801-47605000 (*−0.22*); chr21:47605001-47605200 (*−0.22*); chr6:291801-292000 (*−0.11*); chr6:292001-292200 (*−0.11*); chr6:292201-292400 (*−0.11*); chr6:292401-292600 (*−0.11*);	*CYP2E1; SPRN; SPATC1L; DUSP22*
Female vs male, NAT	chr19:29218001-29218200 (*0.10*); chr2:113992801-113993000 (*0.08*); chr2:113993001-113993200 (*0.08*);	*LOC100420587; PAX8; PAX8-AS1;*		
Female vs male, FZ			chr4:1523001-1523200 (*−0.08*)	

Several adjacent MEs on chromosome 2 were observed to be hypomethylated solely in the male comparisons of ICSI (FH and FZ) groups to NAT. Within the same comparisons, these MEs were not consistently altered in the female cohort. These MEs are located in a potential regulatory region adjacent to SFT2D3, WDR33, and LIMS2. Additionally, these loci were hypomethylated in both FZ as compared with FH males and females. Taken together, this suggests that the type of IVF/ICSI treatment (i.e., cryopreserved or fresh embryo transfer) impacts methylation status to varying degrees, with FZ embryo transfer increasing the level of hypomethylation at this locus in the male cohort. A significant methylation change at this ME in males alone (FH vs. NAT, FZ vs. NAT) suggested a gender-specific effect on establishing the epigenome of this locus. Although Silver et al. [127] identify the top 10 MEs altered by season of conception in Gambian individuals, the DUSP22-, SPATC1L-, and SFT2D3-associated MEs found to be hypomethylated in ART individuals compared with NAT are not among them. Therefore, the differences in nutritional and periconceptional milieu between the rainy and dry season are likely different from those of unassisted and assisted conceptions.

v. Discussion

ART has been successfully applied to produce many phenotypically healthy human children. However, in both humans and model organisms, the impact of ART on the epigenome and long-term health of offspring is disputed. This confusion is compounded by a

lack of understanding of the epigenetic contribution to the health of ART conceptuses. To begin to disentangle the possible epigenetic impact of ART techniques, such as ICSI and cryopreservation, from that of underlying infertility, the present cohort included both NAT- and IUI-conceived newborns as comparative groups. This work verified previous literature indicating that ART does not induce extensive global changes in DNA methylation [103].

In contrast, numerous differentially methylated loci were identified when conception types were compared. The three assisted conception groups also exhibited differences from the natural conception group, thus supporting the view that subtle epigenetic changes occurred in children from parents who sought fertility treatment, relative to those conceived naturally. Consistent with this tenet, 4 of the 18 candidate CpG sites altered in placental tissue from children born to fertile parents compared with infertile parents [99] are located in or near significant clusters of two imprinted genes, growth factor receptor-bound protein 10 (GRB10) and necdin (NDN). While A-clustering recapitulated the hypomethylation of placental GRB10 CpG sites in bloodspots of ART children compared with fertile controls [99], it was not congruent with NDN. This discordance is likely reflective of tissue-specific DNA methylation in peripheral blood and placenta. Methylation differences between ICSI and naturally conceived controls were generally discordant with previous studies' conclusions on imprinted genes [95, 98]. This discordance is likely reflective of study differences in sample size and composition. Overall, a screen of imprinted DMRs does not identify altered DNA methylation profiles in placenta and cord blood samples in ART newborns compared with those spontaneously conceived [132]. This screen agreed with the conclusions drawn from the current cohort, in which the majority of imprinted genes in human do not show consistent or significant methylation changes.

The results of this work suggested that alterations in DNA methylation may explain at least a portion of the increased risk of birth defects in fresh embryo transfer, as well as the

increased birth defect risk in individuals with a history of infertility [89]. As we have shown, the IUI and FZ groups exhibit noted epigenetic similarity.

This is consistent with the recent suggestion that cryopreservation, in conjunction with ICSI, can reduce the risk of birth defects associated with ICSI births to that of the general population [89]. Interestingly, a Drosophila model of intergenerational metabolic reprogramming recently demonstrated a resetting of the fly sperm epigenome after heat shock [12]. It is tempting to suggest that cryopreservation in ART may yield a similar outcome (i.e., the epigenetic resetting of the human embryo). However, the true nature of the mechanism underlying the reduction in aberrant DNA methylation in the FZ newborns remains unknown. In all human studies to date, the combined effects of cryopreservation, embryo quality, uterine receptivity, and parental health have yet to be segregated. Moreover, the endometrial receptivity of women undergoing ovarian stimulation, which is often utilized in fresh embryo transfer, is likely shifted in comparison with that of naturally cycling women undergoing frozen embryo transfer [31]. This suggests that the relative endometrial receptivity may play a role in establishing the implantation environment and ultimately, the epigenome, of the conceptus.

Previous studies considering the impact of ART on the human conceptus have largely been limited in the number and type of genomic loci examined. Uniquely, the present work identified both regulatory regions (i.e., enhancers) and MEs as being affected by infertility and ART. The 450k array locus-specific β-values for various MEs, including PAX8, SPATC1L, and VTRNA2-1 of **Table 2.1**, through comparison to quantitative bisulfite pyrosequencing of MEs, are known to provide a reliable proxy for methylation status [127]. Altered methylation at enhancers, and more broadly, regulatory regions has the potential to impact gene expression and long-term health status. Modulation of MEs was consistent with previous nutritional studies [79, 127], thus indicating that the periconceptional environment of the early human embryo, which includes IVF culture conditions for the IVF and ICSI conceptus, leaves a lasting impression on the epigenome.

Compared with unassisted pregnancies in the general population, ART pregnancies are susceptible to multifetal gestation and adverse pregnancy complications. One must therefore consider the clinical characteristics that may impact DNA methylation and blood cell populations in neonatal studies. These include gestational age, multifetal pregnancies, pregnancy complications, fetal–maternal characteristics, racial characteristics, and sociodemographic status, sample characteristics to which this study was blinded. Nevertheless, although gestational age ranged from 30 to 40 weeks, approximately 80% reached full term; 89.5% were singletons. Accordingly, the majority of each conception group was comprised of full-term, singleton pregnancies.

With current available approaches, the veracity of bloodspot composition estimates from newborn bloodspots is unproven [133]. However, methylation at MEs is known to be established in the early embryo and remain consistent across diverse tissues ranging from blood to hair follicles [79, 127]. Additionally, previous DNA methylation studies suggest that although leukocytes are composed of approximately 54% neutrophils, the remaining approximately 19 different cell types do not affect the majority of loci examined [134]. In the present cohort, leukocyte composition was estimated using the method of Jaffe and Irizarry [135], and the resulting blood cell proportion estimates do not significantly influence the 450k methylation profiles. Hence, any differences in leukocyte composition do not seem to underlie the group differences in ME methylation.

This work serves as a resource for future studies on IVF populations. It provided the first large-scale 450k dataset on cytosine methylation in newborns conceived through three different assisted conception procedures: IUI, ICSI–fresh embryo transfer, and ICSI–cryopreserved embryo transfer, and provided comparison to a naturally conceived cohort. The association of methylation changes in these loci with ART-conceived child's clinical phenotypes and long-term health can now be investigated. Identification of differentially methylated genes between comparisons of conception types suggests that epigenetic

mechanisms may at least partly underlie the observed risks of neurologic and birth defects seen in the ART population and assisted births [89, 90]. The reduction in birth defects from after cryopreserved embryo transfer [89] is suggested here to have an epigenetic basis, given that FZ groups exhibited reduced epigenetic changes. This tenant biologically supports the use of frozen embryo transfers rather than fresh embryo transfers in association with IVF and ICSI.

Chapter 3

"DEFINING THE SPERM TRANSCRIPTOME: RNA ELEMENT DISCOVERY FROM GERM

CELL TO BLASTOCYST"

This chapter was adapted from the following publication:

Molly S Estill, Russ Hauser, Stephen A Krawetz. "RNA element discovery from germ cell to blastocyst". Nucleic Acids Research. 2018 Dec 21. Cover feature. doi: 10.1093/nar/gky1223.

i. Summary

Recent studies have shown that tissue-specific transcriptomes contain multiple types of RNAs that are transcribed from intronic and intergenic sequences. The current study presents a tool for the discovery of transcribed, unannotated sequence elements from RNA-seq libraries. This RNA Element (RE) discovery algorithm (REDa) was applied to a spectrum of tissues and cells representing germline, embryonic, and somatic tissues and examined as a function of differentiation through the first set of cell divisions of human development. This highlighted extensive transcription throughout the genome, yielding previously unidentified human spermatogenic RNAs. Both exonic and novel X-chromosome REs were subject to robust meiotic sex chromosome inactivation, although an extensive de-repression occurred in the post-meiotic stages of spermatogenesis. Surprisingly, 2.4% of the 10,395 X chromosome exonic REs were present in mature sperm. Transcribed genomic repetitive sequences, including simple centromeric repeats, HERVE, and HSAT1, were also shown to be associated with RE expression during spermatogenesis. These results suggest that pervasive intergenic repetitive sequence expression during human spermatogenesis may play a role in regulating chromatin dynamics. Repetitive REs switching repeat classes during differentiation upon fertilization and embryonic genome activation was evident.

ii. Introduction

Expression profiles of known RNAs have been catalogued for a range of cell types, with the use of expression arrays and, more recently though RNA deep-sequencing studies. This has yielded a series of useful databases including GTEx (https://www.gtexportal.org/home/), EMBL-EBI's Expression Atlas (https://www.ebi.ac.uk/gxa/home/), The Human Protein Atlas (https://www.proteinatlas.org/), and ENCODE (www.encodeproject.org) [135-140]. These databases and RNA-seq studies generally focus on annotated genes and transcript variants that are derived from transcript modeling programs such as Cufflinks [141] and are provided as part of the Refseq and Gencode annotation [142, 143].

Both coding and non-coding RNAs play major roles in all cellular processes. In addition to protein-coding RNAs, at present, there are 48 different non-coding and pseudogene classes of RNA documented in the version 27 annotation of the Human Gencode. Approximately 40% of the annotated genes in Gencode correspond to long and short non-coding RNA genes [144]. Non-coding intergenic regions are known to contain regulatory RNAs. These include long intergenic non-protein coding RNA (lincRNA), enhancer RNA (eRNA), piwi-interacting RNA (piRNA) and circular RNAs, with others just beginning to be described [145-148]. The human transcriptome is likely to be more complex than even these annotations indicate, as an estimated three quarters of the human genome is transcribed [149]. This would include novel tissue-specific RNAs, whose roles remain to be established [150].

The palette of RNAs appear enriched in certain specific tissues, with each providing a specialized function, e.g., brain - cognitive and functional system level control, and germline – stem cell – defining development [151-153]. Their corresponding complexity is exemplified in the testis by the collection of unique structural and functional spermatozoal-specific transcript variants [154] that are observed during maturation, as sperm assume their unique

shape. This culminates with the compaction of the sperm nucleus to a transcriptionally and translationally inert structure. The latter is ensured by fragmenting rRNAs [155], as well as others and completes with the expulsion of the majority of the cytoplasm. In addition to the paternal genome and sperm encapsulated RNAs [156], RNA/proteins and other molecules from distant tissues acquired during epidydimal transit [71, 157] are delivered at fertilization. This provides a pathway for soma-to-germline transmission [156, 158, 159] that perhaps conveys signals echoing how other tissues have responded to the environment (reviewed in [160]).

Previous literature has shown that unannotated transcripts corresponding to intronic and intergenic regions of the spermatogenic genome are comparatively abundant in human sperm [154, 161-163]. They vary amongst species and in response to and can provide markers of disease [163-165]. These observations drove the development of an algorithm to systematically identify the genomic locations of RNAs, defined as RNA elements (RE)., i.e., regions transcribed throughout the genome. This unbiased analysis tool is not limited to those RNAs currently defined in the databases as it does not seek to generate gene structures from REs. It is compatible with a range of Next Generation Sequencing (NGS) platforms, RNAs from varied sources, abundance, quality, and levels of fragmentation, i.e., FFPE-like RNAs. The RNA Element Discovery algorithm (REDa) approach only requires the BAM file of genomic alignments to detect transcribed regions of novel loci in conjunction with well-known annotated loci.

RE discovery was applied from the perspective of the human male germ cell to blastocyst paradigm. A series of spermatogenesis and embryogenesis pattern specific intergenic human REs were identified, indicating that the transcriptome extends well-beyond the annotated genes, including those delivered at fertilization. Tissue-specific REs comprised of intronic and intergenic REs were uncovered and, in some cases, exon boundaries extended. Transcribed genomic repetitive sequences, such as simple repeats, HERVE, and

HSAT1, were shown to be associated with RE expression during spermatogenesis, and may play a developmental stage specific role. Similarly, in the human embryo, MER73 was associated with RE transcription at the minor wave of zygotic genome activation and MLT2A1 and SVA-D expressed through the major wave during the transition to the embryonic genome. This provided a deeper understanding of the dynamic transcriptome of human sperm, as well as uncovering the possible role of specific repetitive sequences in the spermatogenesis.

iii. Materials And Methods

RE discovery

The current study used Gencode release 26 (for GRCh38) and the GRCh38 genome for RE discovery, which is detailed in **Appendix E**. RNA-seq samples used in RE discovery are described in **Appendix F**. Sample reads were pre-processed prior to RE discovery with Trimmomatic version 0.36, trimming Illumina adaptors and poly(A$^+$) sequences, where appropriate, with parameters "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15". Reads were aligned to the GRCh38 genome using HISAT2 (version 2.0.6), using the parameters "-p10 -max-seeds 30 -k 2". Read coverage was provided to the RE discovery tool in bigwig format, generated by converting BAM files to bedgraph format, using the bedtools tool genomeCoverageBed, with the parameters "-split –bg", and subsequently bigwig format, using the bedGraphToBigWig program (available from the UCSC Genome Browser utilities). The threshold parameter μ for RE discovery was set to 2.5 reads per million, to minimize their contribution of background noise. Novel REs from each study were combined using custom R commands, which merged overlapping novel REs, re-annotated the merged REs, and added the merged REs to the exonic REs, to produce a collective set of REs. The collective set of REs for the different samples was subsequently used in all analyses. For comparison of RE discovery to established transcript-building software, Cufflinks (v2.2.1) and Stringtie (v1.3.4) were used on the same pre-processed reads previously used for RE discovery [141,

166]. Default parameters for both Cufflinks and Stringtie were employed, using Gencode release 26 (for GRCh38) as the reference annotation.

Differential expression (LMEM, fold change, LM)

A linear mixed-effects model (LMEM) was used to calculate differential expression between poly(A⁺) and total RNA libraries from oocyte through early embryonic development [167-169]. The LMEM was used with a random slope and intercept for each cell type, to consider heterogeneity across cell types [formula= *RPKM ~ RNA.type + (1 + RNA.type | Tissue)*]. Residuals of randomly selected REs were analyzed for homoscedasticity, ensuring that the assumptions of the LMEM were satisfied. Multiple testing correction was applied to P-values for resultant slopes, using Benjamini-Hochberg correction [170].

Differential expression of poly(A⁺) and total RNA libraries in sperm and testis tissue was determined using a fold-change (Fold change= $log2\left(\frac{mean(Total\ RNA)}{mean(poly(A^+))}\right)$). The use of an expression ratio, rather than linear modeling, was necessary due to the technical differences between the total RNA sperm samples [163] and the three poly(A⁺) sperm libraries [171], as well as the absence of multiple independent total RNA testis samples [154].

RE enrichment for repetitive sequences

In cases when median RE RPKM in spermatozoa exceeded 1 RPKM (thus removing REs with low coverage in most samples), peak RE RPKM was 25 RPKM and subsequently used as an expression threshold. REs were first assigned as "Expressed" if the median RPKM for the cell/tissue type was greater than 25 RPKM. The enrichment or depletion of repetitive sequences in the expressed REs was calculated using UCSC's Repeatmasker track (for GRCh38), a hypergeometric test and custom R code. The proportion of each genomic repeat in all available REs was used as input probability, with the number of expressed REs for the given cell type used as the sample size. The probability of drawing the actual number of expressed REs overlapping the given repeat type was adjusted using a Bonferroni correction

[172]. To identify repeats of interest, significantly enriched or depleted repeats were additionally filtered to remove repeats with minimal over- or under-enrichment. Thus, only repeats whose difference between the expected and observed RE count was greater than 10 REs were retained.

Expression clustering

Expression patterns across spermatogenic cell types were identified using the R package Mfuzz [163, 173-175]. Mfuzz is designed for soft clustering of gene expression time-series data. Soft clustering is a form of clustering where a data point (e.g. transcript) can belong to more than one pattern. The samples used in clustering were the Jan et al. spermatogenesis libraries [175], as well as a set of 7 ejaculated sperm samples from fertile males [163]. It should be noted that these 7 samples are a subset of the 52 "Control" sperm samples used for expression comparisons in Chapter 4. The median expression value for the 7 mature sperm samples was used to represent the mature sperm samples as a single value. The RE dataset was then composed of a single library for A_{dark} SSCs, A_{pale} SSCs, Leptotyne/Zygotene, Early Pachytene, Late Pachytene, and Round spermatids, while the library for ejaculated sperm was the median expression value for 7 fertile males. In a step intended to remove universally lowly expressed REs, REs were processed to remove those which did not exceed 25 RPKM in at least one sample. Mfuzz clustering was performed, generating 20 cluster patterns, with a minimal membership of 0.7 required for inclusion of an RE in a pattern.

Paternal/maternal transmission

REs were assigned as maternally transmitted to the zygote with median zygotic level > 10 RPKM, sperm < 2 RPKM, and oocyte > 25 RPKM. REs were assigned as paternally transmitted with moderate confidence with median zygotic level > 10 RPKM, sperm > 25 RPKM, and oocyte < 5 RPKM. REs were assigned as paternally transmitted with greatest confidence with median zygotic level > 10 RPKM, sperm > 25 RPKM, and oocyte < 2 RPKM.

FDR calculation for GTEx and PPV calculation for sperm

The accuracy of the RE discovery algorithm to identify expressed loci was calculated using the Jodar et al. dataset, which consisted of 7 fertile human sperm samples, prepared using total RNA [163]. RE discovery was performed on the 7 samples, at a range of μ from 1 to 10 RPM, at increments of 0.5 RPM. The RPKM of the resulting novel REs for each sample was calculated, along with the median RPKM across the 7 samples. Experimental thresholds for calling a RE as "expressed" ranged from 1 to 200 RPKM, at increments of 1 RPKM. At each expression threshold, the number of REs with a median RPKM at or exceeding the threshold were recorded. The Positive Predictive Value (PPV) at each expression threshold and μ was calculated as

$$PPV = \frac{Novel\ REs > Expression\ threshold}{Novel\ REs > Expression\ threshold + Novel\ REs \leq Expression\ Threshold}$$

The ability of the RE approach to recapitulate tissue expression in the established databases was determined using the testis expression in the GTEx database [138]. The median TPM for all GTEx testis samples was downloaded and was processed to replace duplicated common gene names with the mean TPM for all instances of the gene name. Only gene names found in both GTEx and exonic REs were used in subsequent intersect analysis. The unique gene names with expression exceeding 5 TPM were compared to those of the exonic REs exceeding 25 RPKM.

Gene ontology

Ontological analysis was performed with the Genomatix software suite (https://www.genomatix.de/), version 3.10. The GeneRanker function (using Genomatix Eldorado version 12-2017) generated the ontological enrichment of signaling pathways.

iv. Results and Discussion

RE identification and classification

The RE discovery algorithm was designed to detect expressed regions of the genome using RNA-seq, regardless of the sequencing platform or read structure. A detailed description of RE discovery is presented in **Appendix E**, and the corresponding code is provided online (https://github.com/mestill7/RE_discovery). Briefly, the known gene annotation (e.g., RefSeq, Ensembl, Gencode) for the genome of interest is parsed into individual exon locations. In the current study, Gencode release 26 (GRCh38) was used, with non-coding entries considered as annotated "exons" [144]. As summarized in **Figure 3.1A**, RE discovery first requires the sequenced reads to be processed, e.g., adaptors trimmed and low-quality bases removed, prior to alignment to the genome of interest. For the unannotated regions of the genome, the mean read coverage was calculated for each 10 bp genomic segment and the 10 bp segments with sufficient read coverage, determined by a threshold μ, retained. For the purposes of this study, μ = 2.5 reads per million provided well-balanced signal to noise ratio (**Figure 3.2**) that was suited for RNA libraries generated from low-input, potentially fragmented RNAs, as is often found in clinical Formalin-Fixed Paraffin-Embedded (FFPE) samples and spermatozoa [154, 176]. The overlapping 10 bp regions were subsequently merged to yield the final novel REs for each collection of samples studied. The merging steps allow for a maximum of 150 bp between element bins, intended to allow for gaps in coverage caused by sequencing bias and/or biological fragmentation.
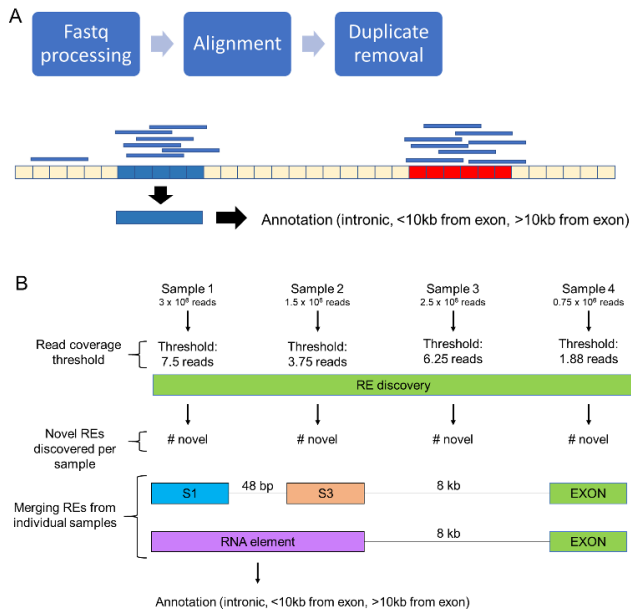
Figure 3.1. Pre-processing for RE discovery. (A) RNA-seq reads, in fastq format are processed to remove low-quality bases and adaptor sequences. The trimmed reads are then aligned to the genome, and the duplicate alignments removed. Read coverage is then used to identify 10 bp segments with read coverage surpassing μ. The expressed 10 bp segments (shown in blue) are merged and annotated according to their adjacency to exons (shown in red). (B) RE discovery workflow for four theoretical RNA-seq samples. Each sample has a different library size, and correspondingly, different read coverage thresholds at a μ of 2.5 Reads per million (RPM). Non-exonic regions of read coverage surpassing the assigned threshold are deemed "Novel REs". Merging novel REs from the four different samples, yields two novel REs, one from Sample 1 (S1) and one from Sample 3 (S3) that are separated by up to 150 bp. Novel REs of the different samples S1 and S3 are merged into a final RNA element, represented in purple. Exonic REs are excluded from this merging step. The final novel RE set for the four samples is then annotated as intronic, near-exon, purple, (<10kb from exon), and orphan REs (>10kb from exon).
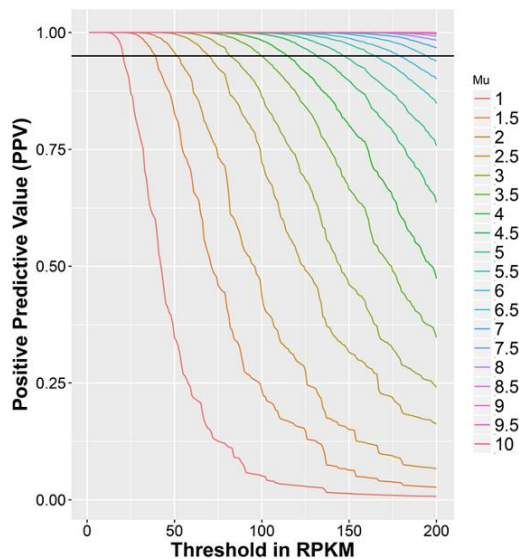
Figure 3.2. Background noise for read coverage thresholds. X-axis represents the experimental threshold for calling a RE as "true positive (TP)". The corresponding Positive Predictive Value (PPV) is calculated as (TP/(TP+FP)). The PPV curve is provided for levels of μ from 1 RPM to 10 RPM. The X-axis represents the expression threshold required for assigning a RE as expressed, and ranges from 1 to 200 RPKM. A PPV of 0.95 (corresponding to a False Discovery Rate (FDR) of 5%), is shown as a black line.

Novel REs were then annotated according to their genomic position, relative to known exons (**Figure 3.1B**). "intronic" REs were located within introns, while any non-intronic REs located within 10 kb of an annotated exon were designated "near-exon" REs; "orphan" REs were at a distance greater than 10 kb from any known exon. An exonic RE was extended into a near-exon RE if they were within 50 bp and the difference in read coverage was < 50%. As summarized in **Figure 3.3**, previously published RNA-seq studies representative of human spermatogenesis, mature sperm, oocyte, embryonic stages, and liver samples, detailed in **Appendix F**, were subject to RE discovery. This set of RNA-seq libraries encompassed both poly(A^+) selected and total RNA preparations. A database of REs across the different tissue and types was created by merging the novel REs from each study with the exonic and non-coding transcript REs and used in all subsequent analyses.

A specific set of RE discovery parameters (μ = 2.5 reads per million and minimum required distances between genomic bins), were used to provide an acceptable signal to noise ratio in the sperm RNA libraries. Due to the exploratory nature of this study, maximizing the number of REs, while still excluding spurious signals, was of particular interest. Modification of the parameters used in the REDa method will allow increased detection stringency (thus reducing the number of REs detected) or more permissive RE detection (thus increasing the number of REs detected). Increased stringency during RE detection can easily be achieved by increasing the required μ (thus increasing the threshold required to mark genomic bins as "expressed") or by increasing the minimum proportion of samples with the required minimum read coverage to retain a genomic bin for RE designation.

The accuracy of the RE approach, which separates exons into individual units, rather than linking exons into a whole transcript, was tested by comparing expressed REs in testis libraries to the testis expression levels given by GTEx. At least 91% of gene names associated with testis-expressed exonic REs overlap with gene names expressed in GTEx testis tissue,

suggesting that the RE approach can recapitulate the patterns designated in established expression databases.
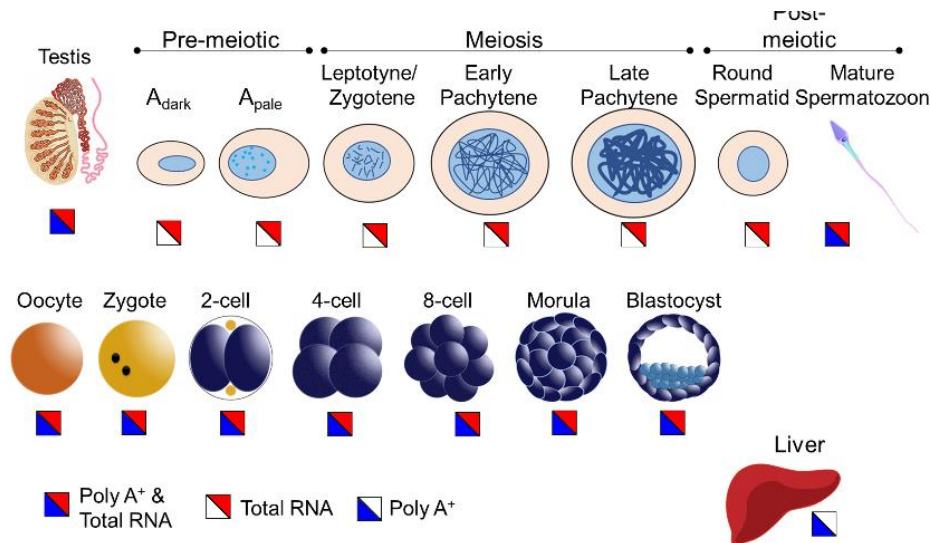
Figure 3.3. Tissue types used for RE discovery. The male germline within the testis tissue is divided into and represented by seven stages of spermatogenesis. The female germline is represented by a single-cell oocyte with embryonic stages that range from zygote to blastocyst. Somatic tissue is represented by the liver sample. Total RNA or poly(A+) enriched RNA-seq libraries are indicated in split squares, with blue representing poly(A+) selected samples, red indicating total RNA samples, and a split blue/red square as both library preparations.

The above RE identification method was developed to ensure accuracy in face of extensive RNA fragmentation, naturally occurring in human sperm. Certain tissue preparations, such as FFPE, also yield compromised RNA preparations. Given that several established transcript-building algorithms are readily available, I compared both Stringtie (v1.3.4) and Cufflinks (v2.2.1) to the RE approach for two random sperm samples and two male human cell lines. RNA-seq datasets from human cell lines, i.e., SRR020288 (h1 hESC) and SRR3192556 (OCI-LY7, derived from a B cell lymphoma), provided independent datasets when testing the RE method. Using minimal thresholds of expression (>10 RPKM in REs, >1 FPKM in Cufflinks and Stringtie), the majority of expressed REs overlap locations of transcripts generated using transcript-building software. Across the 4 samples, 67%-92% of

"expressed" REs overlap Stringtie results, and 81%-90% overlap Cufflinks results at the above thresholds of expression. Notably, regardless of the transcript-building method and required expression thresholds, a majority of REs (complete range 21%-93%) lacking overlaps with Cufflinks and Stringtie results are exonic REs, suggesting that the established transcript-building methods are less than ideal for fragmented or unevenly covered transcripts. Additionally, the presence of spliced reads, a critical component to transcript-building, is reduced in spermatozoal RNAs (36% - 40%) compared to RNAs from cell lines (41%-64%). Other investigators have also employed a targeted Cufflinks [168] discovery approach to identify novel linear embryo transcripts. Reflective of the low level of expression and rigor required for identification, the majority of these linear transcripts were not discovered using the RE strategy (data not shown).

With the function of the novel REs being unknown, I hypothesized that the novel REs may have regulatory roles. To assess this, REs were overlapped with a series of epigenetic marks and regulatory genomic sequences (**Figure 3.4**). For regulatory chromatin marks (proximity to DNase hypersensitive regions, proximity to CTCF binding sites, proximity to Topologically Associating Domains (TADs), and proximity to ENCODE Transcription Factor Binding Sites (TFBS)) [177-182], the novel RE classes largely showed a similar overlap proportion as exonic REs. All RE classes showed very little overlap with piRNA clusters [183, 184]. TADs mark the physical interaction of genomic regions and the minimal overlap of all RE classes with TAD boundaries suggests that novel REs are not primarily involved in establishing TADs. However, it is important to note that the compilation of REs are detected in at least one of a variety of tissues (across spermatogenesis, embryogenesis, and somatic tissue). Therefore, the overlaps summarized in **Figure 3.4** do not preclude a tissue-specific relationship of REs with a given epigenetic mark, such as TADs. Notably, all classes of novel REs had a high overlap (>50%) with repetitive sequences (UCSC's Repeatmasker track (for GRCh38) [185], compared to the approximately 22% overlap in exonic REs.
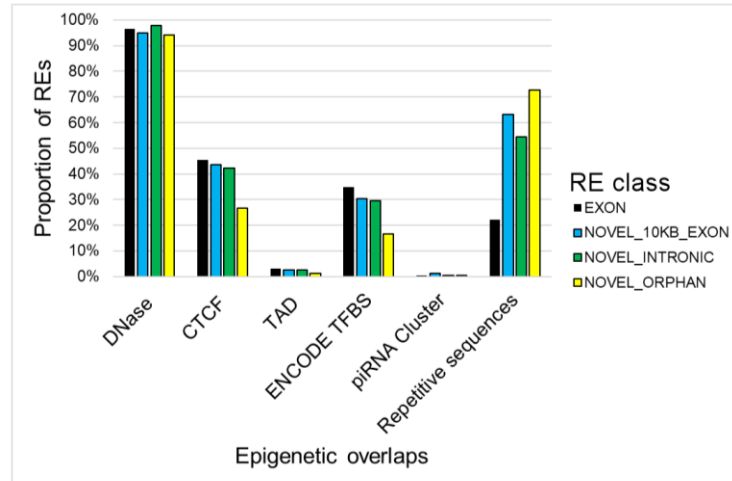
Figure 3.4. Overlap of REs with epigenetic marks and regulatory genomic sequences. The proportion of each RE class classified as overlapping a given epigenetic mark or genomic sequence is indicated on the Y-axis, with the type of epigenetic mark or genomic sequence indicated on the X-axis. The type of RE class is indicated with exonic REs in black, near-exon REs in light blue, intronic REs in green, and orphan REs in yellow. "DNase" indicates overlaps within DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types) or 5 kb of a cluster. "CTCF" indicates overlaps within CTCF binding sites in the GM12878 cell line (determined from ChIP-seq) or 5 kb of a binding site. "TAD" indicates overlaps within 5 kb of a topologically associating domain (TAD) ending site. "ENCODE TFBS" indicates overlaps with Transcription Factor Binding Sites (TFBS) from Encode (ENCODE Mar 2012 Freeze). "piRNA Cluster" indicates overlaps with piRNAs in the human genome. "Repetitive sequences" indicates overlaps with UCSC's Repeatmasker track (last updated 2014-01-10).

RNA-seq complexity of RNA elements

The RE discovery algorithm was developed to identify transcribed intergenic loci from RNA-seq data. Many novel loci (e.g., near-exon and orphan REs) were hypothesized to be derived from non-polyadenylated RNAs, since this class appears underrepresented in the genome and the GENCODE annotations. A series of poly(A+) selected and Total RNAs from a range of cell types that capture fertilization to early embryonic development from oocyte, zygote, 2-cell embryo, 4-cell embryo, 8-cell embryo, and morula (**Figure 3.3**) and the male germline, through ejaculated sperm and testis, were examined [167-169]. Applying a linear mixed-effects model (LMEM) to Total and poly(A+)-selected RNAs from the human oocyte and various stages of early embryonic development, revealed a comparatively lower number of REs detected within the poly(A+) selected fraction (**Figure 3.5**). In general, the number of novel REs that were either increased or depleted by poly(A+)-enrichment do not markedly

differ (**Figure 3.5B**). Interestingly, the number of orphan REs approximately doubled upon poly(A$^+$)-enrichment as compared to the Total RNA fractionation. This suggests that a population of orphan REs belong to a larger, yet unknown set of polyadenylated transcripts. To determine if poly(A$^+$) enrichment of orphan REs reflected a genomic repeat, the distribution of repeats within the 150 poly(A$^+$) enriched orphan REs was assessed and is shown in **Figure 3.5C**. Within the 129 orphan REs that contain a repetitive element, the majority were LTRs and SINEs. It is worth noting that 40 of the 55 LTR-containing REs were ERVL-MaLRs. This is a non-autonomous LTR-retrotransposon element derived from ERV [186, 187] that may function in regulating gene expression during the oocyte-to-embryo transition [188].
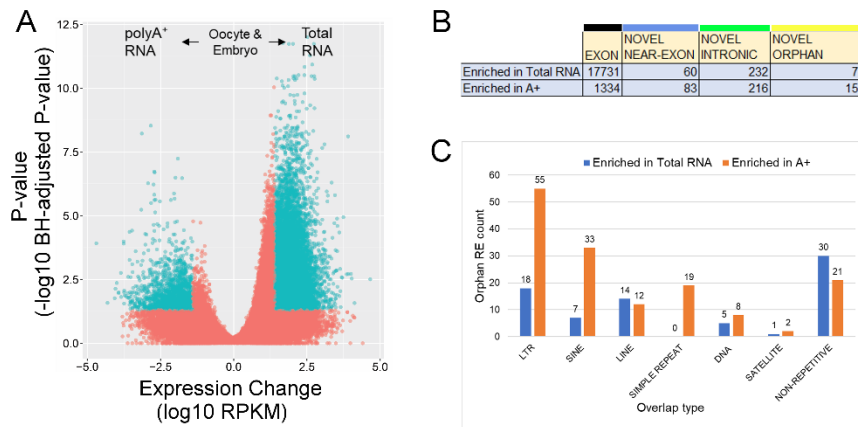
Figure 3.5. Orphan REs are enriched in poly(A+) samples. (A) Volcano plot of slope changes in REs from LMEM in oocyte and embryo, with the X-axis representing slope change in log10-transformed RPKM, and the Y-axis representing the Benjamini-Hochberg-adjusted P-value as a negative log10-transformed P-value. Positive slope and negative slope indicate increased abundance in total RNA and poly(A$^+$) preparations, respectively. Each point represents a single RE, with blue points indicating statistically significant REs (adjusted P-value <0.05) with absolute slope changes exceeding 25 RPKM. (B) Distribution of REs enriched in either total RNA or poly(A$^+$) libraries, according to the annotation class. (C) The distribution of orphan REs enriched in either total RNA or poly(A$^+$) libraries, according to repeat class.

The effect of poly(A$^+$) enrichment was also assessed individually for human sperm and testis samples, providing the other half of the equation to early post-fertilization development. Poly(A$^+$) enrichment has contrasting effects on exonic REs in spermatozoa and testis, with poly(A$^+$) enrichment depleting and enriching exonic REs in sperm and testis, respectively.

However, unlike embryos, novel REs were markedly enriched in sperm and testis total RNA libraries, reflective of the relatively uncharacterized state of this cell type (**Figure 3.6**). Although poly(A$^+$) enrichment does effectively reduce RNA library complexity, it does not appear to select for RNAs of any given biological function or pathway, with many GO terms shared in both the poly(A$^+$)-enriched and Total RNA-enriched gene sets of the human embryo. This reiterates previous studies that indicate a reduction of transcript diversity and exon expression upon poly(A+) enrichment [189].

The number of human zygotic LTR and SINE-associated REs that may be derived from poly(A+) intergenic transcripts is of note. In accord with the data of others [190-193], this could afford transcript stabilization and nuclear export [194] perhaps increasing their retention in a given cell of the dividing embryo. Notably, at least in mouse, the transcription of retrotransposon-derived RNAs is thought to impact chromatin accessibility, and thus embryonic development [195].
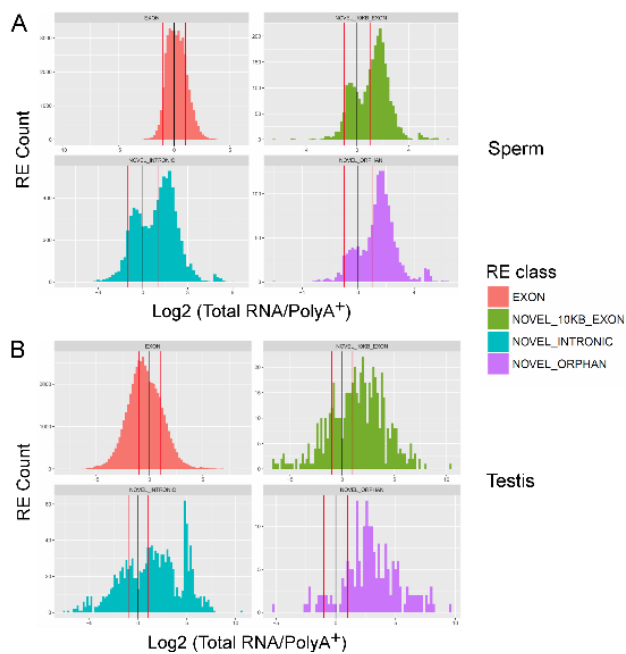
Figure 3.6. Total RNA libraries are enriched for Novel REs in sperm and testes. (A) Histogram of fold changes (log2 transformed) are shown for (A) mature sperm and (B) testes samples. The X-axis indicates the Log2 of the Total RNA/polyA$^+$ ratio, with positive change and negative change representing enrichment in total RNA and poly(A$^+$) libraries, respectively. Note the novel REs all exhibit a shift to the right, indicating increased expression in total RNA samples.

RNA element expression across spermatogenesis

RE expression throughout spermatogenesis was examined as a comparison to previously published patterns of whole transcript expression during the spermatogenic cycle[175]. The spermatogenic stages encompassed 6 cell types (Spermatogenic Stem Cells (SSCs) through Round Spermatids), isolated using laser capture microdissection [175]. Clustering of the various REs expression patterns across spermatogenesis was initially performed using Mfuzz, [173, 174] with the published 6 cell types [175]. RE expression across spermatogenesis recapitulated those patterns previously observed using whole transcripts. To extend the analysis to the final stage of spermiogenesis, RNA-seq from ejaculated sperm datasets from fertile males [163] were included (**Figure 3.3**). The addition of mature sperm enabled the discovery of several patterns specific to early round spermatids and maturing round spermatids, as observed through mature spermatozoa (**Figure 3.7A-D**). The final stages of spermatogenesis involve a burst of transcription, as well as the formation (and eventual loss) of the residual body as the majority of the cytoplasm is expunged from the cell. The burst of transcription in round spermatids was observed as a general increase in transcription of exonic REs that include 34,226 REs found in round spermatids but not in the late pachytene stage spermatocytes. Interestingly, a large portion of spermatid and/or mature sperm-specific clusters were generated from novel REs, suggesting that intergenic and intronic REs play a substantial role in the final stages of spermatogenesis that forms each spermatozoon as summarized in **Table 3.1**. To verify these observations, expressed (median expression >25 RPKM) REs for each spermatogenic stage were partitioned according to RE class (**Figure 3.7E**). The vast majority of REs expressed in pre-meiotic and meiotic stages were exonic (85%±7%). This was followed by a notable increase in the number of novel REs in Round Spermatids and Spermatozoa. The contribution of novel REs to the total transcriptome rose to 47% in mature sperm.
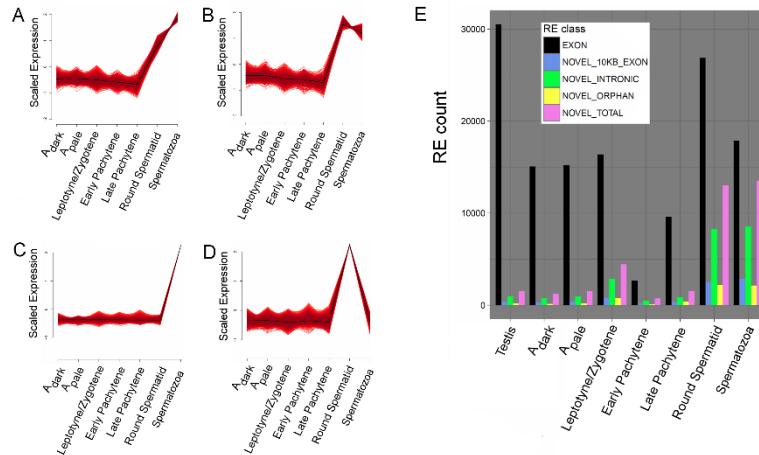
Figure 3.7. Mfuzz clusters highlighting the round spermatid to spermatozoon transition. (A-D) Clusters with increased expression in round spermatids and/or mature spermatozoa. (E) RNA element abundance as a function of annotation class and cell type with median RPKM >25.

Table 3.1. RE class distribution of Mfuzz clusters. The "Pattern" column indicates the expression pattern across spermatogenesis. The count of exonic REs, near-exon REs, intronic REs, and orphan REs belonging to the given pattern are indicated in columns "EXON", "NOVEL_10KB_EXON", "NOVEL_INTRONIC", and "NOVEL_ORPHAN", respectively.

PATTERN	EXON	NOVEL_10KB_EXON	NOVEL_INTRONIC	NOVEL_ORPHAN
High Round spermatids and Spermatozoa	1992	513	1333	286
High Round spermatids	5502	1325	5019	1453
High Spermatozoa	5012	4157	11874	4458
High Leptotene/Zygotene and Round spermatids	1607	59	167	27
High Leptotene/Zygotene	3403	345	1727	333
High Adark and Apale	618	3	11	2
High Adark	1198	23	105	18
High Adark, Apale and Late Pachytene	513	29	32	11
High Apale	996	16	64	10
High Leptotene/Zygotene, Low Round spermatids & Spermatozoa	519	28	33	12

Ontological analysis of the exonic and novel REs (with the exception of orphan REs) showed that the most abundant REs in round spermatids were enriched for genes involved in organelle biogenesis and maintenance. This is in accord with the physiological changes

occurring during spermiogenesis. REs that are abundant in both round spermatids and spermatozoa were enriched for TNF-alpha signaling, associated with maintaining a homeostatic state [196, 197]. The TNF-alpha signaling-associated REs enriched throughout the post-meiotic phase of spermatogenesis may be another part of a surveillance mechanism to ensure an optimal contribution [165]. REs that were primarily abundant in spermatozoa were associated with a range of signaling pathways, such as Glutamate Receptor signaling, WNT signaling, NGF signaling, EGFR1, and Signaling by Rho GTPases. WNT signaling has several roles in spermatogenesis, from maintenance to maturation, and thus motility [198-200]. The role of NGF signaling in spermatogenesis in humans is unclear but has been implicated in mammalian Sertoli-germ cell signaling, sperm motility, and the acrosome reaction [201, 202]. Sperm EGFR activation is a major driver of sperm capacitation [203, 204], while Rho GTPases are likely to aid as mediators of the acrosome reaction [205]. Odorant receptors may be required for sperm chemotaxis in mammals [171], while glutamate receptors may also be involved in capacitation and/or sperm chemotaxis [206, 207] although such functions have yet to be demonstrated in mammalian systems.

Sex-chromosome expression during spermatogenesis

Meiotic sex chromosome inactivation (MSCI), the process by which genes located on the X-chromosome are repressed during meiosis, is essential for successful meiosis during human spermatogenesis [208]. However, abundant evidence suggests that numerous X-linked genes escape post-meiotic X chromosome silencing (PMCI), a process that may be less effective in humans than other species [209, 210]. In comparison, most classes of Y-linked REs undergo silencing during MSCI, with the exception of Y-linked orphan REs that are present throughout spermatogenesis.

As shown in **Figure 3.8**, repression of exonic X-linked REs during spermatogenic MSCI is evident. This is followed by de-repression of X-linked exonic and novel REs, that return to pre-meiotic levels in mature sperm (**Figure 3.8A**). Notably, several X-linked REs

were intensely expressed (at a threshold of 25 RPKM) in solely one spermatogenic stage, including the post-meiotic stages, i.e., round spermatids and to a greater extent, mature sperm (**Figure 3.8B**). Overall, the patterns of X-linked REs across spermatogenesis imply a larger upregulation of genes and novel REs in the post-meiotic stages than previously thought, with the number of expressed X-linked REs largely following the patterns laid by autosomes. Of the 289 paternally transmitted REs, two were located on the X-chromosome, and both were exonic REs. The two REs are located (in hg38 coordinates) at chrX_2717605_2717652 and chrX_149929645_149930127, corresponding to CD99 and XX-FW81066F1.2, respectively. The spermatogenic roles of CD99, a cell surface glycoprotein involved in T-cell adhesion processes, and XX-FW81066F1.2, a poorly described transcript with a putative protein structure or antisense lncRNA function [135], are unknown. Although few paternally derived zygotic RNAs are X-linked, the expression patterns of REs located on the X chromosome were largely congruent with the current paradigm of Meiotic sex chromosome inactivation (MSCI) and reactivation during spermatogenesis [210]. However, the data also suggest that the process of post-meiotic X chromosome silencing (PMCI) during human spermatogenesis is selective, as many genes and novel REs escape silencing.
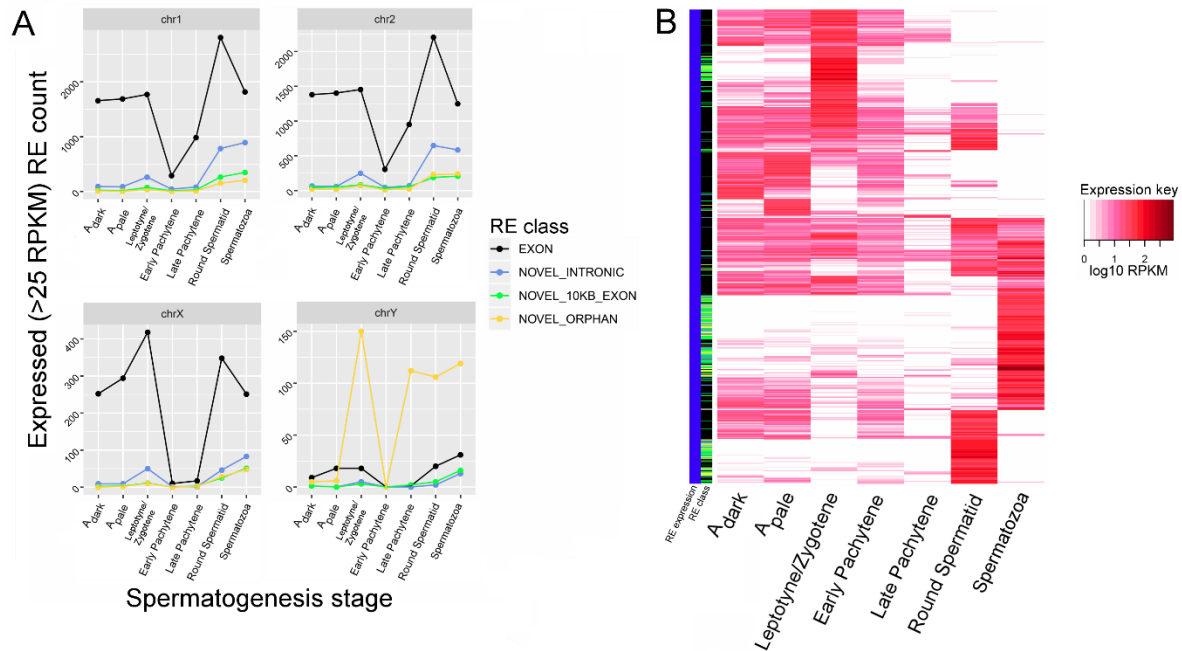
Figure 3.8. X-chromosome expression during spermatogenesis. (A) The number of expressed REs across each spermatogenic stage, for two representative autosomes (upper panels chr1 and chr2) and the sex chromosomes (lower panels chrX and chrY). The connected points are colored according to the RE class, with exonic REs in orange, intronic REs in green, near-exon REs in light blue, and orphan REs in purple. The X-axis of each graph presents the spermatogenic stage, with the pre-meiotic stages represented by A_{dark} and A_{pale}, the meiotic stages represented by Leptotene/Zygotene, and early/late Pachytene, and the post-meiotic stages represented by round and mature sperm. (B) An expression heatmap of X-chromosome REs that are primarily expressed (>25 RPKM) at one spermatogenic stage. RE class, shown adjacent to the RE expression column, shows exonic REs in black, intronic REs in green, near-exon REs in light blue, and orphan REs in yellow.

Paternal transmission of REs to the human embryo

It has been proposed and shown *in vitro* that human sperm deliver a cadre of RNAs upon fertilization [154, 160, 211, 212]. In the current study, the series of human RNA-seq profiles from sperm, oocyte, and embryo allowed for the identification of REs that are transmitted to the human oocyte solely by sperm. These are in addition to those 26,740 zygotic REs (5% FDR), associated with a total of 6,118 individual named genes, which are essentially provided by the oocyte, but not present in sperm (**Figure 3.9**). Up to 289 sperm REs were identified as a majority contributed by paternal transmittance, with an FDR of ~3.4%, and 75 REs essentially provided by the sperm, at an FDR of ~2.7% (**Figure 3.10A,B**).

Interestingly, the 289 REs were enriched for "cycling of RAN in nucleocytoplasmic transport" (p=8.36 x 10-8) and the Unc 51 Like Kinase (p=1.47 x 10-3). RAN cycling is required for effective translocation of RNA and proteins across the nuclear pore. The human sperm REs contain RANGAP1, XPO7, XPO6, NUP210, and NUP214, as members of the nucleoporin complex. Interestingly others have shown that at least in embryonic stem cells, the nucleoporin complex may regulate parentally imprinted genes [213]. In comparison, the Unc 51 Like kinase is associated with autophagy a process that is essential for the oocyte-to-embryo transition [214]. These observations are consistent with the view that the paternal RNAs may contribute to re-establishing nuclear transport in the zygote and clearance of extraneous cellular complexes post-fertilization, when cell lineages begin to be established. Compared to the oocyte's maternal contribution, relatively few paternal full-length RNAs are likely to be exclusively contributed to the embryo [156]. Of note, the genes associated with the paternally transmitted REs did not overlap those long RNAs suggested to be paternally derived in mouse [215]. This is likely due to the differences in genome activation, which occurs in the late 1- cell zygote in mouse [216], compared to the later 4-8 cell stage of human embryos, or other sperm derived RNAs providing a substitutive function [154, 217].
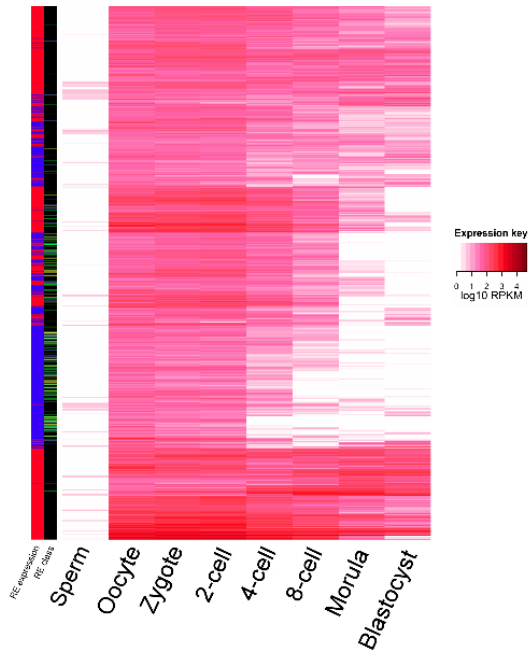
Figure 3.9. Expression heatmap of maternally derived REs. The overall expression level is represented in "RE expression", with red indicating a median expression exceeding 25 RPKM. RE class, shown adjacent to the RE expression column, shows exonic REs in black, intronic REs in green, near-exon REs in light blue, and orphan REs in yellow. The REs presented are supplied by the oocyte to the zygote (Sperm < 2 RPKM; Oocyte > 25 RPKM; Zygote > 10 RPKM)
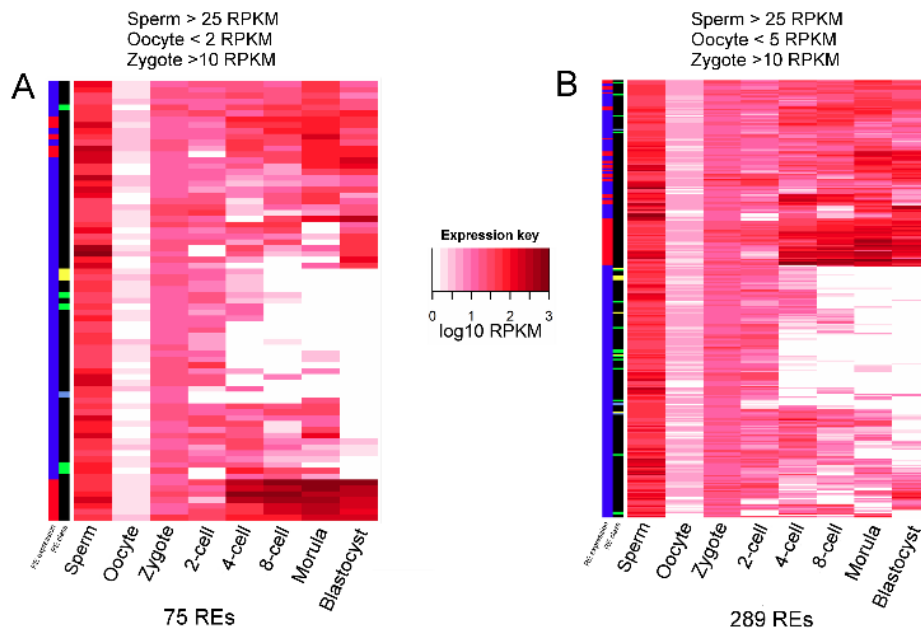
Figure 3.10. **Expression heatmap of paternally derived REs.** The overall expression level is represented in "RE expression", with red indicating a median expression exceeding 25 RPKM. RE class, shown adjacent to the RE expression column, shows exonic REs in black, intronic REs in green, near-exon REs in light blue, and orphan REs in yellow. The REs presented are supplied by the sperm to the zygote, with strong sperm preference (A) Sperm > 25 RPKM; Oocyte < 2 RPKM; Zygote > 10 RPKM and moderate sperm preference (B) Sperm > 25 RPKM; Oocyte < 5 RPKM ; Zygote > 10 RPKM

Differential gene expression during embryogenesis

Transcriptomic changes across mammalian embryogenesis have been well-studied, using both microarrays and RNA-seq [218-222]. However, these experiments have not addressed the contribution of intergenic RNAs to embryogenesis and, importantly, during human embryogenesis. Towards filling this gap, examination of expression changes of novel REs from oocyte to blastocyst, while considering the contribution of the spermatozoon, tested the hypothesis that both exonic and novel REs would exhibit distinct patterns.

To identify differentially expressed REs, a linear model was applied to the single-cell oocyte and embryonic RNA-seq datasets [168, 223]. Differential expression with REs reiterated previous analysis of RefSeq-annotated genes suggested that the oocyte, zygote, and 2-cell embryo contain a similar distribution of transcripts [168]. Few differences (59 REs) were identified between oocyte and zygote, and no differential REs were identified between zygote and 2-cell embryo (**Figure 3.11**). As expected, exonic REs exhibited characteristics of maternal genes, which are supplied by the oocyte and diluted as the embryo develops in anticipation of the 4 & 8-cell stage extensive Embryonic Genome Activation (**Figure 3.12**) [224]. This included a set of novel maternal REs specific to the early zygote (maternal genes) and EGA (the 4 & 8-cell stage). The majority of these novel maternal REs were intronic, suggesting e.g., (1) incomplete processing, (2) expression within an intron, (3) retention of circular RNA, or some other form. They were supplemented by a series of maternal intergenic orphan REs. Interestingly, novel REs also followed similar patterns, defining clusters of REs that are present during the minor first wave of human ZGA, as well as clusters that are active during EGA (**Figure 3.13**). While the novel REs with a maternal gene pattern are enriched for neuronal genes (Neuronal system, p=2.12e-05), those expressed during EGA are associated with protein metabolism (p=4.90e-06), consistent with the energy requirements of the early embryo.

	Oocyte	Zygote	2-cell	4-cell	8-cell	Morula	Blastocyst
	↑ 15 ↓ 44	↑ 0 ↓ 0	↑ 1820 ↓ 1220	↑ 1 ↓ 6		↑ 94 ↓ 45	↑ 54 ↓ 2450
Exon	46	0	2579	5		127	2224
Near-Exon	4	0	70	1		27	69
Intronic	3	0	277	1		7	146
Orphan	6	0	114	0		3	65
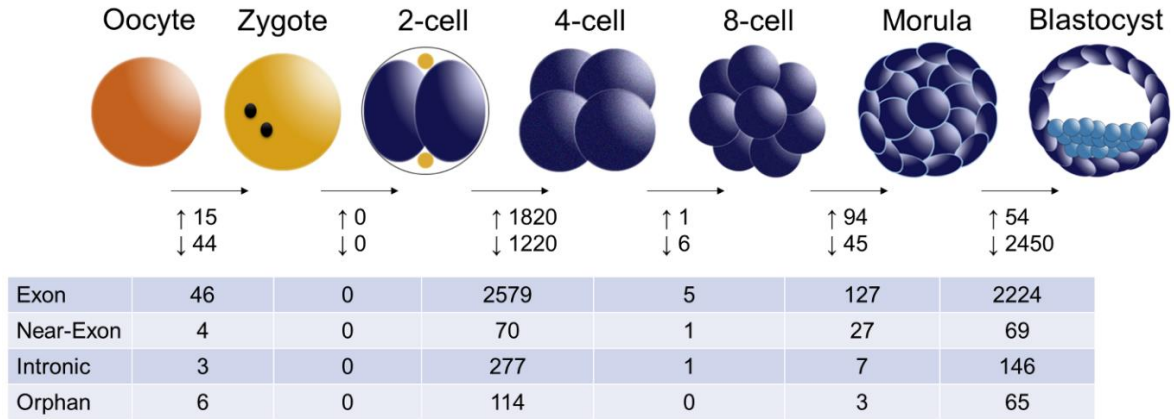
Figure 3.11. Differential RE expression across early embryonic development. The count of differential up- and down-regulated REs as embryogenesis proceeds is shown below the diagrams of cell types. The annotation classes of the total differential REs are summarized in the bottom table.
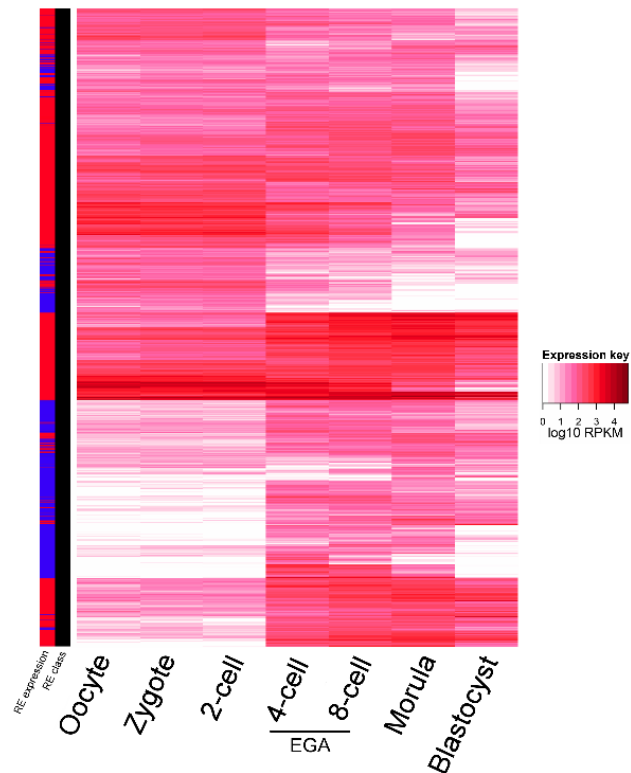
Figure 3.12. Expression heatmap of differentially expressed exonic REs across early embryogenesis. The overall expression level is represented in "RE expression", with red indicating a median expression exceeding 25 RPKM. RE class, shown adjacent to the RE expression column, shows exonic REs in black. The REs presented are differentially expressed across at least one stage.
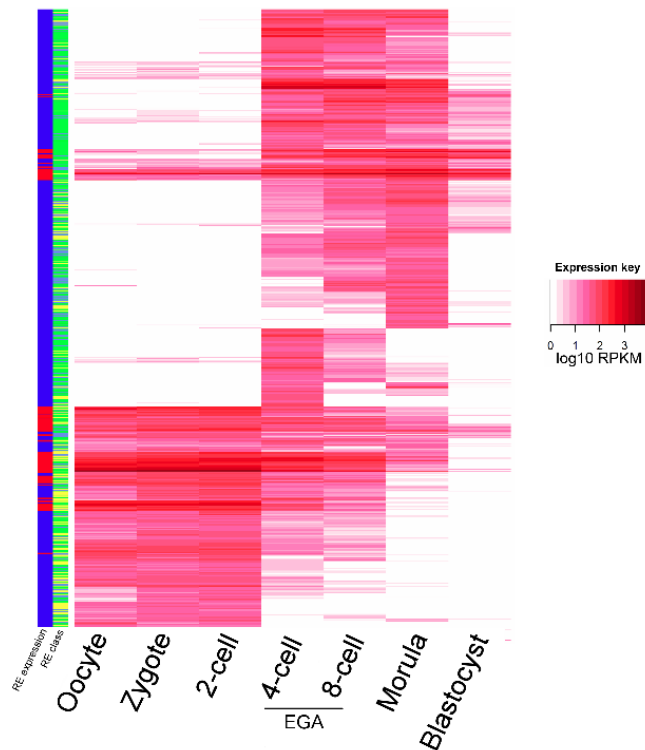
Figure 3.13. Differential novel REs across embryogenesis. The overall expression level is represented as left panel "RE expression", with red indicating a median expression exceeding 25 RPKM. RE class, shown adjacent to the RE expression column, shows intronic REs in green, near-exon REs in light blue, and orphan REs in yellow. The REs presented from oocyte to blastocyst are differentially expressed across at least one developmental stage.

Expression of genomic repeats across spermatogenesis and embryogenesis

Genomic repetitive elements and small non-coding RNAs are thought to play a role in confrontation-consolidation of the maternal and paternal genomes after fertilization [162, 225]. As novel REs tend to overlap genomic repetitive sequences, I examined RE expression to determine what genomic repeats may influence RNA expression throughout spermatogenesis and early human embryo development (**Figure 3.14**). The relative enrichment or depletion of repetitive sequences in the expressed REs was calculated for each available cell type. Briefly, the number of instances of genomic repeats overlapping expressed REs in each cell type were compared to an expected random distribution, with the random distribution drawn from the repeat occurrence in all available REs. Using a hypergeometric test, both relative

enrichment and depletion of repeat families were calculated across cell types. Despite the many instances of repeat depletion, there were relatively few instances of enrichment.
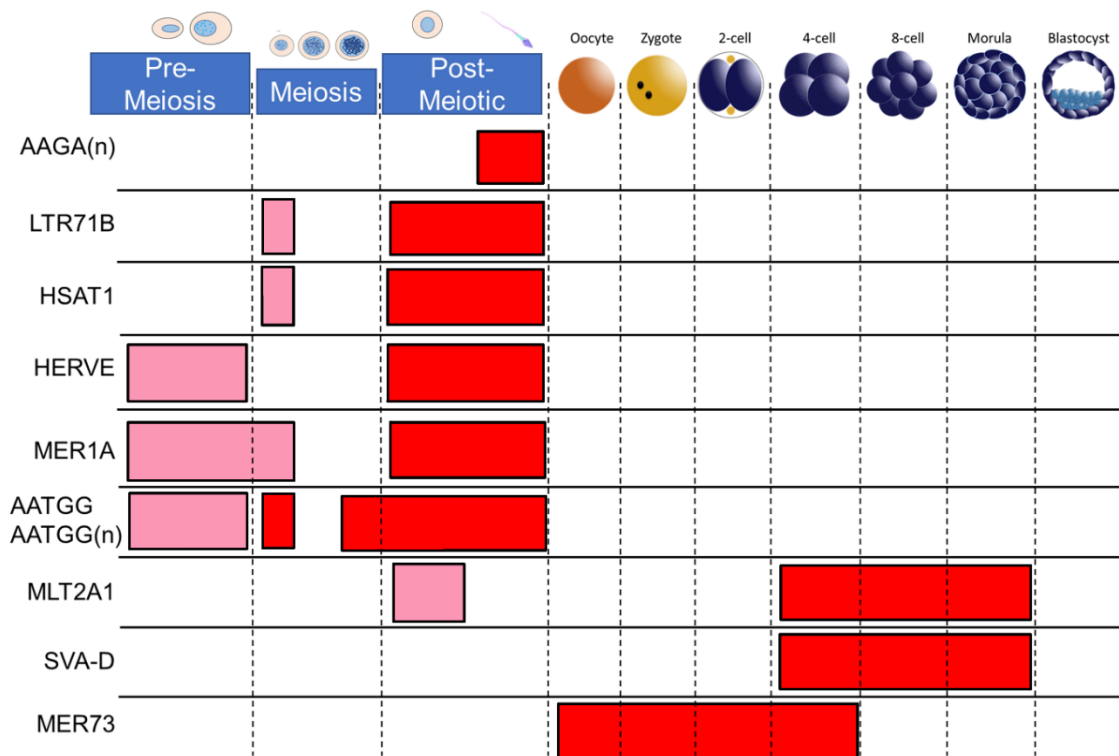
Figure 3.14. Expression of repetitive sequences across spermatogenesis and embryogenesis. Moderate enrichment (mean RE expression > mean expression across all cell types) is shown in pink, and strong enrichment (mean RE expression is an upper outlier) is shown in red. The name of the genomic repeat is given on the left of the diagram, and the cell type is shown at the top of the diagram.

Although several studies have examined the influence of environment on epigenetic marks, such as DNA methylation, at genomic repeats in spermatozoa, much less is known about genomic repeat expression during spermatogenesis and if genomic repeats are in part driving spermatogenesis, perhaps through transcriptional regulation or chromosomal reorganization [226, 227]. Four repeat families, LTR71B, HERVE-int, HSAT1, and MER1A were primarily expressed in both round spermatids and mature spermatozoa, while the centromeric repeat AATGG(n) showed greatest expression in the leptotyne/zygotene and late pachytene stages through the post-meiotic phase [228]. The simple repeat AAGA(n) was enriched solely in mature spermatozoa. The genomic repeats identified here as expressed

during spermatogenesis suggest that different repeats have different roles in spermatogenesis. For example, the centromeric repeat AATGG(n) likely plays a role in establishing stage specific chromosomal structure and position throughout spermatogenesis [229, 230]. The simple repeat AAGA(n) and HSAT1, primate-specific Satellite repetitive element, may also play a role in organizing sperm nuclear structure through Matrix-Associated Regions (MARs) of sperm, which are enriched in TTCT(n) and TCTT(n) repeats [230]. The remaining spermatogenesis-associated repeats LTR71B, HERVE, MER1A are all members of the HERV family of retroviruses or DNA transposons. The murine embryo and sperm are known to express a LINE-1-encoded Reverse Transcriptase (RT) that may serve to reverse transcribe the sperm-supplied retroviral and transposon RNAs for integration into the genome [44, 231, 232]. Insertion by retrotransposition might then act to provide regulatory networks, or genetically/epigenetically modify the developing embryo [44, 231] during syngamy. However, the presence of LINE-1-encoded RT in mature murine spermatozoon does not appear to extend to an enrichment of LINE1 RNAs in human sperm or zygote. This likely reflects a species differences, although one cannot exclude the influence of differing methodologies. However, MLT2A1 and SVA-D were both present during EGA, while MER73 was strongly enriched in oocyte and the early embryo (**Figure 3.14**). Both MLT2A1 (primate-specific) and MER73 are LTRs for ERVL endogenous retrovirus, while SVA-D is a hominid-specific composite retroelement (SINE-R + VNTR + Alu) [187]. Although SVA-D is a marker of naive human ESCs, consistent with the enrichment from 4 cell to morula stage, it is not enriched in blastocyst stage, from which human ESC cell lines are derived [233]. The ERVL retrotransposon has been previously implicated in mammalian embryonic development [188]. Notably, the presence of an RT in the early embryo would provide the opportunity for LTR71B, HERVE-int, and MER1A, components of HERVs and DNA transposons, to undergo transposition [234].

Overall, the above data suggest that the mechanism(s) driving spermatogenesis may involve the use of repetitive sequences as regulators of transcription and/or chromatin states [195, 235, 236]. Its nuclear architecture reflects the complex and orchestrated compaction and restructuring of its chromatin via protamination. This is linked through the nuclear matrix/lamina in a non-random manner [237], consistent with the current 3D models [238]. The enrichment of centromeric AATGG(n) repeat RNAs appears in the leptotyne/zygotene and late pachytene stages through the post-meiotic phase [228]. This repeat can form a double folded hairpin [228], that in mice can promote RNA:DNA hybrids mediating heterochromatin formation [239]. Perhaps this aids in excluding large repetitive DNA domains from homology searching enhancing the fidelity of meiosis as observed by the clustering of pericentromeric chromatin during meiosis [240].

In summary, this work introduced a RE discovery algorithm (REDa) that identifies tissue and cell type specific expression in both exonic and intergenic REs. Expression patterns of REs were identified across human spermatogenesis, extending the current knowledge of the transcriptome in developing human sperm. In addition to observing considerable effects of poly(A$^+$) enrichment, the sheer abundance of intergenic RNAs suggests that they play a large role in spermiogenesis. Of note, extensive expression of repetitive elements during spermatogenesis, suggests that perhaps these are driving spermatogenesis, while sperm-delivered repeat-derived RNAs may play more of a regulatory role in the human embryo.

CHAPTER 4

"THE EFFECTS OF DI-BUTYL PHTHALATE EXPOSURE FROM MEDICATIONS ON

HUMAN SPERM RNA"

This chapter was adapted from the following publication:

Molly Estill, Russ Hauser, Feiby L. Nassan, Alan Moss, and Stephen A. Krawetz. "The effects of high exposure to di-butyl phthalate from medications on human sperm RNA among men", In review with Scientific Reports.

i. Summary

Endocrine disruptors, chemicals that perturb hormonal function, are suspected of affecting reproductive function. Low-dose exposures to endocrine disruptors such as phthalates, widely used as plasticizers, is widespread. Patients with Inflammatory Bowel Disease (IBD) are often prescribed mesalamine, that in some forms is encapsulated within a di-butyl phthalate (DBP)-containing coating. The Mesalamine and Reproductive Health Study (MARS) was designed to address the physiological effect of *in vivo* phthalate exposure on male reproduction. As part of this effort, sperm RNA profiles among men with IBD and their relationship to DBP was longitudinally assessed across binary (high or background) DBP crossover exposures. As the level of DBP was altered as +/- DBP (H, high vs B, background), numerous changes in the composition of sperm RNA elements were detected. This was observed when the acute and recovery phases were compared, suggesting that, exposure to, or removal from high DBP, produces effects that require longer than one spermatogenic cycle to resolve, if at all. While the two study arms exhibit enrichment of different biological pathways, the arm which initiates the study on high-DBP mesalamine displayed activation of oxidative stress and DNA damage response pathways. However, DBP exposure has a minimal effect on small RNAs. Small RNAs were negatively correlated with specific genomic repeats, suggesting that they may contribute to repeat regulation. This provided insight into

both the influence of phthalate on the male germline, as a dynamic function of RNA during human spermiogenesis.

ii. Introduction

Endocrine disruptors, exogenous chemicals that can mimic or alter hormonal responses, are a prevalent feature in urban environments. A heterogeneous collection of natural and synthetic chemicals have been identified as likely EDs, including several well-publicized pesticides, such as dichlorodiphenyltrichloroethane (DDT), and plastic components, such as bisphenol A (BPA) and phthalate esters [55]. Phthalates, suspected endocrine disruptors, are commonly used as solvents and plasticizers in consumer products, such as polyvinyl chloride. They are also incorporated into coatings used in medications [56, 57]. Phthalates have been noted to act on peroxisome proliferator-activated receptors (PPAR) [58, 59]. Additionally, different phthalate species, including phthalate metabolites, have different capacities for modifying an endocrine response [58-60]. Although considerable literature suggests that gestational and neo-natal phthalate exposure is detrimental to reproductive function [61], the health effects of phthalates at environmentally relevant doses in adult humans is uncertain, particularly in the adult male.

To directly address the effect of physiological, *in vivo* phthalate exposure on male reproduction, the Mesalamine and Reproductive Health Study (MARS) (NCT01331551), designed as a clinical trial (https://clinicaltrials.gov/ct2/show/NCT01331551), was initiated. The MARS study, implemented from 2010-2018, was designed to assess semen quality and hormone levels in human males with longitudinally alternating DBP exposures, administered via mesalamine-containing medication for the treatment of Inflammatory Bowel Disease and Ulcerative Colitis. Patients with Inflammatory Bowel Disease (IBD) are often prescribed mesalamine, a medication which, in some formulations, has di-butyl phthalate (DBP) in the coating to allow for release of the active ingredient in the distal small intestines and colon [56, 57]. The DBP-coated medication(s), at maximal dosages, range from 300% - 700% of the

designated EPA Reference Dose (RfD) for a 150-pound individual [241-243]. The use of the DBP-coated medication produces urinary monobutyl-phthalate (MBP) levels 1000x higher than the levels found in the general U.S. population [69].

 With the MARS study longitudinal data structure, each individual acts as their own control, thus mitigating the potential exposure fluctuation, genetic variation and minimizes within environmental variation that often complicates accurate causal assessment of epidemiological data. As shown in **Figure 4.1**, the recruited subjects provided semen, urine, and blood samples across the longitudinal study. The cycle of human spermatogenesis and subsequent ductal transport takes approximately 90 days [36]. The MARS study required a minimum medication duration of 90 days, ensuring that an entire spermatogenic cycle occurred on a single medication. This duration was necessary to ensure that collected semen samples had developed solely from the current medication period. A total of 73 individuals were recruited for the MARS study. A subset of the individuals provided longitudinal samples across alternating DBP exposures, with a total of 60 individuals enrolled in the full protocol (baseline, crossover, and crossback visits collected) [69].
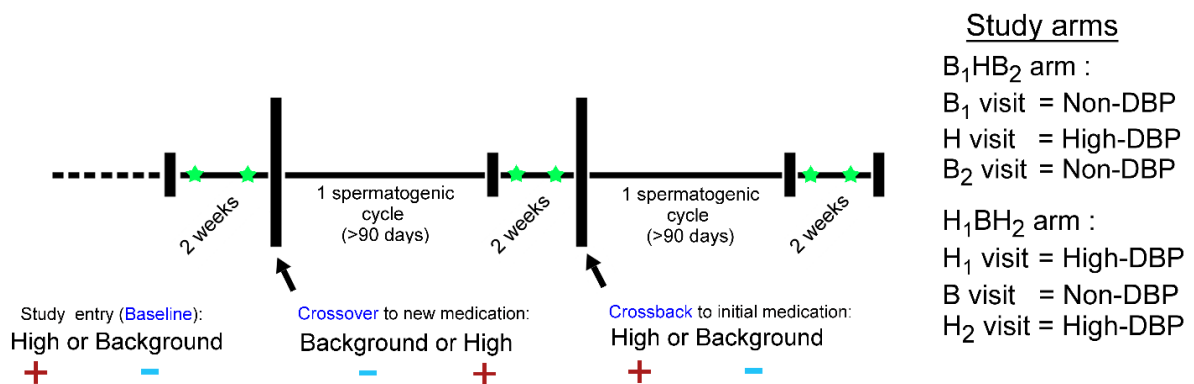
Figure 4.1. Crossover study design. Men enter the study having taken a mesalamine medication coated with (+) or without (-) DBP for at least 3 months. Semen, blood and urine were collected twice (green star) at baseline, after 4 months on an alternate drug (crossover), and after a final 4 months on the original drug (crossback).

Ejaculated spermatozoa, and their RNAs provide a snapshot of transcriptomic processes, capturing the influence of the paternal environment [17, 71, 73, 244] during spermatogenesis. MARS sperm samples were processed for RNA-seq, generating both a series of long RNA (>200 nucleotides) and small (<200 bp) RNA libraries to elucidate the biological processes being modified through phthalate exposure. The transcriptomic effects of both IBD and DBP exposure were assessed as a function of sperm RNA elements (REs), to provide a robust quantitative measure of effect [245]. Differential responses to high-DBP exposures were readily apparent in DBP-naïve men and men chronically exposed to high-DBP mesalamine. RNA levels of genomic repeats were examined to determine both which genomic repeats were well-represented in human sperm and which genomic repeats were part of the response to high-DBP exposure. While transcription is a dynamic process, the interactions of the sperm transcriptome until now remained unknown. Correlations between small RNAs and genomic repeats suggest a dynamic regulatory relationship.

iii. Materials and Methods

Sperm purification and library construction

To perform spermatozoal purifications, MARS samples were brought up to 1 ml volume with PureSperm Buffer (Nidacon) and laid onto a 50% PureSperm (Nidacon) gradient, then centrifuged at 300xg for 20 minutes [246]. The pellet was removed and washed in PureSperm Buffer. Cell counts were performed with a hemacytometer and microscope including a visual inspection for somatic cells and round spermatids.

Each sample was added to a 2ml tube containing 500ul RLT buffer (Qiagen) with 7.5ul BME (14.3M), 100 mg 0.2 mm nuclease free SS beads. The sample were homogenized, 500 ul Qiazol (Qiagen) added and homogenized again. 200 ul chloroform (0.124M) was added and samples shaken by hand then centrifuged at 12,000 x g for 20 minutes at 4ºC and the upper aqueous layer removed. RNA was isolated from the aqueous layer using a customized Qiacube (Qiagen) protocol [246]. RNA samples were treated with 4 U Turbo DNase (Ambion)

for 20 minutes at room temperature. Samples were tested for DNA contamination by RT-PCR and intron spanning PRM1 primers [246]. Samples still containing DNA were treated a second time and tested again by RT-PCR. The RNA samples void of DNA were quantified with a fluorescent assay as described [247]. An aliquot of RNA (5 ng or 10 ng) of each sample was used to make cDNA with SuperScript III RT (Invitrogen). The cDNA's were used as template for RT-PCR with intron spanning PRM1 primers providing verification of RNA.

Two nanograms of RNA per sample was used with the Seqplex (Sigma-Aldrich) amplification kit prior to library construction. Fifty nanograms of Seqplex cDNA product was used with the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs) to create sequencing libraries. Complete barcoded libraries were quantified and pooled for sequencing at a concentration of 2 nM per sample. All samples were subject to paired-end sequencing using either the NextSeq 500 (Illumina) sequencer, HiSeq 2500 (Illumina) sequencer or the HiSeq 4000 (Illumina) sequencer.

One nanogram of small RNA per sample was used with the NEXTflex Small RNA-Seq Kit v2 (Bioo Scientific) to create small RNA sequencing libraries. Complete barcoded libraries were quantified and pooled for sequencing. Samples were subject to paired-end sequencing using the MiSeq (Illumina) sequencer.

RNA-seq data processing methods

As a control cohort, RNA-seq datasets from males of idiopathic infertile couples who fathered a child were downloaded from the Gene Expression Omnibus (GEO), accession number GSE65683 [163]. A total of 52 RNA-seq datasets were downloaded from the Gene Expression Omnibus (GEO), accession number GSE65683 [163]. The MARS long RNA libraries were processed similarly to the GSE65683 samples. Paired-end reads were trimmed of adaptors and low-quality bases using Trimmomatic (version 0.36) [248], using default parameters (2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15), and requiring a minimum read length of 50 bp (MINLEN:50). The TruSeq Universal adaptor

(AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) was used as input to Trimmomatic. Paired and unpaired reads were aligned to the consensus human ribosomal RNA (GenBank: U13369.1) using HISAT2 (version 2.0.6) and the non-default parameters (-p10 --max-seeds 30 -k 2), then aligned to the human genome (hg38) and exogenous RNA spike-in sequences, using the same HISAT2 parameters. Reads without alignments in either U13369.1, the human genome (hg38), or an exogenous RNA were further assessed for alignments to repeat sequences (RepBase, February 2017 release), using HISAT2, with the following parameters (-p10 --no-spliced-alignment --max-seeds 10 -k 3). Reads without alignments in either U13369.1, hg38, an exogenous RNA, or RepBase were aligned to bacterial and viral genomes using Kraken (version 0.10.5-beta) and Jellyfish (version 1.1.10), implementing the full Kraken library and filtering alignments with a threshold of 0.15. Read alignments to the human genome, exogenous RNAs, and U13369.1 were processed to remove duplicated reads using Picardtools MarkDuplicates (version 1.129).

QC of the MARS study's long RNA samples was accomplished by examination of alignment statistics, allowing for the quantitative classification of samples failing QC (**Table 4.1**) into one of five categories (Category 1: Low genomic alignment, Category 2: High intergenic reads, Category 3: High bacterial and viral reads, Category 4: High spike-in reads; Category 5: High unmapped reads).

Table 4.1. Summary of sample quality. The first two columns ("Study arm" and "RNA type") indicate the study arm(s), and RNA type (small RNA or Long RNA library) for the given table row. "Total Samples" and "Pass QC samples" indicate the number of RNA-seq samples sequenced and the number of sequenced samples that passed quality control, respectively. "Subjects with samples" and "Subjects with pass QC samples" indicate the number of patients for which sequenced samples and sequence samples that passed quality control measures, respectively, were available. "Average pass QC samples per person" indicates the average number of sequence samples that passed quality control measures per patient. "Subjects with a Pair/trio" indicate the number of patients for whom the sequenced samples that passed quality control measures formed either a continuouous pair (baseline & crossover visits or crossover & crossback visits) or a visit trio (baseline, crossover, and crossback).

Study arm	RNA type	Targeted Insert Size	Total Samples	Pass QC samples	Subjects with samples	Subjects with pass QC samples	Average pass QC samples per person	Subjects with a Pair/trio
H_1BH_2 arm	Long RNAs	150 bp	124	93	27	27	3.4	20
B_1HB_2 arm	Long RNAs	150 bp	82	59	21	21	2.8	16
All arms	Long RNAs	150 bp	206	152	63	61	2.5	36
H_1BH_2 arm	Small RNAs	13-50 bp	62	58	21	21	3.0	12
B_1HB_2 arm	Small RNAs	13-50 bp	24	23	12	12	1.9	6
All arms	Small RNAs	13-50 bp	86	81	33	33	2.6	18

RNA element (RE) discovery algorithm (described in [245]) was applied to the MARS and GSE65683 (control) samples. Expression (in Reads Per Kilobase per Million - RPKM) for the RE loci was then calculated for all MARS and GSE65683 samples. Due to the use of REs, rather than whole transcripts, Paired-end read alignments were treated as individual (single) reads, and the common FPKM value was replace by RPKM. Depending on a read pair's insert size and RE length, this approach ran the risk of inflating the read count for a given RE. To mitigate this risk, read counts were assessed for the forward and reverse reads separately, with the read count from single (non-paired due to mate loss during quality control) reads were added to the forward and reverse read counts. The averaged read count between the forward and reverse reads were then used for generating the RPKM values.

Small RNA data processing methods

MARS small RNA libraries were trimmed of adaptors and low-quality bases, followed by removal of reads smaller than 13 bp. sncRNAbench (version 10.14) [249] was used for assigning reads to small RNA species and repeat classes, followed by a custom code for generating normalized expression values (RPM- Reads Per Million). Among the 86 samples, only one had an insufficient number of reads for reliable analysis (a threshold of 1000 aligned sense reads were required). Common small RNA species of interest, such as miRNAs, piRNAs, tRNAs, tRNA fragments, and siRNA can be detected with small RNA libraries (miRNAs ~22 bp) [250], piRNA ~ 24-31 bp [251], and tRNA fragments ~28- to 34-nt [71]).

Differential long RNAs

When comparing IBD samples to normal samples, the control cohort was composed of couples with idiopathic infertility. Therefore, to reduce the potential effect of infertility on the overall RNA expression profiles and concurrently maximize the size of the control cohort, only control sperm samples which presented with live birth (LB) (52 samples) were considered for use in differential expression [163]. Due to the idiopathic infertility of the couples composing the control cohort, the male component of the control cohort is not considered phenotypically normal. However, with respect to inflammatory diseases and in particular, inflammatory bowel disease or ulcerative colitis, the control samples were assumed to be disease free, and were thus labeled "Normal" in differential analyses.

To identify REs modified by IBD, a LM was used to compare the Normal sperm to the B_1HB_2 arm of the MARS study, with three total comparisons being performed (Normal vs B_1; Normal vs H; Normal vs B_2). The following formula was used for all three comparisons: "lm(value ~ seqset + lib + age + protamine_ct + sigma_ct + RNA_conc + cellcount_millions,data=input_data). Multiple-testing correction was applied as Benjamin-Hochberg. REs modified in a consistent manner (e.g. IBD-enriched or Control-enriched) in any two of the three visits (B_1, H, and B_2) were considered for further investigation. In the

current study, the Control samples were all present in a single sequencing batch, so a batch effect would be indistinguishable from the designation as a control sample. Several of the differential REs initially identified as IBD-enriched were differential solely due to near-zero expression values in all Control samples, suggesting a batch effect. Consequently, such REs were removed from consideration if the mean Control expression was less than 1 RPKM. This step removed 6 of the 32 REs enriched in IBD for at least two visits.

To identify REs modified by DBP exposure, a Linear Mixed-Effects Model (LMEM) was used to detect REs that changed with DBP exposure. Models were applied to each study arm independently. Two comparisons were done for each study arm, in order to find the changes occurring from Baseline visit to Crossover visits, and again from Crossover visit to Crossback visit. The following formula was applied to the H_1BH_2 arm: "rpkm ~ visit_simp + lib + period_asacol + bmi + season_warm + smokstat + age_bq + sigma_ct + percent_genomic_duplicated + percent_genomic + (1 | patient)". The following formula was applied to the B_1HB_2 arm of the MARS study: "rpkm ~ visit_simp + lib + bmi + season_warm + smokstat + age_bq + sigma_ct + percent_genomic_duplicated + percent_genomic + (1 | patient)". Due to the large number of REs tested in each comparison, standard multiple testing corrections removed all statistical significance. Therefore, a bootstrapped P-value was generated using random resampling. One thousand iterations were performed by permuting the RE expression value and running the given model, while maintaining the same sample order and covariate order. The empirical P-value was thus defined as the proportion of the 1000 random iterations that resulted in a P-value less than the original P-value.

$$Empirical\ P.value = 1 - \frac{\#\ iterations_{P.value > P_0}}{\#\ iterations}$$

REs were subsequently classified into eight unique expression patterns, with significance determined if the absolute value of the slope exceeded 10 RPKM and the empirical P-value was less than 0.05. REs which changed explicitly with DBP exposure would

display a pattern of increased expression from Baseline to Crossover visits, then decreased expression from Crossover to Crossback visits, or vice versa. Patterns of acute change were defined as those where an RE was significantly up- or down-regulated in the Baseline to Crossover comparison, but not altered from Crossover to Crossback. Patterns of recovery were defined as those where an RE was unchanged in the Baseline to Crossover comparison, but was significantly up- or down-regulated from Crossover to Crossback. Patterns of additional interest were those that continuously increased across the study arms or continuously decreased across the study arms.

Differential small RNAs

The human sperm samples investigated for small RNAs (<50 bp) are described above in **Table 4.1**. A total of 81 small RNA libraries was subsequently used in modeling, using an LMEM. Due to the low power in the B_1HB_2 study arm and insufficient numbers of complete trios, the predictive variable used was the DBP state (e.g. High DBP and Baseline DBP levels), whereas the long RNA analysis used the study visit as the predictive variable. The formula used for both study arms' small RNAs was "rpm ~ med_simp + bmi + season_warm + age_bq + sigma_ct + percent_genomic_duplicated + percent_genomic + (1 | patient)". In this formula, the influence of medication (high or low DBP) on small RNA expression was being corrected for patient BMI, seasonal warmth, patient age, sigma_ct from long RNA libraries, genomic duplication rate from the long RNA libraries, and proportion of long RNA reads aligning to the autosomal and sex chromosomes. For concordance of small RNA methods with those of the long RNAs, multiple testing correction was applied using a bootstrapped P-value, generated using random resampling. The given model was run 1000 times, permuting the RE expression value each time, while maintaining the same sample order and covariate order. The empirical P-value was thus defined as the proportion of the 1000 random iterations that resulted in a P-value less than the original P-value. Differential small RNAs were defined as those whose absolute value of the slope exceeded 5 RPM and empirical P-value was less than 0.05.

Repeat enrichment

Repeat enrichment was measured using the following formula, where "R" indicates the REs associated with the repeat of interest, "A" indicates the REs associated with any repeat, and the required median expression threshold for a study visit is 25 RPKM. Repeat enrichment is the change in contribution of the given repeat to the repeat population, when a given expression threshold is applied to both the repeat of interest and the total repeat population.

$$\Delta\ ratio = \frac{\#\ R_{expressed}}{\#\ A_{expressed}} - \frac{\#\ R}{\#\ A}$$

The statistical significance of an enrichment or depletion (indicated with a positive or negative Δ ratio, respectively) was tested using a hypergeometric test, implemented in *stats* R package. This method is similar to the one implemented in Estill et al [245]. The repeat enrichment analysis merely indicates if a repeat type is relatively under- or over-represented in the expressed REs, relative to the expected proportion when no expression threshold is applied.

Small RNA-only Correlations

Expression levels of genomic repeats and small RNA species were determined using the small RNA libraries. The total number of small RNA species and genomic repeats totaled 12,779 individual sequences. In order to remove spurious results due to poorly expressed small RNAs, the compilation was subsetted to those small RNAs with expression exceeding 5 RPM in a portion of each sample set. In the B_1HB_2 arm, with a total of 24 Samples, using a threshold of 5 RPM in at least 10 samples, this threshold allowed 602 small RNAs for correlation. In the H_1BH_2 arm, with a total of 62 Samples, this threshold allowed 939 small RNAs for correlation. Spearman correlation was applied to each individual study arm for correlational analysis of small RNAs.

Correlation of long RNA genomic repeats and small RNAs

To ensure that the conclusions drawn from the small RNA correlations were consistent, repeat expression estimated from long RNA libraries was generated. The 371,805

available REs overlapped a total of 4,449 repeat names. The expression of each repeat instance was averaged across a given sample, producing a resulting expression matrix of 4,449 rows (representing 4,448 repeats and 1 dummy classification of no repeat overlap) and n sample columns (where n changes according to study arm size). This was merged with a series of small RNAs (generated previously from sncRNAbench), which were expressed (greater than 5 RPM) in at least 10 of the 67 small RNA libraries, producing a set of 1,829 small RNAs. The 1,829 small RNAs were correlated with the 4,449 repeats in both the H_1BH_2 arm (n=50 samples) and B_1HB_2 arm (n=16 samples), with one small RNA library omitted due to being a replicate sample. It is important to note that, in order to generate the expression values for repeats in the long RNAs, the repeats were summarized across all REs overlapping the given repeat. This approach has the disadvantage of negating locus-specific regulatory effects (such as cis-regulatory effects), due to the averaging of all individual repeat loci. Therefore, in the case that repeats do have a regulatory effect, the summarization across all genomic instances (in context of REs) may introduce inaccuracies.

Ontological enrichment

Gene Ontology (GO) enrichment was generated using the GeneRanker function of Genomatix (Eldorado version 12-2017), from the Genomatix software suite (https://www.genomatix.de/), version 3.10. The gene names associated with differential exonic, novel near-exon, or novel intronic REs were compiled and used as input to GeneRanker. Pathway enrichment was assessed using Ingenuity Pathway Analysis (version 1-13, Content build 46901286). The expression changes occurring in differential exonic REs were compiled and used as input for IPA. The compilation method used was to average the expression changes of all differential exonic REs belonging to a given gene. This produced a single slope, p-value, and computed log2ratio for each gene name.

iv. Results

Expression – RE discovery

RE discovery was performed on MARS samples and external sperm samples [163], with REDa parameters allowing an RE to be detected if it was present in solely a single sample. Therefore, downstream analyses used REs that surpass a minimum expression value (in RPKM) in several samples, thus reducing the computational burden of the downstream analyses. RE expression was then calculated for all MARS samples and external sperm datasets described in GSE65683. Novel REs and exonic REs have largely similar length distributions (**Figure 4.2**). However, novel REs tend to have much more uneven length distributions, compared to the exonic REs, which is likely due to the lesser number of novel REs. Interestingly, each of the novel RE classes exhibited several REs exceeding 1 kb, suggesting that at least a small number of intergenic and intronic are expressed along a long stretch of the genome (**Appendix G**). There were 138 novel REs with a width exceeding 1 kb (78 near-exon, 50 intronic, and 10 orphan RE classes). Of the 138 novel REs, 103 were associated with a genomic repeat, with the most common repeat families being Simple repeats, Endogenous RetroViruses (ERVs), and L1 repeats (30, 16, and 16 REs, respectively). As expected, approximately 90% of the very long novel REs were within 5 kb of DNase site [245]. The very long novel REs were not associated with known Topologically Associating Domains (TADs). Approximately a third of the very long novel REs were within 5 kb of a GM12878 CTCF binding site and/or a known sperm MNase hypersensitive site. The proportion of sperm MNase-localized very long novel REs was slightly higher than observed for the complete set of RE classes (ranging from 5% to 25%). However, the very long novel REs did not appear to be associated with a particular genomic structure, as the overlaps of very long novel REs with epigenetic marks and regulatory genomic sequences appear to reflect the expected proportions of the complete set of RE classes [245].
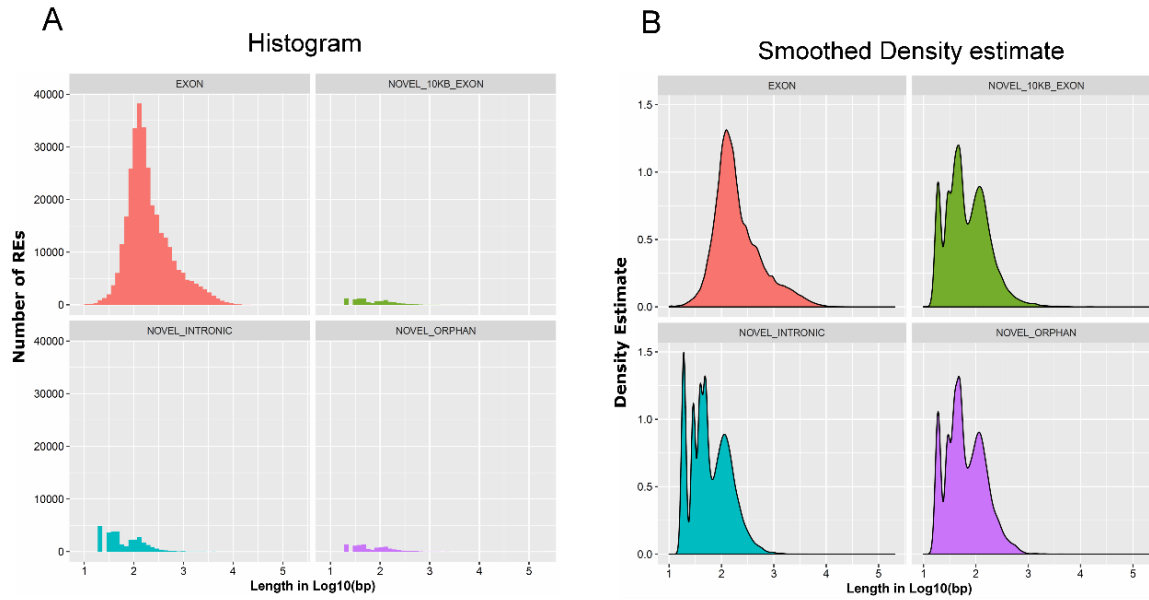
Figure 4.2. RE length distribution. (A) Histogram of RE lengths, segregated according to RE type (exonic, intronic, near-exon, and orphan). The X-axis indicates the RE length in log 10 scale, and the Y-axis indicates the number of REs. (B) Smoothed density estimates of the RE lengths, segregated according to RE type. Exonic REs are noted in red, near-exon (<10 kb from exon) are in green, intronic REs are noted in blue, orphan REs are noted in purple.

Quality control of sequenced sperm RNAs

Quality control of sperm RNA libraries, prior to sequencing, was assessed using the protocol described previously [252, 253]. Suspected low-quality sperm RNA libraries were excluded from sequencing. However, despite this precaution, a portion of each set of sequenced RNA samples often failed post-sequencing quality control. Within the MARs study, quality control determination of the MARs samples was necessary to remove samples with poor sequencing performance from subsequent modeling efforts. For QC of the MARS study's long RNA samples, a quantitative approach was developed by first qualitatively assessing which samples appeared either deficient in read coverage or had excessive intergenic reads. The initial judgement of poor quality samples was based on read profiles and read coverages at several key loci, such as the human protamine locus (which is expected to have thousands of reads) and ACSBG2 (which is well-expressed and has few intronic reads). Subsequent examination of alignment statistics allowed for the quantitative classification of samples failing

QC into one of five categories (Category 1: Low genomic alignment, Category 2: High intergenic reads, Category 3: High bacterial and viral reads, Category 4: High spike-in reads; Category 5: High unmapped reads). The specific classification algorithm used in assigning the 5 categories is provided in **Appendix H**. The proportion of samples failing quality control for the MARS samples are shown in **Table 4.1**. It is important to note that small RNA libraries use different criteria for assessing sample quality.

Notably, poorly performing samples tended to have overlapping alignment statistic ranges. Therefore, the five classifications are not mutually exclusive. In the case of a sample with potential to belong to multiple classifications, the classifications were assigned in a sequential order (from Category 1 to Category 5). A sample assigned to one of the five classifications may still exhibit an acceptable read distribution (e.g. high reads at expected loci, such as PRM1, and relatively few intergenic reads). This is particularly true for Categories 3 and 5 (Category 3: High bacterial and viral reads; Category 5: High unmapped reads). However, such visually acceptable samples were still removed from consideration in order to eliminate the possibility that the samples had altered RNA profiles.

Clinical characteristics were statistically indistinguishable between the study arms, with the exception of Total_score, a measure of IBD severity symptoms, which was on average higher in the B_1HB_2 arm. Both study arms were considered to have very low Total_scores, with a maximum score of 5 in the current study (observed for a single patient). By design, all recruited individuals in the MARS study were required to have a mild IBD score on the simple clinical colitis activity index (five or less for UC) or Harvey-Bradshaw index (four or less for CD) [69].

Transcriptomic similarity of human sperm to Testis

In addition to the previously described quality control measures, the Krawetz laboratory developed a method of assessing sample quality through correlational analysis of RE expression to GTEx expression values [163, 254]. The process first examines each GTEx

tissue to identify the most highly expressed genes. The top 5,000 most abundant genes were chosen as an optimal threshold for including both housekeeping genes (which are expected to be highly expressed) and genes unique to the tissue type. The exonic REs associated with each gene name were then averaged across a sample to produce a single value. A Spearman correlation was then implemented across the top 5,000 gene names to determine the relative similarity of a given sample to each of the 53 tissues. The Spearman rho was subsequently transformed into a rank, with a rank of 1 given to the tissue with the highest (positive) rho and 53 given to the lowest rho value. A sperm sample would be expected to be most similar to testis, and therefore have a rank of 1. The resulting tissue similarity distribution for MARS samples is shown in **Figure 4.3**. Nearly all quality-controlled sperm samples, such as those used in the MARS study, were expected to be most similar to Testis. Accordingly, GTEx correlation of the MARS study samples (**Figure 4.3**) indicated that the gene expression in the samples (in the form of averaged exonic REs) was most similar to GTEx testis tissue. The only two samples without Testis ranking at 1 are ranked as most similar to "Cells - EBV-transformed lymphocytes", with 2nd and 3rd ratings at Stomach and Prostate tissue. However, sperm purification protocols make it extremely unlikely that lymphocytes, stomach cells, or prostate-derived cells would contaminate the purified sperm. Therefore, this Spearman correlation with non-testis tissue (for the two indicated samples) does not indicate a contamination of the sperm sample and subsequent failure to sufficiently purify spermatozoa. The two samples are in the H_1BH_2 arm, which is the most populous study arm, and are members of a crossover/crossback set and a trio (baseline, crossover, crossback). The samples in question are otherwise of acceptable quality, and the diagnostic value of the GTEx correlations has not been proven. Based on the overall sample quality and large size of the H_1BH_2 arm, the inclusion of the two samples in modeling was not expected to adversely influence or bias the outcome.

Top 5000 genes, out of 33182 genes
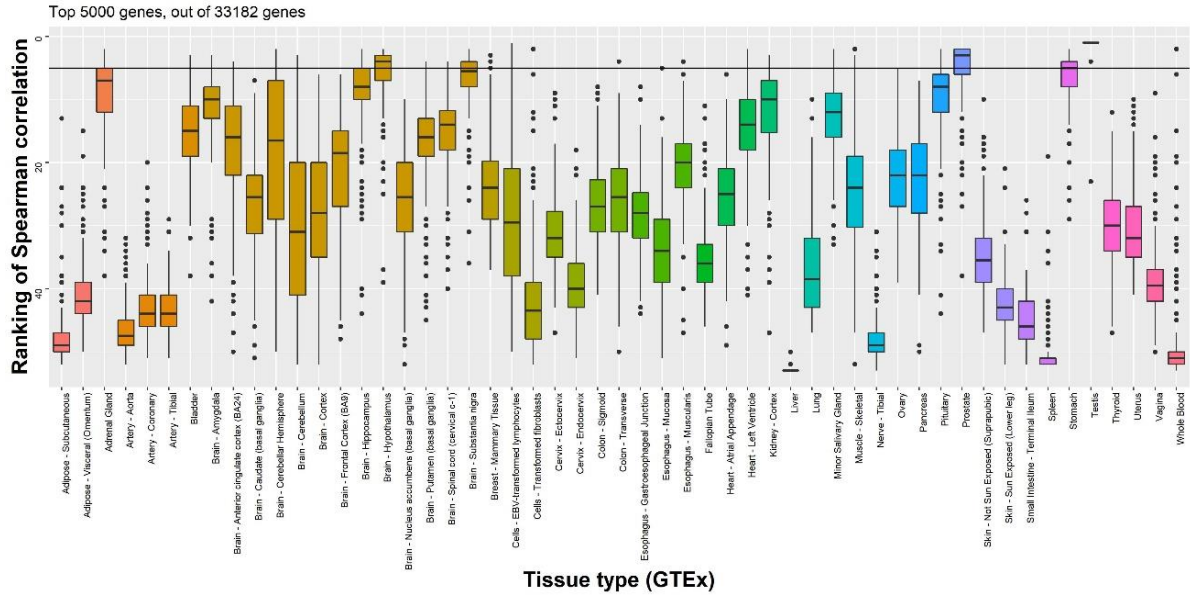
Figure 4.3. GTEx tissue distribution for quality controlled MARS samples. Ranking of Spearman correlation for the top 5000 genes of each tissue are shown for 152 quality controlled MARS sperm samples. The X-axis represents the individual GTEx tissues, while the Y-axis represents the relative ranking of Spearman rho.

In addition to the previous GTEx correlations, the most abundant RNA elements in the pass-QC samples from the MARS dataset were examined, with the expectation that the RNA elements should reflect transcripts known to be highly expressed in testis [136]. **Table 4.2** presents the top 50 most expressed REs, ordered from highest RPKM to lowest. PRM1, PRM2, TNP1, SMCP, CRISP2 are expressed primarily in testis, while GIGYF2 is most highly expressed in testis. This basic analysis indicates that the RNA-seq datasets are indeed spermatozoal, and therefore, any differentially expressed REs reflect changes in the sperm transcriptome. As indicated in the methods used for processing the MARS samples, precautions were taken to ensure that somatic cells were excluded prior to RNA extraction. The previous GTEx analysis identified two sperm samples were not most similar to testis, but to EBV-transformed lymphocytes. However, based on purification protocols, this similarity was not expected to be due to contamination with lymphocyte cell lines. At this time, further optimization of the correlation approach to GTEx tissues may be needed.

Table 4.2. Highly expressed exonic REs across the MARS study. The top 50 most expressed exonic REs, ordered from highest RPKM to lowest, are listed. For REs that encompass more than one transcript, the gene names are separated by a comma. REs highlighted in bright yellow indicated a series of genes that are known to be primarily expressed in testis and sperm.

Element Identifier	Gene Symbol	Median expression across MARS study	Note	Tissue expression	Median testis GTEx expression
chrM_577_3304	MT-TF	352637.5	Mitochondrial	Low in all tissues	0.71
chr6_52995620_52995950	RF00100,RN7SK	6053.6		Present in all tissues	50.03
chr16_11280836_11281035	PRM1	5693.2	Protamine 1, Exon 1	Testis specific	20530
chrM_5904_7445	MT-CO1	5124.1	Mitochondrial	Present in all tissues	20780
chr14_49586579_49586878	AL139099.4,RN7SL1	4284.3		Present in all tissues	70.93
chr16_11275639_11275981	PRM2	4272.6	Protamine 2, Exon 1	Testis specific	21190
chr14_49862550_49862849	RN7SL2	4121.2		Present in all tissues	58.13
chrM_7518_8269	MT-TD	3233.5	Mitochondrial	Low in all tissues	0
chr5_7304232_7304261	AC091951.1	3147.1		Testis specific	144.2
chr16_11281127_11281350	PRM1	3002.6	Protamine 1, Exon 2	Testis specific	20530
chrM_8366_14148	MT-ATP8,MT-ATP6,MT-CO3	2886.1	Mitochondrial	Present in all tissues	16720;27980;34450
chrM_3307_4331	MT-ND1	2672.1	Mitochondrial	Present in all tissues	22760
chrM_4402_5579	MT-TM	2381.4	Mitochondrial	Brain specific	0.71
chrM_14149_14742	MT-ND6	1986.0	Mitochondrial	Present in all tissues	1774
chr1_152878317_152878446	SMCP	1905.9	Sperm Mitochondria Associated Cysteine Rich Protein, Exon 1	Testis specific	1606
chr16_11276100_11276480	PRM2	1887.2	Protamine 2, Exon 2	Testis specific	21190
chr1_16740516_16740679	RNU1-4	1777.4		Unknown	0
chr2_216859896_216860064	TNP1	1764.3	Transition Protein 1, Exon 2	Testis specific	8839
chr9_9442060_9442380	RN7SL5P	1737.4		Low in all tissues	0
chr2_216859458_216859695	TNP1	1601.2	Transition Protein 1, Exon 1	Testis specific	8839
chrM_14747_15953	MT-CYB	1536.0	Mitochondrial	Present in all tissues	15430
chr12_112267077_112267394	RN7SKP71	1458.7		Low in all tissues	0
chr5_7303912_7303962	AC091951.1	1425.8		Testis specific	144.2
chr5_7302224_7302282	AC091951.1	1372.1		Testis specific	144.2
chr3_15738515_15738809	RN7SL4P	1368.6		Low in all tissues	0.27
chr6_49712501_49712604	CRISP2	1326.8	Cysteine Rich Secretory Protein 2, Exon 2	Testis specific	1194
chr2_88788318_88788346	ANKRD36BP2	1279.4		Present in multiple tissues, highest in testis	13.94
chr15_22771692_22771750	AC011767.1	1270.8		Unknown	
chr1_16514122_16514285	RNU1-1	1241.7		Low in all tissues	0.51

chr14_66488987_ 66489037	CCDC196	1199.3		Testis specific	181.4
chr6_49697860_4 9697957	CRISP2	1193.9	Cysteine Rich Secretory Protein 2, Exon 8	Testis specific	1194
chr6_34697159_3 4697470	AL451165.2	1122.3		Present in all tissues	86.31
chr14_49853616_ 49853914	RN7SL3	1094.1		Low in all tissues	0.98
chr2_216871456_ 216871639	LINC01921	1085.3		Low in all tissues, possible testis specificity	0
chr22_20341524_ 20341548	AC007731.1	1073.8		Unknown	0
chr5_7305074_73 05203	AC091951.1	1049.6		Testis specific	144.2
chr1_152884403_ 152885047	SMCP	1045.0	Sperm Mitochondria Associated Cysteine Rich Protein, Exon 2	Testis specific	1606
chr15_28846220_ 28846249	GOLGA6L7	1041.0		Testis specific	42.6
chr22_18424977_ 18425035	FAM230A	980.6		Testis specific	15.52
chr2_88788442_8 8788514	ANKRD36BP2	919.6		Present in multiple tissues, highest in testis	13.94
chr22_18736559_ 18736583	LINC01662	889.9		Testis specific	0.36
chr5_7301753_73 01894	AC091951.1	867.6		Testis specific	144.2
chr2_232816871_ 232817032	GIGYF2	846.9	GRB10 Interacting GYF Protein 2, Exon 21	Present in multiple tissues, highest in testis	45.91
chr5_177729023_ 177729053	FAM153A	833.9		Present in all tissues	18.94
chr16_14997748_ 14997826	PDXDC1	815.9		Present in all tissues, highest in testis	116.9
chr1_144560666_ 144560829	RF00003	807.9		Unknown	0
chr14_60245752_ 60246046	PPM1A	754.7		Present in all tissues	30.22
chrM_5761_5891	MT-TC,MT-TY	748.9	Mitochondrial	Low in all tissues	1.79;3.41
chrX_103712236_ 103712360	TMEM31	746.6		Testis specific	147.7
chr2_232790698_ 232790915	GIGYF2	732.6	GRB10 Interacting GYF Protein 2, Exon 10	Present in multiple tissues, highest in testis	45.91

Influence of sample characteristics on sperm RNA expression

Spermatozoa are considered transcriptionally and translationally inert. RNA profiles of mature sperm therefore do not reflect active transcriptional activity, but are considered to reflect transcriptional activity during spermiogenesis and/or transcripts acquired during epididymal transit. Therefore, the use of the term "expression" refers to the RNA profiles of the ejaculated spermatozoa. A primary purpose of generating expression values for REs was for performing differential expression. Therefore, it was essential to determine the sample

characteristics (patient phenotype and sequencing statistics) that exerted a major influence on RE expression, and subsequently account for these characteristics when modeling differential expression. Correct covariate selection is necessary for accurately modeling and interpreting an outcome (in this case RNA levels). For example, initial differential expression attempts of the MARS dataset used covariates chosen primarily for their assumed technical and biological importance. However, partly due to an overall increased variance within the H_1BH_2 arm, the outcome of the initial analyses showed a disparity between the two study arms, with the B_1HB_2 arm having a far higher amount of differential expression than the H_1BH_2 arm. An unbiased assessment of the sample characteristics that exhibited the most influence on RE expression, using Single Value Decomposition, Pearson correlation, and statistical significance in linear modeling, identified influential covariates. The implementation and results of each of the steps are provided below. When influential covariates were included in the expression model, the B_1HB_2 arm and H_1BH_2 arm presented a similar number of differential REs. Therefore, correct covariate selection reduced the disparity in differential expression between the two study arms.

Covariate selection for the MARS dataset proceeded as described in the following. The RE dataset was first reduced to the MARS samples that passed QC, and additionally reduced by removing REs that failed to surpass 25 RPKM in at least a third of the samples of interest (e.g. Quality-controlled H_1BH_2 MARS samples). Single Value Decomposition (SVD) analysis was performed, with the result visualized as shown in **Figure 4.4**. In the H_1BH_2 arm MARS samples, SVD analysis revealed that Principle Component (PC) 1, PC-2, and PC-3 accounted for 93.0%, 5.4%, and 0.49% of the data variance, respectively. In the B_1HB_2 arm MARS samples, SVD analysis revealed that Principle Component (PC) 1, PC-2, and PC-3 accounted for 91.1%, 7.6%, and 0.46% of the data variance, respectively. In a dataset combining both study arms (thus being composed of all quality-controlled MARS samples, SVD analysis revealed that Principle Component (PC) 1, PC-2, and PC-3 accounted for

92.2%, 6.3%, and 0.40% of the data variance, respectively. In all quality-controlled MARS samples, PC-1 was significantly correlated with library duplication rates (in both ribosomal and genomic reads), percentage of repeat sequence-associated reads, percentage of genomic (autosomal and sex chromosomes) alignments, percentage of autosomal alignments, percentage of ERCC spike-in reads, and percentage of ribosomal alignments. Potential co-linearity of the sample characteristics, which would be suggested by the presence of highly positive or negative correlations, was tested using a Pearson correlation, shown in **Figure 4.5**. The characteristics found to be potentially collinear were the following: library duplication rates and protamine_ct; percentage of ribosomal alignments and sigma_ct; percentage of ribosomal alignments and percentage of genomic (autosomal and sex chromosomes) alignments; percentage of repeat-associated reads and percentage of unmapped reads; percentage of repeat-associated reads and percentage of adaptor-derived reads.
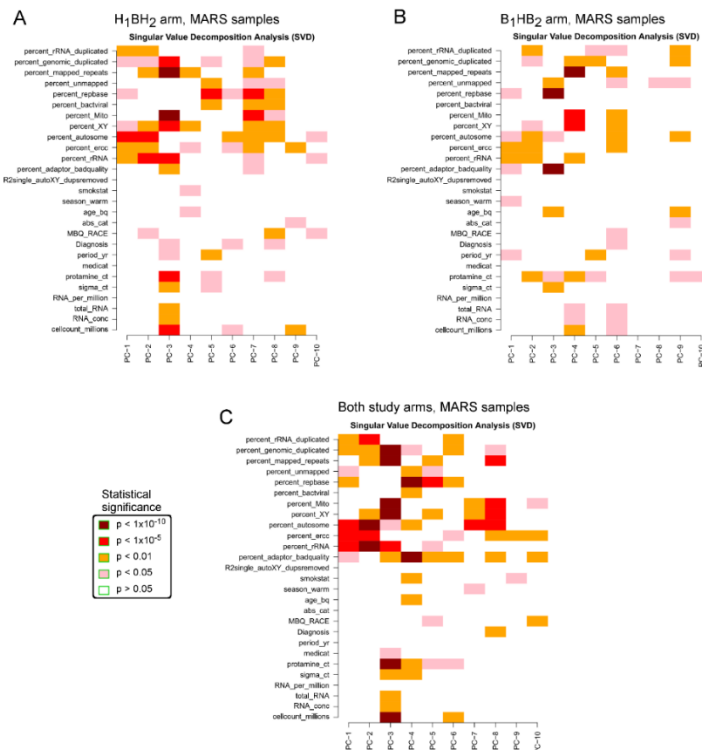
Figure 4.4. Singular Value Decomposition Analysis of quality-controlled MARS samples. (A) Association of sample characteristics for H_1BH_2 arm. (B) Association of sample characteristics for B_1HB_2 arm (C) Association of sample characteristics for all quality-controlled MARS samples, regardless of study arm. The association of a sample characteristic with a Principle Component (PC) is measured using a standard linear model (lm) for numeric characteristics, while categorical or integer characteristics (eg. Medication, diagnosis, race, season, smoking status, R2single_autoXY_dupsremoved) used a Kruskal-Wallis rank sum test (kruskal.test). The P-value for the linear model or Kruskal-Wallis test is color-coded for each pairwise test of Principle Component and sample characteristic, with the color legend provided in the left-bottom corner of each graph.
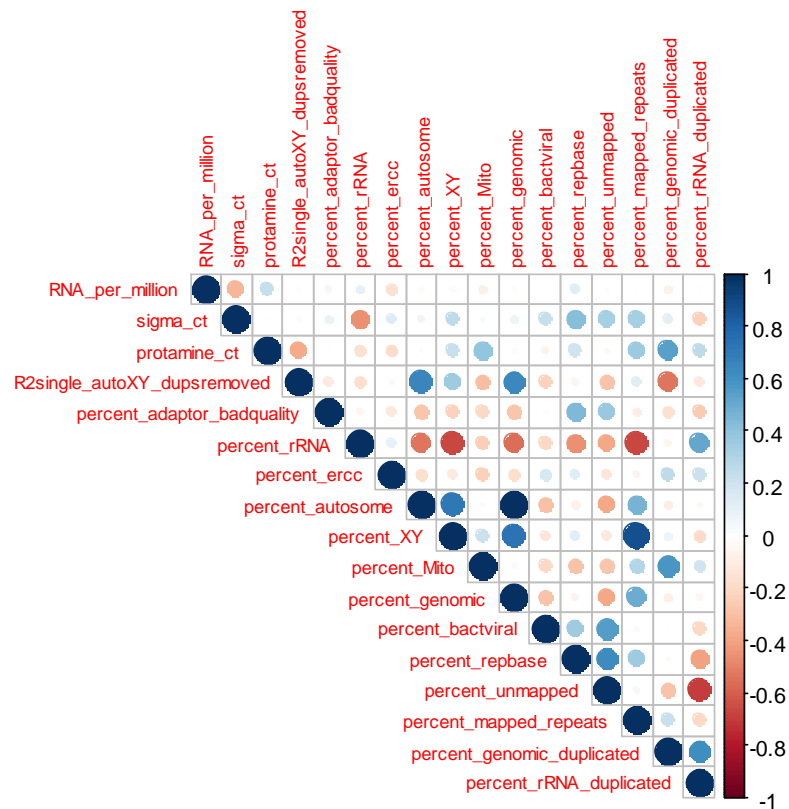
Figure 4.5. Pearson correlation of numeric sample characteristics for all quality-controlled MARS samples. The Pearson's rho value is plotted as a color gradient on the right-hand side of the graph, with deep red representing a negative correlation and dark blue representing a positive correlation as indicated in the bar.

In order to determine which characteristics would be influential in context of the research question of interest (expression change with phthalate levels), a basic linear model was applied to all quality-controlled MARS samples (formula: rpkm ~ visit_cat + lib + RNA_per_million + period_yr + bmi + season_warm + smokstat + age_bq + sigma_ct + percent_genomic_duplicated + percent_adaptor_badquality + percent_ercc + percent_genomic + percent_bactviral + percent_repbase). The RNA-seq library parameters sigma_ct, percent_genomic_duplicated, percent_ercc, and percent_genomic presented a significant P-value (P<0.05) for their concomitant beta estimate. sigma_ct, percent_genomic_duplicated, and percent_genomic were subsequently used as technical (RNA-seq library) covariates in linear mixed effects models and basic linear models of the MARS and pilot datasets. Although ERCC (spike-in) percentage did have a significant P-value

in the linear model (P=0.027), ERCC RNA spike-ins were previously noted to have extreme variability in the final sperm RNA-seq libraries. Additionally, ERCC spike-in proportions after sequencing may correspond to insert quality, as the ERCC mixture may provide a better template than the sperm RNA. However, ERCC mixtures were added at different concentrations to the different MARS sequencing batches (in an effort to reduce the amount of sequencing materials dedicated to the ERCC spike-in). Therefore, ERCC proportion was omitted from expression modeling. The final components of the formulas for Linear Mixed-Effects Modeling (LMEM) and Linear Models (LM) accounted for a series of clinical parameters (Sequencing batch, Time on High-DBP drug, Body Mass Index (BMI), Season, cigarette smoking, and patient age) and technical parameters (sigma_ct, duplication rate of genomic (autosomal and sex chromosomes) reads, and percentage of alignments located on autosomal and sex chromosomes. In the case of datasets missing any of the above parameters (e.g. external dataset from Jodar et al.), the missing parameters were omitted from the given model [163].

IBD-induced changes on RNA

Several studies have applied RNA-seq to IBD (previously known as Irritable Bowel Syndrome (IBS)), with a primary focus on intestinal biopsies. While IBD is an inflammatory condition that is limited to gastrointestinal tissue, altered brain-gut interactions do occur [255], and additional direct or indirect effects on peripheral tissues are possible. This is underscored by work by Gupta et al, who identified 288 dysregulated genes in peripheral blood mononuclear cells (PBMCs) between IBD and healthy control groups, with 12 of the dysregulated genes being associated with the immune system (4 anti-inflammatory, 8 pro-inflammatory). Iborra et al. [256] identified a series of serum miRNAs with altered expression in IBD. Machine learning approaches have also uncovered a peripheral blood miRNA signature indicative of IBD [257]. Plasma-induced signature analyses performed using blood plasma from IBD patients [258] also suggest that the plasma of CD and UC patients contain

an immunoregulatory plasma milieu that can induce immune activation in a healthy leucocyte population, as measured using a high-density microarray. Given the growing interest in the male germline as a mediator of intergenerational/transgenerational effects, I hypothesized that the IBD condition may alter the RNA profiles of the paternal human germline. Using the MARS B_1HB_2 study arm as an IBD cohort, the differences between sperm samples from control males and the B_1HB_2 study samples were determined. The control males used in the study were males from idiopathic infertile couples who produced a live birth after infertility treatment.

Relatively few REs (40 REs) were altered in at least two study visit comparisons to healthy controls (**Figure 4.6**). Interestingly, the B_2 visit of the B_1HB_2 study arm exhibits the highest number of differential REs, with the majority of those REs being unique to the B_2 visit comparison to healthy controls. Within the B_1HB_2 study arm, the Crossback visit, B_2 visit, represents the return to a low-phthalate condition after an acute exposure to high-DBP medication. Therefore, the large number of unique differential RES in the B_2 visit comparison is likely due to biological processes initiated after the high-DBP exposure. Overall, the few differential REs suggest that the IBD condition alone does not substantially alter the transcriptome of ejaculated spermatozoa. However, small RNA libraries were not available for the Control cohort. Small RNA species, such as miRNAs, piRNAs, and fragmented tRNAs, were not compared between IBD and healthy controls. Therefore, I cannot exclude the possibility that the abundance of small RNAs, which often have regulatory roles, may still be altered in IBD.
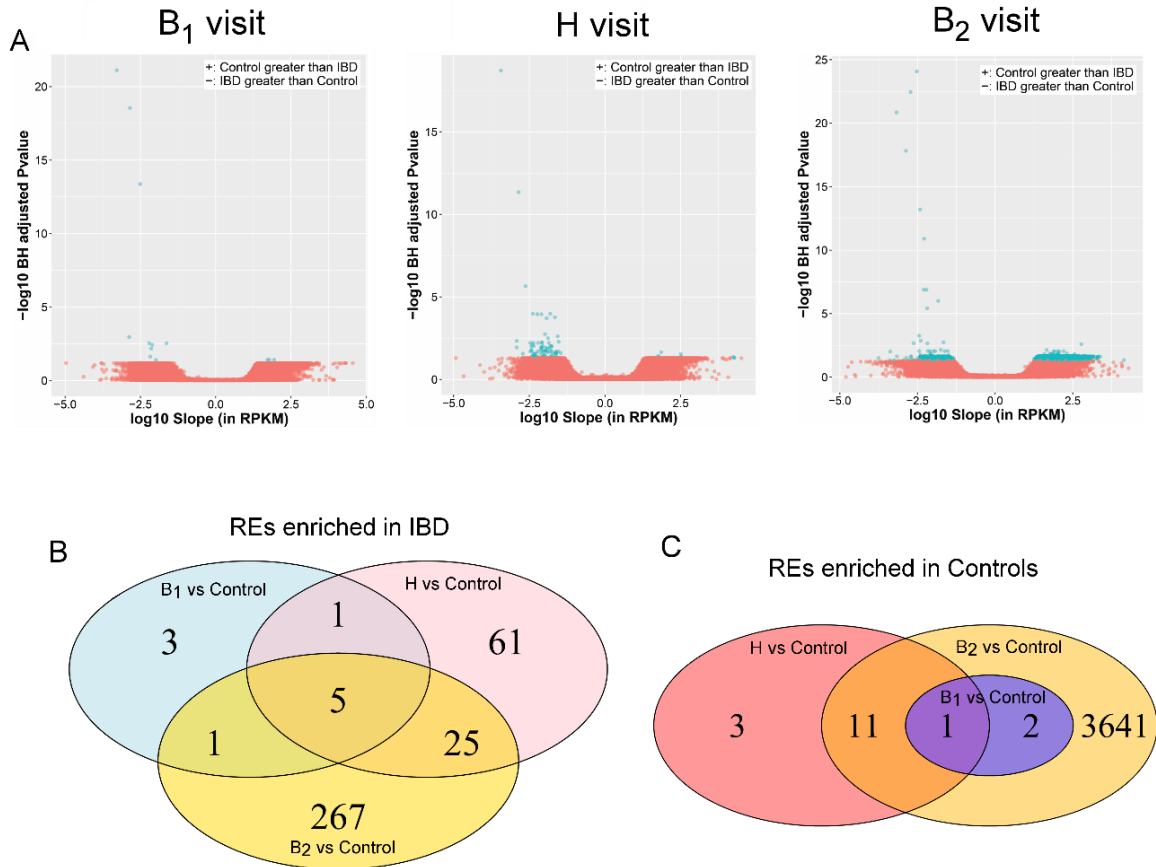
Figure 4.6. REs that differ between Normal and IBD individuals. (A) Volcano plots of linear model results for all tested REs. X-axis indicates the log10 expression change (slope) in RPKM, while the Y-axis indicates the negative log10 Benjamini-Hochberg adjusted P-value. Red points represent REs that are not statistically significant, while blue points represent REs that have a Benjamini-Hochberg adjusted P-value less than 0.05 and an absolute slope of at least 10 RPKM. REs with a positive slope exhibit higher expression in Control individuals, while those with a negative slope have higher expression in IBD individuals. (B) Venn diagram of the REs enriched in the IBD individuals, across the three study visits. (C) Venn diagram of the REs enriched in the Control individuals, across the three study visits.

Table 4.3. REs consistently altered by IBD. Genes with multiple affected REs are indicated in bold.

Expression change	Element Identifier	Gene Symbol	RE class
Enriched in IBD	chr10_49703080_49703099	C10orf53	NOVEL_INTRONIC
Enriched in IBD	chr11_22674773_22675014	GAS2	EXON
Enriched in IBD	chr1_245677221_245677520	KIF26B	NOVEL_INTRONIC
Enriched in IBD	chr15_33244895_33245438	TMCO5B	EXON
Enriched in IBD	chr1_90497041_90497080	NA	NOVEL_ORPHAN
Enriched in IBD	chr19_37551371_37551665	ZNF571-AS1,ZNF540	EXON
Enriched in IBD	chr2_178535732_178535828	TTN-AS1	EXON
Enriched in IBD	chr21_8259686_8259815	RNA5-8S5	NOVEL_10KB_EXON
Enriched in IBD	chr2_200797499_200797575	AC007163.1	EXON
Enriched in IBD	chr2_202346319_202346358	AC064836.2	NOVEL_10KB_EXON
Enriched in IBD	chr22_41174591_41174824	EP300-AS1	EXON
Enriched in IBD	chr2_30091379_30091408	AC016907.2	NOVEL_INTRONIC
Enriched in IBD	chr2_97149295_97149361	**ANKRD36**	EXON
Enriched in IBD	chr2_97151879_97151939	**ANKRD36**	EXON
Enriched in IBD	chr2_97183582_97183654	**ANKRD36**	EXON
Enriched in IBD	chr3_113805354_113805525	ATP6V1A	EXON
Enriched in IBD	chr3_47772817_47772936	SMARCC1	EXON
Enriched in IBD	chr3_52800965_52801146	ITIH3	EXON
Enriched in IBD	chr4_141231439_141231485	ZNF330	EXON
Enriched in IBD	chr4_15007305_15007595	CPEB2	EXON
Enriched in IBD	chr5_176586651_176586842	**CDHR2**	EXON
Enriched in IBD	chr5_176589031_176589182	**CDHR2**	EXON
Enriched in IBD	chr6_147001441_147001543	STXBP5-AS1	EXON
Enriched in IBD	chr9_34839387_34839416	FAM205BP	NOVEL_10KB_EXON
Enriched in IBD	chrX_126731815_126731844	MTCYBP38	NOVEL_10KB_EXON
Enriched in IBD	chrY_12915883_12916027	DDX3Y	EXON
Enriched in Control	chr15_29716611_29716838	**TJP1**	EXON
Enriched in Control	chr15_29718266_29719138	**TJP1**	EXON
Enriched in Control	chr19_10118668_10118727	EIF3G	EXON
Enriched in Control	chr19_11024331_11024438	SMARCA4	EXON
Enriched in Control	chr20_38517784_38518000	RALGAPB	EXON
Enriched in Control	chr20_410412_410764	RBCK1	EXON
Enriched in Control	chr20_62256057_62256221	OSBPL2	EXON
Enriched in Control	chr3_57268398_57268487	APPL1	EXON
Enriched in Control	chr4_42070526_42070691	SLC30A9	EXON
Enriched in Control	chr7_35872666_35872766	SEPT7	EXON
Enriched in Control	chr7_99353909_99354121	**ARPC1A,AC004922.1**	EXON
Enriched in Control	chr7_99358340_99358415	**ARPC1A,AC004922.1**	EXON
Enriched in Control	chr7_99359534_99359738	**ARPC1A,AC004922.1**	EXON
Enriched in Control	chrM_14747_15953	MT-CYB	EXON

Only 14 REs showed higher levels in Controls and 26 REs were upregulated in IBD, the majority of which were exonic REs (**Table 4.3**). However, in the IBD-enriched category, ANKRD36 and CDHR2 had two or more differential exonic REs. In the Control-enriched

category, had two or more differential exonic REs. This suggests that select genes or isoforms are consistently up-regulated or down-regulated in sperm from DBP-naïve individuals. ANKRD36 is a predicted intracellular protein, with potential nuclear localization, whose function has not yet been characterized. CDHR2 is a non-classical cadherin, which mediates formation of intermicrovillar adhesion complexes, necessary for formation of the murine intestinal brush border [259]. CDHR2, through work on cancer cell lines, has been hypothesized to act as a molecular switch for contact inhibition of epithelial cells [260]. TJP1 is a tight junction adaptor protein. ARPC1A is a subunit of the human Arp2/3 protein complex. AC004922.1 is an uncharacterized protein, formed from an ARPC1A and ARPC1B readthrough. Of the four genes (ANKRD36, CDHR2, ARPC1A, and TJP1), ANKRD36 exhibits the highest expression. However, all four genes are relatively lowly expressed in spermatozoa, with a median expression less than 25 RPKM, as ANKRD36, CDHR2, ARPC1A, and TJP1 have median expression values in sperm of 19.14, 7.32, 0.86, and 2.40 RPKM, respectively.

Of the 36 gene names up- or down-regulated in IBD, 23 of them were expressed in cultured human Sertoli cells [261], when requiring a threshold of at least 100 baseMean reads (assigned using DESeq). At this threshold, 14,843 gene names from the Ribeiro et al. dataset were considered expressed. Three of the 4 noted genes (ANKRD36, TJP1, and ARPC1A) in **Table 4.3** were among the 23 Sertoli-expressed genes. This suggests that the effects of IBD on spermatozoa may work through Sertoli cells, rather than through absorption during epididymal transit. Both CDHR2 and TJP1 are associated with cell-cell interactions, further suggesting that IBD may alter the interactions of Sertoli cell during spermatogenesis and/or spermiogenesis.

To determine if the 14 Control-enriched and 26 IBD-enriched REs were defined as differential in previous studies of other laboratories, the differential RE gene names and genes names identified as altered in previous microarray or RNA-seq studies were compared. The

exact gene names were compared, so it is important to note that if the published gene names presented in the literature were updated prior to this analysis, an overlap would not be identified. The 40 REs' gene names did not overlap with several previous studies that used colon biopsies for microarray or RNA-seq, in either human or mouse [262-270]. This is likely due to the vastly different transcriptomes between intestinal tissue types and male reproductive tissues (**Figure 4.3**). The few studies describing changes in peripheral blood [255, 258] do not provide a comprehensive list of differential genes in the publication, and so could not be compared to the gene names identified in the current study. Regardless, spermatozoa are known to have a vastly different transcriptome (compared to somatic cell types), and so likely would not exhibit overlaps with the peripheral blood studies.

Differential expression analyses of Long REs in MARS study arms

The MARS study design (**Figure 4.1**) allows for the detection of longitudinal changes occurring as a result of shifting DBP exposure. A Linear Mixed-Effects Model, when applied to the repeated longitudinal samples from the MARS dataset, can define the DBP-induced expression changes, while also taking into account any repeated samples. The LMEM model formulas were developed for defining the expression changes across study periods (e.g. Baseline, Crossover or Crossback), while considering biological and technical covariates. LMEM modeling across the MARS study arms revealed considerable differential expression (defined as empirical P-value < 0.05 and minimum absolute threshold of 10 RPKM) in each study arm, as shown in **Figure 4.7**.
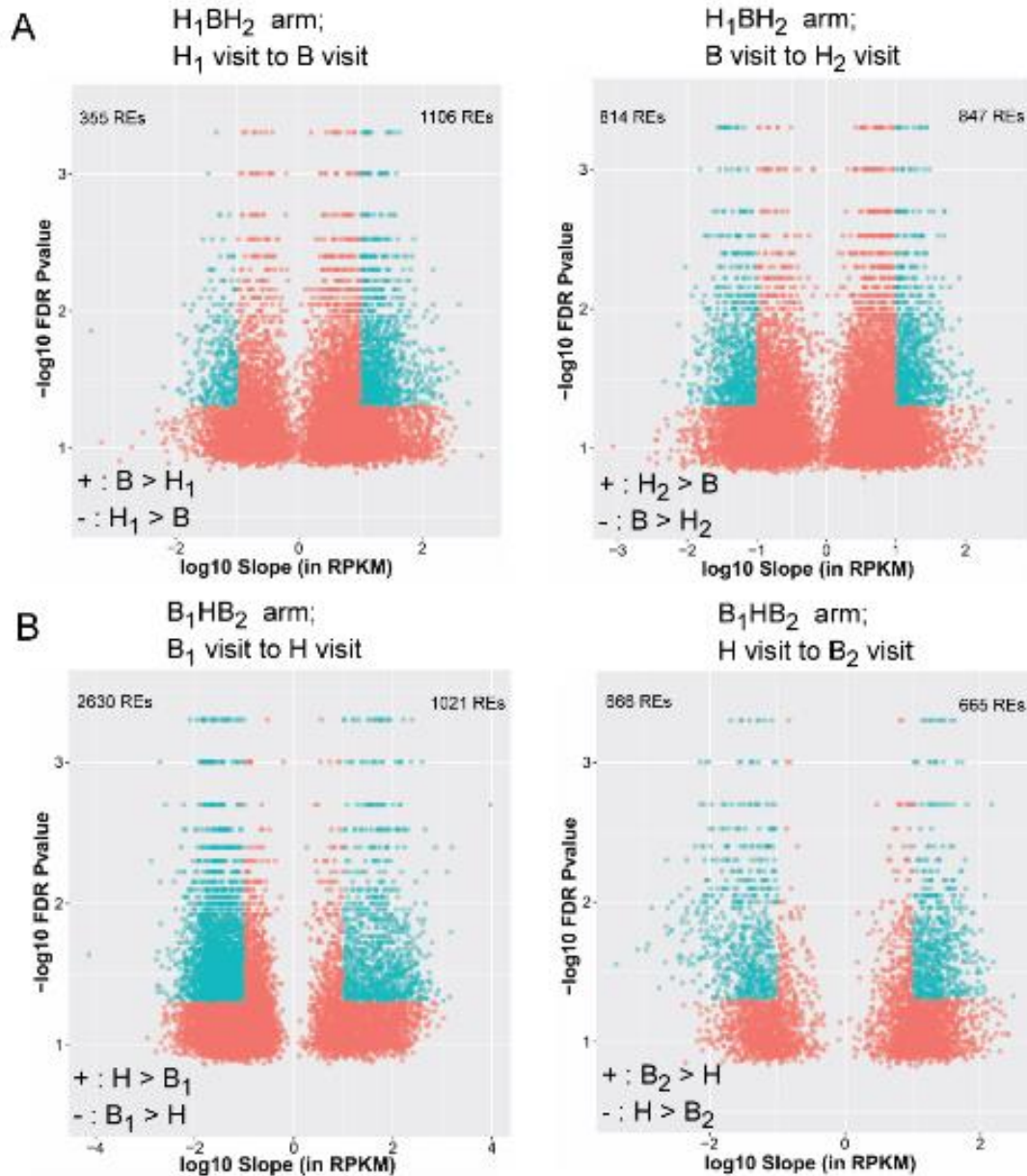
Figure 4.7. Volcano plots of REs altered across MARS study arms. (A) H_1BH_2 study arm. (B) B_1HB_2 study arm. Blue points represent REs with an empirical P-value < 0.05 and a minimum absolute threshold of 10 RPKM, while red points represent REs with empirical P-value > 0.05 or a minimum absolute threshold less than 10 RPKM.

The majority of all REs were exonic (321,207 REs), with intronic REs being the most numerous of the novel classes (near-exon: 9,730 REs; intronic: 30,853; orphan: 10,015). Correspondingly, the majority of REs altered by DBP were either exonic or intronic (**Table 4.4B**). However, novel REs, highlighted by intronic REs, were major components of all

observed transcript patterns, indicating that DBP exposure(s) affects far more than known transcripts. Few REs were differentially expressed in both study arms (H_1BH_2 arm and B_1HB_2 arm), suggesting that the alternating DBP exposures affects DBP-naïve males differently from males chronically exposed to a high-DBP mesalamine. This arm-specific effect was also observed in the sperm motility and hormonal responses of the MARS individuals [69, 70, 271].

Table 4.4. Expression patterns of REs altered across MARS study arms. (A) Total RE count for the given expression pattern. Expression patterns are presented to the left of the table, with blue lines indicating significant slopes and grey lines indicating non-significant expression changes (slopes).

Expression Patterns	B_1HB_2	H_1BH_2
	70	45
	12	105
	589	802
	646	709
	2552	310
	1003	1001
	8	0
	6	0

Expression Patterns	B1HB2 arm					H1BH2 arm				
	Exon	Novel 10kb	Novel Intron	Novel Orphan	Total	Exon	Novel 10kb	Novel Intron	Novel Orphan	Total
	44	7	15	4	70	15	3	20	7	45
	2	3	5	2	12	35	17	36	17	105
	355	45	155	34	589	377	73	286	66	802
	175	79	262	130	646	213	94	325	77	709
	1265	204	772	311	2552	88	35	133	54	310
	288	151	363	201	1003	288	131	468	114	1001
	1	1	5	1	8	0	0	0	0	0
	0	1	5	0	6	0	0	0	0	0

Differential REs were then classified into specific response patterns, defining REs which either changed explicitly with high-DBP exposure, i.e., acute response REs that changed from baseline to crossover, followed by an opposite recovery response in REs that changed from crossover to crossback, and REs that increased or decreased across all study visits. Both study arms had relatively few REs altered concomitantly with DBP exposure across both transitions (1.7% and 5% in the B_1HB_2 arm and H_1BH_2 arm, respectively). Unexpectedly, the majority of differential REs in either study arm were significantly altered in a single comparison (i.e., baseline to crossover - acute response or crossover to crossback - Recovery). Semen analysis of the MARS subjects described in Nassan et al [69], showed a continuous decline in sperm motility due to a carry-over effect of high-DBP exposure in the B_1HB_2 arm. In accord, as indicated in **Table 4.5**, several exons of sperm-motility associated genes (ATP1A4, WDR66, TEKT2, TEKT5, DRC7, CFAP44, DDX4, DNAJA1) were downregulated in the B_1HB_2 arm, during the initial high-DBP insult, consistent with the B_1HB_2 arm's observed decline in sperm motility [69]. In contrast, in the H_1BH_2 arm, semen parameters (including sperm motility) did not change across the study visits and few sperm-motility associated genes were linked with differential exonic REs in the H_1BH_2 arm. Notably, a small number of B_1HB_2 arm REs that were not directly associated with motility continuously increased (6 REs) or decreased (8 REs) across the study visits. The H_1BH_2 arm REs did not follow a continuous pattern. All 6 continuously increasing REs were novel, with 5 intronic REs (AC012531.3, WDR20, AC007993.2, RAE1, GUSBP1) and one near-exon RE (NPIPB10P). In comparison, nearly all 8 continuously decreasing REs were also novel, with 5 intronic REs (TMEM104, AC068594.1, AC016590.1, XRN1, FER), one near-exon RE (AC107958.2), one orphan RE, and a single exonic RE (PRSS21). Interestingly, PRSS21 (Serine Protease 21) is a cell-surface anchored serine protease present within elongating spermatids that may also be involved in spermatocyte development [272, 273].

Table 4.5. DBP-altered exonic REs overlapping genes associated with sperm motility. All murine genes in MGI database with the associated Gene Ontology term "sperm motility" (GO:0097722, http://www.informatics.jax.org/go/term/GO:0097722) were downloaded and transformed into the HGNC (human) gene symbol using custom R code and the BiomaRt package. The gene symbols associated with the differential exonic REs, partitioned according to expression change, were then overlapped with the list of HGNC gene symbols. All displayed gene names represent the gene symbol overlaps.

B_1HB_2 arm			
B_1 to H transition; Increased levels	B_1 to H transition; Decreased levels	H to B_2 transition; Increased levels	H to B_2 transition; Decreased levels
CATSPERD	ATP1A4	WDR66	CATSPERD
TTLL1	WDR66	GAPDHS	DNAH1
DNAH1	TEKT2	IQCF1	
	TEKT5		
	DRC7		
	CFAP44		
	DDX4		
	DNAJA1		
H_1BH_2 arm			
H_1 to B transition; Increased levels	H_1 to B transition; Decreased levels	B to H_2 transition; Increased levels	B to H_2 transition; Decreased levels
CELF3	IFT88	IFT88	TEKT5
ATP1A4		DNAI1	UBE2B

Biological Response to DBP exposure

The biological impact of DBP on REs was assessed through Gene Ontology (GO) enrichment. The gene names associated with differential exonic, novel near-exon, or novel intronic REs in the expression patterns of interest were compiled to resolve GO categories associated with both novel REs and exonic REs. **Table 4.6** summarizes the primary affected signaling and literature-based pathways. Few of the top GO pathways were shared between the two study arms, suggesting that shifts in DBP levels differentially affect DBP-naïve males compared to chronically exposed males. Within the B_1HB_2 arm, acutely downregulated REs were associated with "RAN-GAP cycling", "Focal adhesion kinase signaling", and "Ras GTPase binding". Focal adhesion kinase signaling facilitates integrin-mediated signal transduction, which has a clear role in maintaining the seminiferous tubule structure [274].

RAN-GAP cycling, which is a critical component of nucleo-cytoplasmic transport, likely is involved in epigenetic regulation during spermatogenesis, via movement of regulatory RNAs [275]. During mammalian spermiogenesis, Ran GTPase may mediate kinesin localization, which is needed for producing the unusual spermatid form [276, 277]. Interestingly, REs upregulated and downregulated in recovery of the B_1HB_2 arm were enriched for NGF signaling and EGFR signaling. NGF protein is found throughout male reproductive tissues [272, 278-280], including mammalian spermatozoa [281], and outside of the suspected regulatory roles in Sertoli cells [282], likely facilitates sperm motility [202, 283]. A disruption of the NGF signaling-mediated motility in the B_1HB_2 arm is thus in agreement with previously noted decreased motility [69] after administration of a high-DBP mesalamine to high-DBP-naïve participants. EGFR plays a regulatory role in mammalian spermatogenesis, mediating Sertoli-germ cell crosstalk [284], non-classical testosterone signaling [285], and may mediate the acrosome reaction in mature spermatozoa [203].

Table 4.6. Gene ontology enrichment summary of differential MARS REs. The gene names associated with differential exonic, novel near-exon, or novel intronic REs were compiled and used as input to Genomatix's GeneRanker function. The signaling pathways and literature-based pathways were summarized and grouped into common ontological themes.

Study arm	Acute change	Recovery
	Upregulated REs	*Upregulated REs*
B_1HB_2	Amino acid metabolism	NGF signaling
	TNF-alpha	EGFR signaling
	14-3-3 protein signaling	RAN signaling
	Downregulated REs	*Downregulated REs*
	RAN cycling	NGF signaling
	Focal adhesion kinase signaling	EGFR signaling
	Ras GTPase binding	Protein Kinase D and N
H_1BH_2	*Upregulated REs*	*Upregulated REs*
	Arf6 trafficking	Lipid metabolism
	Chromatin organization	Calmodulin
	Organelle organization	Organelle biogenesis and maintenance
	Downregulated REs	*Downregulated REs*
	Cell cycle	Cyclin D3
	Coregulation of Androgen receptor activity	Chromatin organization
	Organelle organization	Coregulation of Androgen receptor activity

Within the H_1BH_2 arm, acute response REs were associated with organellar and chromatin organization, which are requisite for extreme cellular and nuclear remodeling during spermatogenesis. The corresponding downregulated acute and recovery REs were enriched for "Coregulation of Androgen receptor activity", which is coherent with previous mammalian literature indicating the importance of testosterone in phthalate-induced testicular dysgenesis syndrome [61, 286]. However, while testicular dysgenesis syndrome reflects pre-natal

endocrine disruptor exposure, the MARS study suggests that responding to *in vivo* adult exposures also elicits androgen disruption, providing another inroad to this pathology. Adult rat models of high-DBP exposure also indicate a disruption of androgenic activity [287, 288]. Recovery in the H_1BH_2 arm suggests an involvement of "lipid metabolism" and "calmodulin". In the male, Leydig cells take up lipids for testosterone production [289], perhaps indicating a shift in steroidogenesis. Calmodulin, a calcium-binding protein activated in the presence of calcium, was previously known to be involved in mammalian sperm motility [290-292].

The above GO analysis suggests that specific signaling pathways, such as NGF signaling, EGFR signaling, RAN cycling, and Androgen receptor signaling, may be altered due to DBP exposures. To complement the above, Ingenuity Pathway Analysis (IPA) was applied to the differential exonic REs to resolve the enrichment of signaling pathways, as well as suggesting the direction of pathway modulation (**Figure 4.8**).
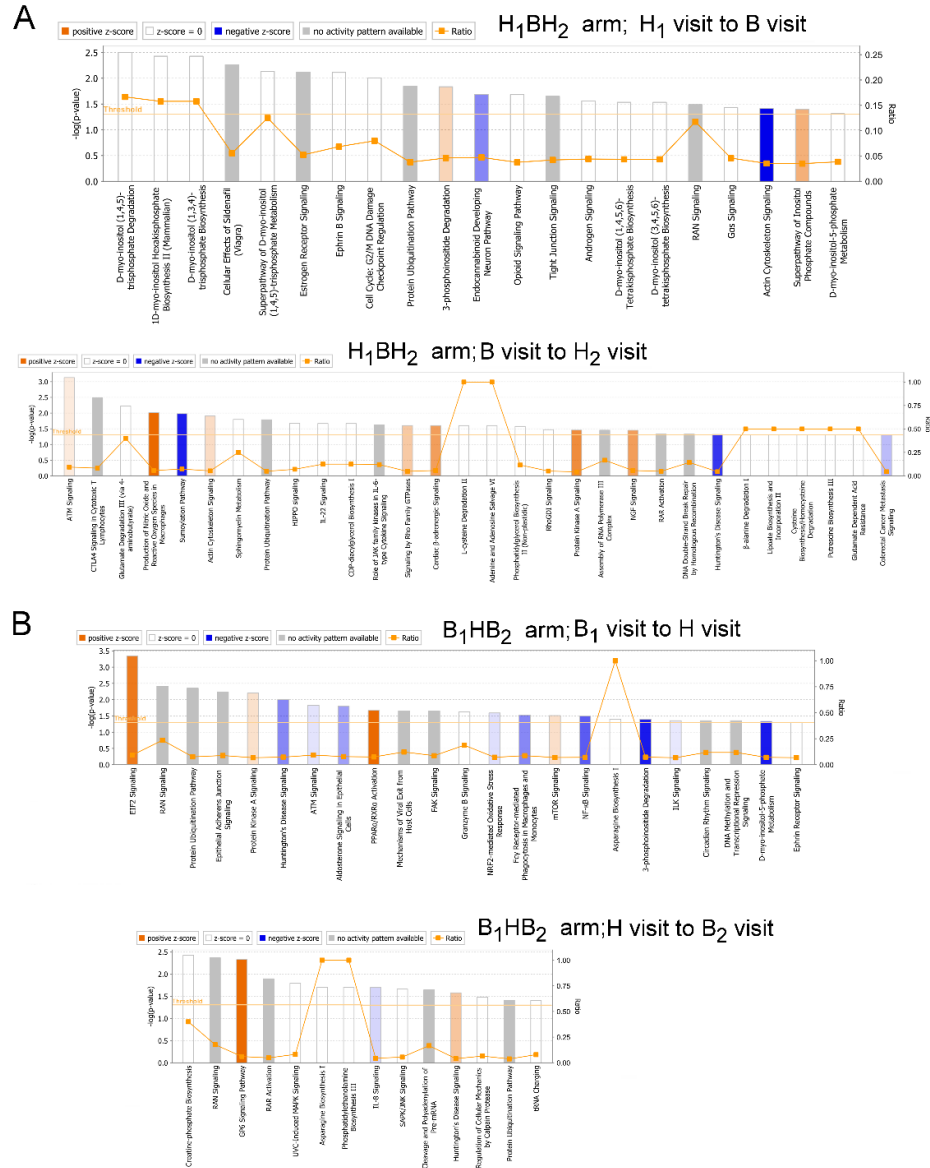
Figure 4.8. IPA pathways of REs altered across MARS study arms. (A) Pathway enrichment for H_1BH_2 study comparisons. (B) Pathway enrichment for B_1HB_2 study comparisons. Enriched pathways are ordered according to the relative significance, with most significantly enriched pathways displayed on the left. Pathways highlighted in orange hues and blue hues indicate predicted pathway activation or repression, respectively, with darker colors indicating greater confidence in the activation/repression prediction. Briefly, pathway activation/repression prediction is based on the correlation of inputted gene expression changes with the pathway's known activity patterns.

IPA signaling pathway enrichment revealed several interesting associations. Notably, in the H_1BH_2 arm, the transition from crossover (B visit) to crossback (H_2 visit), which represents the return to high-DBP mesalamine, displayed activation of oxidative stress and

DNA damage response pathways. This is consistent with previous reports of DBP-induced spermatozoal damage and oxidative stress [67, 68]. However, the B_1HB_2 arm's enriched pathways do not strongly implicate oxidative stress and DNA damage response. The transition from baseline (B_1 visit) to crossover (H visit) was associated with several spermatogenesis-related pathways, including activation of EIF2 signaling and the PPAR-alpha/RXR-alpha signaling. These pathways were not strongly associated with either the H_1BH_2 arm or in the H visit vs B_2 visit comparison of the B_1HB_2 arm, suggesting that a concerted shift in the PPAR-alpha and EIF2 pathways only occurs upon the initial high-DBP exposure. The detrimental effects of peroxisome proliferators, i.e., DBP, on germ cells likely acts through Sertoli cells [58-60, 293]. The transition from crossover (H visit) to crossback (B_2 visit) yielded a strong activation of GP6 signaling [294, 295]. Additionally, the retinoic acid receptor (RAR) pathway, which is a known mediator of germ cell differentiation [296], was also enriched, and although no concerted activity was observed, levels of several of the altered pathway members (CARM1, SWI/SNF, NCOR1, PKC) were consistent with an activation of the Retinoic acid nuclear receptor (RAR) and Retinoid receptor (RXR) (via binding of retinoic acid). Notably, several of the RAR's altered pathway members (CARM1, SWI/SNF, NCOR1) associated with changes in chromatin structure [43, 297-300] have the potential to mediate epigenetic effects of intergenerational inheritance.

Upstream regulators of DBP-altered genes

In addition to pathway enrichment, IPA provided the relative enrichment and associated activation/repression states for the upstream regulators of differential genes. The enriched upstream regulators are shown in **Table 4.7**. The upstream regulators are not present in the seminal plasma proteome, from Barrachina et al. [301]. In the spermatozoal proteome, only Myc and MAPK9 are present [43]. All upstream regulators indicated in **Table 4.7** were lowly expressed across the MARS samples, with a median expression value less

than 1 RPKM. Only two genes, EGLN2 and FOXM1, had a median expression value

exceeding 1 RPKM, with values of 2.22 and 2.25, respectively (**Table 4.8**).

Table 4.7. Upstream regulators from IPA. Positive Z-scores are indicated in shades of red, while Negative Z-scores are indicated in shades of blue. Deeper shades of the given color represent more extreme Z-scores. All presented Z-scores represent a statistically significant (p<0.05) enrichment of the regulator's associated genes. A negative Z-score indicates that the actions of the upstream regulator are predicted to be repressed.

Upstream regulators	$B_1HB_2\_B_1\_to\_H$	$B_1HB_2\_H\_to\_B_2$	$H_1BH_2\_H_1\_to\_B$	$H_1BH_2\_B\_to\_H_2$
MYC	-3.65	N/A	N/A	N/A
CST5	1.71	0.00	1.00	-0.50
ERBB2	N/A	N/A	-1.00	-1.55
EGLN	2.14	N/A	N/A	N/A
REST	N/A	N/A	-1.98	N/A
HR	N/A	N/A	N/A	-1.41
MAPK9	N/A	N/A	-1.07	N/A
TP53	N/A	N/A	-1.00	N/A
CCND1	N/A	N/A	N/A	-1.00
mir-10	N/A	N/A	N/A	-0.97
miR-122-5p (miRNAs w/seed GGAGUGU)	N/A	N/A	N/A	0.45
CDK4/6	N/A	-0.45	N/A	N/A
FOXM1	N/A	N/A	-0.33	N/A

Table 4.8. Summarized expression values of IPA's upstream regulators. The column "Gene Symbol" indicates the gene name, while the columns "Median expression" and "Mean expression" indicates the median and mean expression value, respectively, across the MARS samples.

Gene Symbol	Median expression	Mean expression
MYC	0	0.32283
CST5	0	0.018956
ERBB2	0	0.001302
EGLN1	0	0.111128
EGLN2	0	1.423787
EGLN3	0	0
REST	0	0
HR	0	0.00872
MAPK9	0	0.041987
TP53	0	0.083205
CCND1	0	0.55502
CDK6	0	0.003696
CDK4	0	0.1018
FOXM1	0	0.544666

In the current regulatory analysis, MYC, a well known transcription factor that acts as a proto-oncogene, has a negative Z-score (-3.65), suggesting that the actions of Myc are repressed by the transition to high DBP in the B_1HB_2 arm. This effect does not carry over to any of the remaining three comparisons. Higher Myc expression is associated with increased cell growth [302]. In mouse, spermatogonial stem cell renewal is regulated by Myc/Mycn-mediated glycolysis [303].

Interestingly, Cystatin D (CST5 gene) was found to be enriched in all four comparisons. In the B_1HB_2 arm, CST5 was moderately activated in the transition to high DBP, and having no predicted activation or repression during the return to low DBP. In the H_1BH_2 arm, CST5 was lowly activated during the transition to low DBP, and lowly repressed during the return to high DBP. The role of Cystatin D in spermatogenesis is unknown, and is not highly expressed in testis. However, CST5, as a secreted cysteine proteinase inhibitor, may play a protective role against proteinases [304].

Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2), commonly known as HER2, is a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases. It was predicted to be mildly repressed in the H_1BH_2 arm comparisons, suggesting a continual repression across the changing DBP levels.

EGLN is a family of prolyl hydroxylase enzymes, with canonical HIFα targets, which act as an ancient oxygen-monitoring machinery [305]. In the B_1HB_2 arm, the transition to high DBP was concurrent with a predicted moderate activation of EGLN. REST (RE1 Silencing Transcription Factor), a transcriptional repressor, and HR (Lysine Demethylase And Nuclear Receptor Corepressor) were both predicted to be repressed in the H_1BH_2 arm, with REST enriched in the transition to low DBP and HR enriched in the return to high DBP. HR is a histone demethylase, which acts to demethylate mono- and dimethylated Lys-9 of histone H3. Although the B_1HB_2 arm's enriched pathways do not strongly implicate oxidative stress and

DNA damage response, HIF is integral to sensing and responding to hypoxia. Uncontrolled hypoxia can result in oxidative stress.

Interestingly, two microRNAs, mir-10 and mir-122-5p, were enriched upstream regulators during the return to high DBP in the H_1BH_2 arm. While neither microRNA was itself altered by DBP exposure, shown in the following section "*Small RNAs altered by DBP exposure*", mir-10 was predicted to be repressed, while mir-122-5p was predicted to be mildly activated, as shown in **Table 4.7** and **Figure 4.9**. The mir-10a-5p, mir-10b-5p, and mir-122-5p were all well expressed in human sperm, with median expression values of 105.5, 191.7, 170.6 RPM, respectively. The 3p sections of mir-10 and mir-122 were poorly expressed, being zero RPM in nearly all samples. Reference values from SpermBase (http://spermbase.org), which used a different sequencing platform (Ion Proton system), also indicate that mir-10a-5p, mir-10b-5p, and mir-122-5p are well expressed in human sperm [306]. mir-10a-5p is expressed at 57 RPK and 91.3 RPK in whole sperm from human and mouse, respectively. mir-10b-5p is expressed at 31.5 and 65.8 RPK in whole sperm from human and mouse, respectively. mir-122-5p is expressed at 88.2 RPK in human whole sperm, but is poorly expressed in mouse sperm (0.7 RPK).
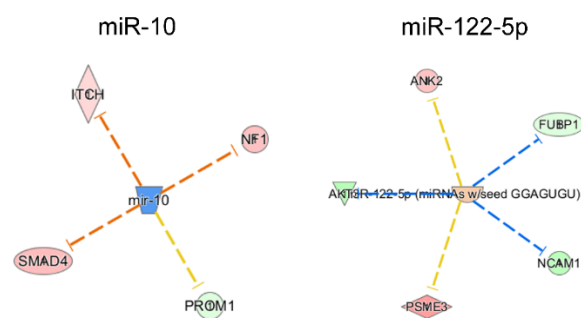
Figure 4.9. Downstream effectors of mir-10 and mir-122. Differential downstream regulators of (A) mir-10 and (B) mir-122 are shown.

DBP exposure promotes expression of simple repeats

Several classes of repetitive element associated RNAs, including simple repeats, endogenous retroviruses, and centromeric RNAs, have been identified within the population

of human spermatozoal RNAs [156, 245]. The effect of high-DBP exposure in spermatozoa for each study arm and study visit was assessed as a function of relative enrichment/depletion of REs that overlapped a genomic repeat. As shown in Estill et al., centrometric repeats and MER1A were enriched in mature human spermatozoa [245]. As expected, the centromeric repeat, (AATGGAATGG)n, was enriched across all MARS study arms (see **Figure 4.10A**), with no distinct differences across the study arms. This centromeric enrichment was primarily due to novel orphan REs. Interestingly, the abundance of differential REs overlapping the (AATGGAATGG)n repeat decreased from the B_1 visit to the H visit. As shown in **Figure 4.10B**, this suggests that in the B_1HB_2 arm, the transition to the high-DBP exposure reduces the levels of (AATGGAATGG)n-associated REs. Centromeric RNA has been shown to facilitate the localization of nucleoproteins and the chromosomal passenger complex (CPC) [307, 308]. Interestingly when nuclear structures are resolved, sperm centromeres are located towards the nuclear periphery [309]. This is consistent with the view that centromeric repeat RNA may represent residual transcripts that, in some manner, guide sperm differentiation and/or guide mitotic progression of the early human embryo.

Simple repeats, such as GA-rich repeats and variations of TC-rich repeats (e.g. (TTTC)n) were highly enriched across many of the MARS sperm samples as shown in **Figure 4.10C**. Interestingly, simple repeats were highly enriched in all study visits and arms, with the exception of the B_1 visit of the B_1HB_2 study arm. Within the MARS study set, the B_1 visit is the sole timepoint for which sperm samples have not been exposed to high-DBP mesalamine. All novel RE classes (near-exon, intronic and orphan), but not the exonic REs, exhibit this DBP-specific pattern of repeat enrichment. Differential repeat analysis verified this DBP-specific pattern for several of the simple repeats (**Appendix I**), with the primary exceptions of GA-rich and A-rich repeat classes.

These results suggest that high-DBP exposure elicits an immediate and acute response, represented by a dramatic increase in the expression of simple repeats in the male

gamete. This showed that, in addition to being enriched in spermatozoa, RE-associated genomic repeats are selectively modified by DBP exposure. As these repeats are compartmentalized in sperm [156], perhaps they also have a role in sperm chromatin organization. In this manner, their modification by DBP, that is known to increase DNA nicking [310], may specifically alter chromatin states [311].
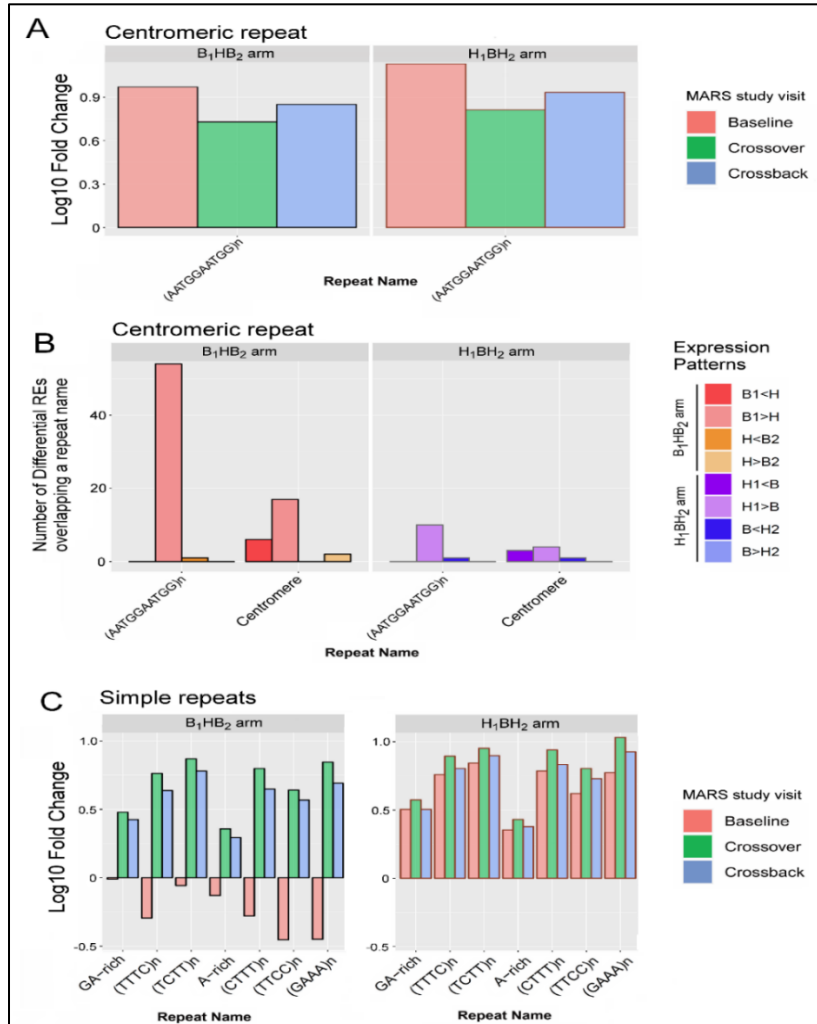
Figure 4.10. Enrichment of repetitive element expression. (A) Repeat enrichment (positive log10 fold change) or depletion (negative log10 fold change) of centromere-associated repeats. X-axis provides the repeat name, while the Y-axis indicates the relative enrichment (positive log10 fold change) or depletion (negative log10 fold change). (B) The number of differential REs that overlap centromeric repeats. X-axis provides the repeat name, while the Y-axis indicates the number of differential REs for each significant expression change. (C) Repeat enrichment (positive log10 fold change) or depletion (negative log10 fold change) of Simple repeats. High-DBP exposure within the past spermatogenic cycle enriches simple repeats in spermatozoa.

Small RNA composition of spermatozoa

Several types of small RNAs, such as miRNAs and piRNAs, have known roles in regulating mRNA and transposable element-derived RNAs and levels of repetitive elements. piRNAs have been proposed to play a regulatory role in spermatogenesis and the mammalian embryo, perhaps as part of the confrontation-consolidation of the embryo. miRNAs are well-known to regulate their target genes through mRNA degradation and translational repression. Therefore, miRNAs delivered to the embryo may play a regulatory role in regulating the early use of maternal RNAs before zygotic genome activation. Among the other small RNA species available in the mammalian genome (e.g. snoRNAs), tRNA fragments delivered by sperm have been suggested to regulate the endogenous retroelement MERVL in the murine embryo [71].

Common small RNA species of interest, such as miRNAs, piRNAs, tRNAs, tRNA fragments, and siRNA can be detected with small RNA libraries (miRNAs ~22 bp) [250], piRNA ~ 24-31 bp [251], and tRNA fragments ~28- to 34-nt [71]). Accordingly, small size-selected RNA (sncRNA)-Seq libraries, were prepared and sequenced to assess the impact of DBP exposure. **Figure 4.11** shows the expected lengths of several small RNA species, in context of an exemplar MARS small RNA sample. Several of the small RNA libraries show peaks around the expected miRNA, piRNA, and tRNA fragment (tRF) sizes. As an initial analysis to determine if small RNA species have different distributions between the two study arms, the read sizes were binned into approximately 5 bp segments, for a total of 6 bins. **Figure 4.12** shows that the two study arms and their individual visits (Baseline, Crossover, Crossback) were not noticeably different.
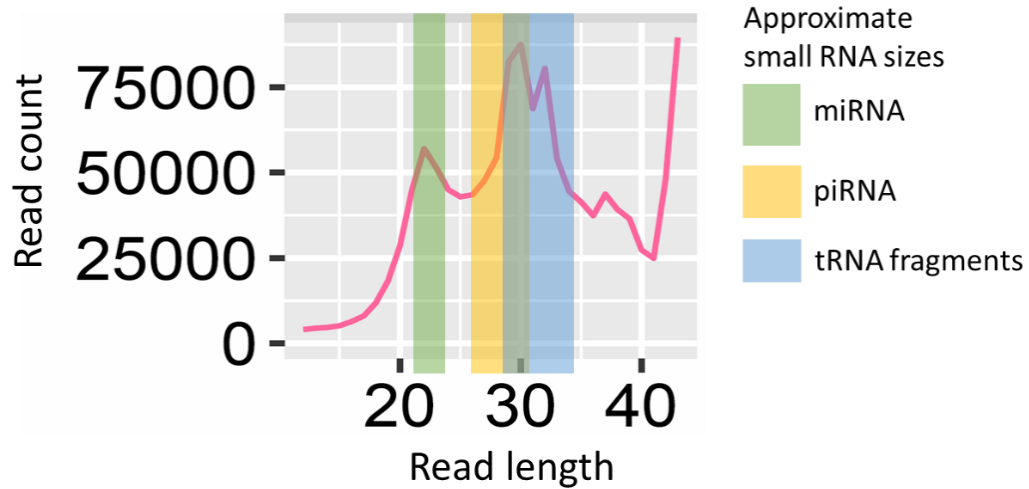
Figure 4.11. Expected read lengths of small RNA species. The read length distribution of an example sample is used as background. The Y-axis provides the read counts for a given read length, while the X-axis indicates the read length. Expected miRNA, piRNA, and tRNA fragment lengths are shown in green, yellow, and blue, respectively.
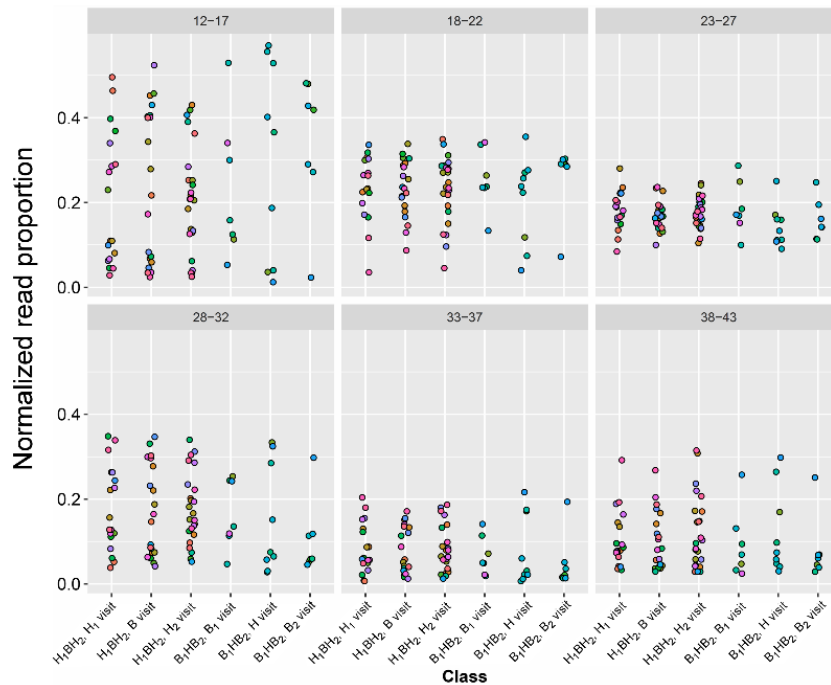
Figure 4.12. Small RNA read length comparisons. (A) Read count distributions. Y-axis indicates the read count for each sample, and the. (B) Read count proportions. The Y-axis indicates the proportion of a given sample belonging to the read length bin of interest. X-axis indicates the study arm and study visit. Points are colored according to the patient that the sample was sourced from. Panel headers indicate the read lengths included in the given panel.

Several previous studies have assessed the small RNA component of mammalian sperm. In order to compare the small RNAs in the human sperm from the MARS study, the highly expressed small RNAs were assessed, as shown in **Table 4.9 and Figure 4.13**. The top 50 small RNAs were primarily tRNAs (15/50, 30%), piRNAs (13/50, 26%), rRNA (8/50, 16%) and miRNAs (7/50, 14%). When considering all small RNAs that were highly expressed (using an arbitrary threshold of median RPM exceeding 50 RPM), 153 small RNAs were identified as highly expressed. As with the more limited "Top 50" set shown in **Table 4.9**, the highly expressed set was dominated by piRNAs (42/153, 27%), miRNAs (37/153, 24%), tRNAs (25/153, 16%), small nuclear RNAs (16/153, 10%), and rRNA (8/153, 5%), as shown in **Figure 4.13**.
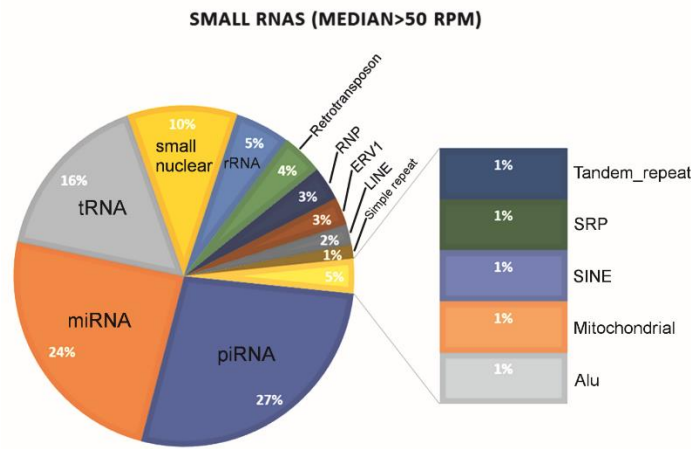
Figure 4.13. Distribution of small RNA families in highly expressed small RNAs. Highly expressed small RNAs were defined as having a median RPM exceeding 50 across all MARS sperm samples.

Table 4.9. Top 50 small RNAs in MARS small RNA libraries. "small RNA ID" indicates the small RNA species, while the adjacent "Median RPM" column indicates the median expression value in RPM, across the MARS small RNA libraries.

small RNA ID	Median RPM	small RNA ID	Median RPM
LSU-rRNA_Hsa	419782.8	ENSG00000283293|ENST00000636484	1073.7
SSU-rRNA_Hsa	109860.6	tRNA-Met	1012.6
chrM	77858.1	tRNA-Met-i	1012.6
hsa_piR_016735	20128.9	tRNA-Ser-AGY	933.3
TRNA_GLY	16437.5	hsa_piR_019914	855.5
5S	11970.6	hsa-let-7b-5p	755.5
LSU-rRNA_Cel	11908.2	hsa_piR_008114	710.2
SSU-rRNA_Dme	9280.0	hsa_piR_008113	677.8
hsa_piR_000823	8419.9	tRNA-Leu-CTA_	677.0
LSU-rRNA_Dme	8214.1	hsa-miR-375	669.8
hsa_piR_000765	6957.7	tRNA-SeC(e)-TGA	567.8
Y4	6693.5	tRNA-Val-GTA	514.6
hsa_piR_020326	5798.1	tRNA-Leu-CTA	507.5
tRNA-Asp-GAY	5675.3	hsa_piR_019825	440.1
TRNA_GLU	5505.7	hsa_piR_006046	431.5
HY1	3908.0	hsa-miR-21-5p	428.2
tRNA-Leu-CTY	3381.0	U4B	427.9
7SL	3332.4	hsa-let-7a-5p	416.5
SSU-rRNA_Cel	2705.0	tRNA-Pro-CCA	393.9
RRNA45	2521.4	hsa-miR-26a-5p	352.8
hsa_piR_004153	2288.6	hsa-miR-30a-5p	324.1
LOR1I	1550.4	hsa_piR_015249	302.7
TRNA_VAL	1411.4	tRNA-Leu-TTG	295.5
tRNA-Lys-AAG	1279.6	hsa-miR-191-5p	288.2
hsa_piR_017716	1090.8	hsa_piR_009294	288.0

MARS small RNA populations were largely concordant with the small RNA proportions in human sperm indicated by Donkin et al., who found that when rRNAs were excluded, piRNAs, tRNAs, and miRNAs comprised a large proportion of the remaining small RNAs [46]. While Donkin et al. identified 37 piRNAs with a false discovery rate (FDR) below 0.1, those 37 piRNAs do not overlap with the piRNAs identified above as being highly abundant. This likely reflects the different experimental paradigms (i.e phthalate exposure and bariatric surgery for weight management).

Work in mice has previously shown an abundance of tRNA fragments in mature sperm [312], with an excess of fragments generated from tRNA-GLY and tRNA-GLU (Glycine and

Glutamic Acid). In human sperm, I find tRNA-GLY and tRNA-GLU to be the 1st and 3rd most highly abundant tRNAs, respectively, as shown in **Table 4.9**. As noted in Chen et al., sperm transfer RNA–derived small RNAs (tsRNAs) and miRNAs exhibit expression changes after High-Fat Diet (HFD) in a paternal mouse model, and parallel inherited metabolic disorders in offspring. The noted highly expressed tsRNAs in sperm (which accounted for ~70% of sperm tsRNAs) were derived from tRNA-GLU, tRNA-GLY, and tRNA-VAL (Glutamic Acid, Glycine, and Valine) [72], which correspond to the 3rd, 1st, and 5th most abundant tRNAs in the current study. Additional murine sperm analysis [71] indicates that tRNA fragments are accumulated by maturing sperm during epidydmal transit. Paternal protein restriction affects caudal sperm's tRNA fragment levels and miRNA levels. Notably, tRNA-GLY, which is highly expressed in sperm, is more abundant in protein restricted mice. tRNA-GLY fragments delivered by sperm have been proposed to regulate the target genes of the MERVL endogenous retroelement in the murine embryo.

Previous work on spermatozoal small RNAs (18–30 nucleotides) by the Krawetz laboratory, using ejaculates of three fertile individuals [313], suggested that between 20 and 60% of the sequenced reads were donor specific, indicated extensive sample heterogeneity in sperm's small RNAs. The MARS small RNAs were analyzed for expression specificity, and revealed that of the 12,779 small RNAs, 48% were expressed at a level exceeding 5 RPM in only a single sample, as shown in **Figure 4.14**. Overall, this suggests that a majority of the small RNAs measured in the MARS study will exhibit extensive sample specificity, regardless of the expression threshold used.
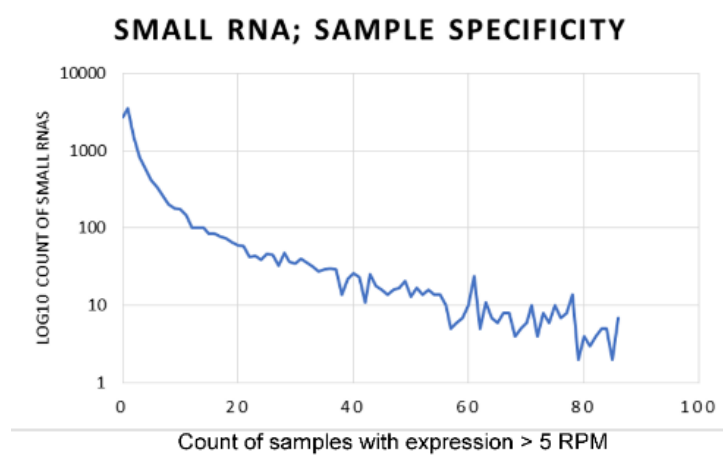
Figure 4.14. Heterogeneity of small RNAs. For 12779 small RNAs, the Y-axis indicates the count (in log10) of the small RNAs, and the X-axis indicates the count of small RNA libraries with an expression value exceeding 5 RPM.

piRNA cluster expression

Small RNAs have a known role in genomic imprinting and establishment of the embryo in plant species [314]. Given this suspected importance of piRNAs in gene regulation and confrontation/consolidation within the embryo, the expressed piRNAs were examined to determine if any particular piRNA clusters were transcribed. piRNA alignments were obtained from https://www.pirnadb.org/ for hg38. piRNA cluster information was obtained from the piRNA cluster database (http://www.smallrnagroup.uni-mainz.de/piRNAclusterDB.html), using the pooled generic testis dataset from Homo sapiens [183, 315]. Most piRNAs have alignments in relatively few parts of the genome, with a few piRNAs occurring many times (>100 times) in the human genome. A total of 14 piRNA clusters had two or more piRNAs present (median RPM > 1 RPM) in the sperm small RNAs, as shown in **Appendix J**. Interestingly, two clusters (on different chromosomes) had 6 expressed piRNAs, suggesting that the two clusters may be active in late human spermatogenesis.

Small RNAs altered by DBP exposure

A LMEM was applied to the individual study arms to assess the impact of DBP exposure on small spermatozoal RNAs under high-DBP and background-DBP conditions

across the entire arm. This approach, did not differentiate between study visits (e.g. baseline, crossover, and crossback), but instead treated samples as replicates of the associated high-DBP or background-DBP conditions. This use of this approach was necessary due to small sample sizes of the individual study visits. Both the use of an empirical P-value and Benjamini-Hochberg correction produced similar numbers of significant small RNAs. For concordance with adjustment strategy employed in the long RNAs, an empirical P-value was used for the small RNAs. As shown in **Figure 4.15** and detailed in **Appendix K**, in comparison to non-DBP medication, the B_1HB_2 arm showed upregulation of 3 small RNAs in response to high-DBP mesalamine and downregulation of 12 small RNAs. In the H_1BH_2 arm, exposure to high-DBP mesalamine upregulated 9 small RNAs and downregulated 77 small RNAs. The difference in detection of differential small RNAs between study arms was likely due to the smaller sample size of the B_1HB_2 arm. CHARLIE3, a hAT-Charlie DNA transposon, and hsa_piR_019675 were differentially regulated in both study arms. Of note, CHARLIE3 was upregulated in the B_1HB_2 arm, yet down regulated in the H_1BH_2 arm upon high-DBP exposure. In contrast, hsa_piR_019675 was down regulated in both study arms upon high-DBP exposure. Although the biological significance of hsa_piR_01967 is not yet known, its genomic loci overlap several SSU-rRNA loci, suggesting that this piRNA may serve a regulatory role for rRNA [316].
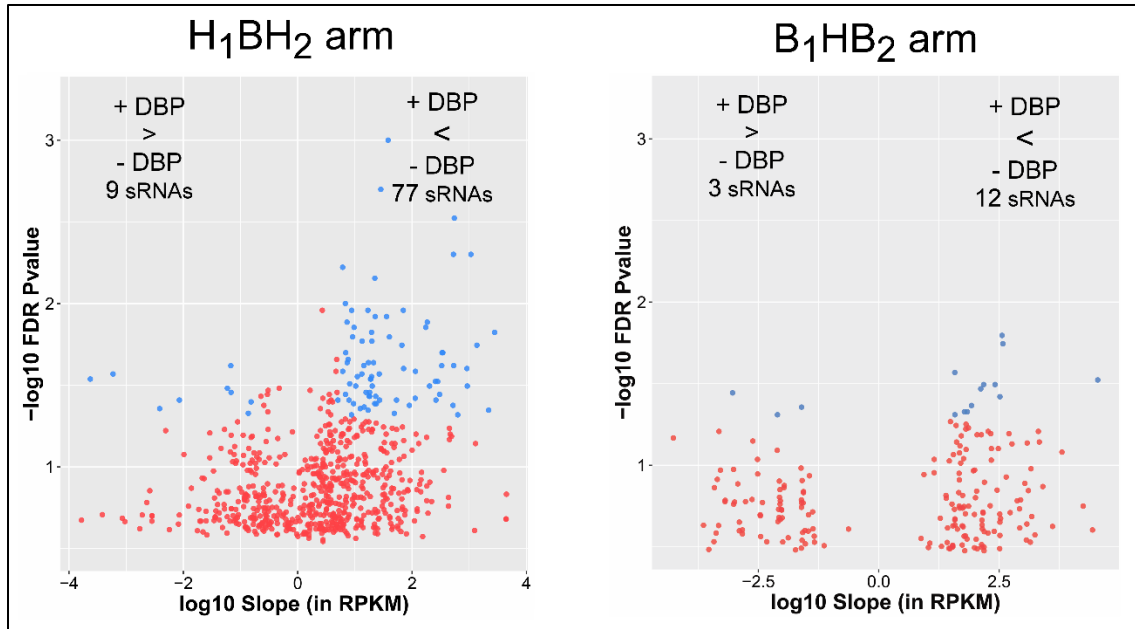
Figure 4.15. Volcano plots of differential small RNAs. The left and right panels show the volcano plots for the H_1BH_2 arm and the B_1HB_2 arm, respectively. The X-axis indicates the log10 expression change (slope) in RPM, while the Y-axis indicates the negative log10 empirical P-value.

Regulatory roles of small RNAs

Small RNAs, e.g., miRNAs and piRNAs, are known to be regulators of transposable elements and mRNA expression [317-321]. Relatively few small RNAs were modified by DBP exposure (**Figure 4.15** and **Appendix K**), and do not appear to underly gene expression changes observed in the long RNAs. However, given the presence of both complex and simple genomic repeats that are compartmentalized within the sperm [156], their association with sncRNAs was considered. The association of sncRNAs with genomic repeats were assessed by correlation in each individual study arm. The use of a series of expressed small RNAs that are physically linked to repeat sequences within the small RNA libraries enabled their exact positioning. Independent of the study arm, greater than 90% of the small RNA Spearman correlations were positive. In comparison, within the H_1BH_2 arm, 6.8% (37,098 of 548,628) and the B_1HB_2 arm 8.4% (18,880 of 224,115) of small RNA correlations were negative, potentially reflecting inhibition.

The majority of the negative small RNA pairs were concordant between both study arms, suggesting that their effect remains constant across exposures (**Figure 4.16**). For example, RNA pairs shared between the B_1HB_2 arm and H_1BH_2 arm include piRNA:rRNA, miRNA:rRNA, rRNA:snRNA, rRNA:SINE1/7SL, rRNA:tRNA, ERV1;rRNA, while the B_1HB_2 arm also included ERV3:rRNA, L1:rRNA, rRNA:snoRNA, Endogenous Retrovirus:rRNA, and ERV2:rRNA. Interestingly, the small rRNA fragments were negatively correlated with many of the other small RNAs (piRNA, miRNA, snRNA, tRNA, snoRNA) and genomic repeats (SINE1/7SL, ERV1, ERV2, ERV3, L1, Endogenous Retrovirus). piRNAs exhibited a series of strongly negative correlations with Endogenous Retrovirus, L1, ERV1, and Simple repeats, likely reflective of a suppressive role, while miRNAs were negatively correlated with Endogenous Retrovirus. This is consistent with the view that piRNAs and miRNAs, either directly or indirectly, act to suppress RNAs generated from Endogenous Retroviruses.
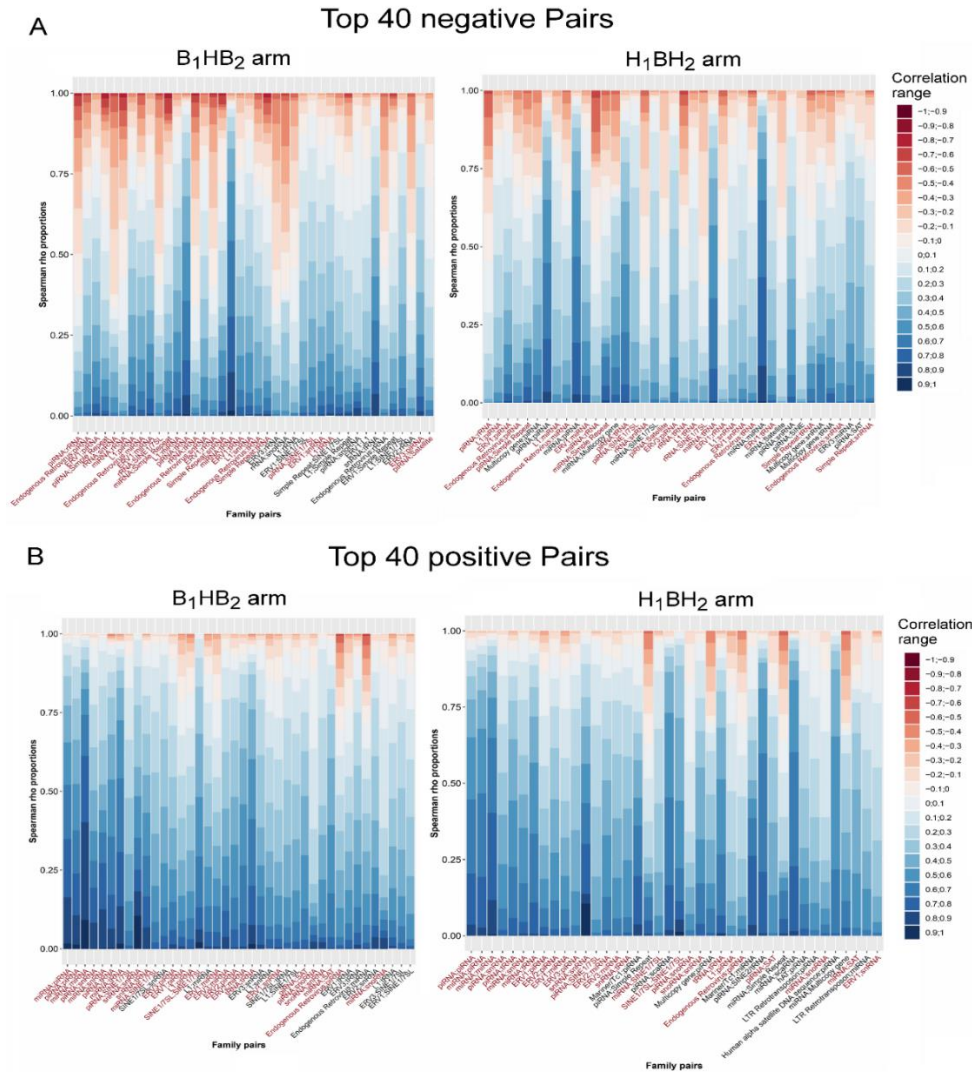
Figure 4.16. Top 40 positive and negative small RNA pairs. (A) The RNA pairs with the highest count, among the RNA pairs with a Spearman rho less than -0.2, are shown in order of most counts to least counts, from left to right. The proportion of each RNA pair belonging to given correlation range (e.g. a bin of 0.8;0.9 contains correlations with a rho greater than or equal to 0.8 and less than 0.9.) are stacked along the Y-axis and colored according to the correlation range. The RNA pairs in the top 40 negative pairs for both study arms (B_1HB_2 arm and H_1BH_2 arm) are indicated in red text along the X-axis. For example, piRNA;rRNA contains the greatest number of negative correlations in both study arms and is thus shown on the leftmost side of each of the figures in panel A. The dominance of negative correlations is displayed as a series of stacked bars in red shading. (B) The RNA pairs with the highest count, among the RNA pairs with a Spearman rho greater than 0.5, are shown in order of most counts to least counts, from left to right. The proportion of each RNA pair belonging to given correlation range (e.g. a bin of 0.8;0.9 contains correlations with a rho greater than or equal to 0.8 and less than 0.9.) are stacked along the Y-axis and colored according to the correlation range. The RNA pairs in the top 40 positive pairs for both study arms (B_1HB_2 arm and H_1BH_2 arm) are indicated in red text along the X-axis.

The interaction of long RNA repetitive elements with small RNAs

To assess the veracity of the above, genomic repeat associations were extended to the long RNA libraries, thus excluding the possibility that the interactions were merely observed in fragmented RNAs. As shown in **Table 4.10**, the majority of repetitive elements in human sperm were simple repeats, followed by ERV1, tRNA, and ERVL. This is in accord with both the repetitive elements observed in small RNA libraries (**Figure 4.13C**) and previous repeat enrichment analyses through spermatogenesis to zygotic genome activation [156, 245].

Table 4.10. Overall repeat expression in long RNA libraries. "Repeat Family" column indicates the repeat class. "Total" column indicates the total number of REs that overlap genomic repeats from the given repeat class. "Median >1 RPM" and "Median >5 RPM" indicate the number of REs overlapping the given genomic repeat that have a median expression value greater than 1 RPM or 5 RPM, respectively.

REPEAT FAMILY	Total	B_1HB_2 ARM		H_1BH_2 ARM	
		Median>1 RPM	Median>5 RPM	Median>1 RPM	Median>5 RPM
SIMPLE_REPEAT	3292	578	315	582	322
ERV1	285	27	19	26	14
TRNA	46	15	9	16	11
ERVL	118	6	4	6	4
UNKNOWN	39	5	2	4	1
ERVL-MALR	78	5	3	4	2
ERVK	38	4	2	4	2
DNA?	18	3	2	3	2
HAT-CHARLIE	73	3	1	3	0
SATELLITE	10	2	2	2	2
RRNA	3	2	2	2	2
HAT?	6	2	2	2	2
SCRNA	5	2	1	2	1
L1	129	1	1	1	1
LTR	5	1	1	1	0
GYPSY	17	1	0	1	0
HAT-TIP100?	3	1	0	1	0
MERLIN	1	1	0	1	0
DNA	11	1	0	0	0
L2	9	1	0	0	0

The combined small RNA:long RNA correlations primarily yielded positive correlations (71.2% and 73.1% in the B_1HB_2 arm and H_1BH_2 arm, respectively), concordant with the small RNA-only correlations (small RNAs:small RNAs). As previously observed for the small RNA-only analysis, the RNA pairs with the greatest number of negative correlations or positive correlations were largely shared between the two study arms. However, unlike the small RNA-only analysis (**Figure 4.16**), nearly all of the top 40 negative pairs exhibited a high background of positive correlations, among comparatively fewer negative correlations.

piRNAs can act as direct regulators of transposable elements [317] and miRNAs also have indirect roles in modulating transposons and direct roles in modulating mRNA [318-321]. Both miRNA and piRNA were primarily correlated to long RNA repetitive elements in a positive manner (**Figure 4.17**). However, a small portion of the miRNA pairing to Simple repeats, ERV1, ERVL, and tRNA were highly negative. Similarly, a small portion of the piRNA pairings to Simple repeats, ERV1, ERVL, ERVL-MaLR, and tRNA were highly negative. Interestingly, despite a similar read coverage of the rRNA locus by long and short RNA libraries, rRNA fragments in the long >200 bp RNA libraries were largely positively correlated with piRNA, miRNA, and scaRNAs (**Figure 4.18**).
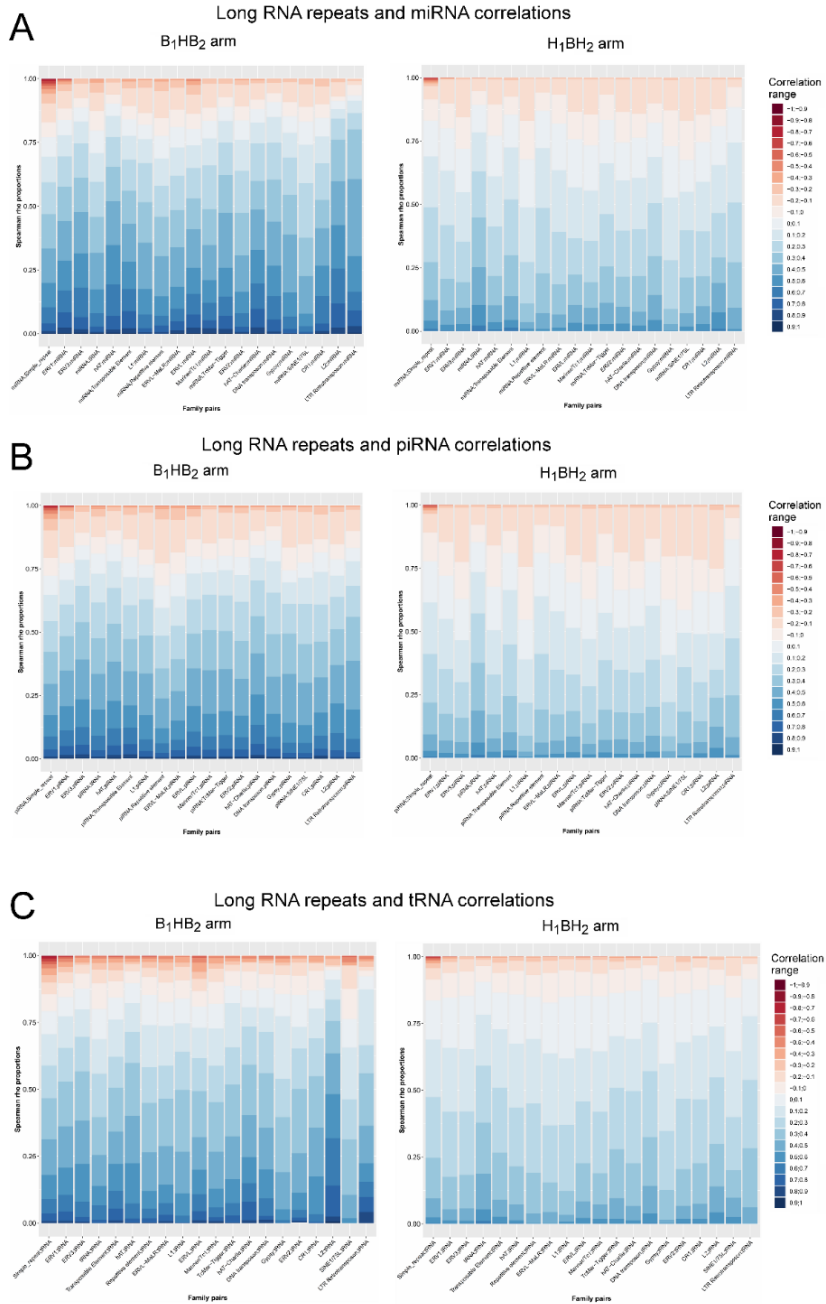
Figure 4.17. miRNA, piRNA and tRNA correlations to genomic repeats in long RNA libraries. (A) Pairs for which a miRNA was correlated with a repeat in the long RNA samples. (B) Pairs for which an piRNA was correlated with a repeat in the long RNA samples. (C) Pairs for which an tRNA in the small RNA libraries was correlated with a repeat in the long RNA samples. The top 20 most numerous pairs and their given Spearman correlations are shown, with the pairs identically graphed in the same order for both study arms. The proportion of each RNA pair belonging to given correlation range (e.g. a bin of 0.8;0.9 contains correlations with a rho greater than or equal to 0.8 and less than 0.9.) are stacked along the Y-axis and colored according to the correlation range.
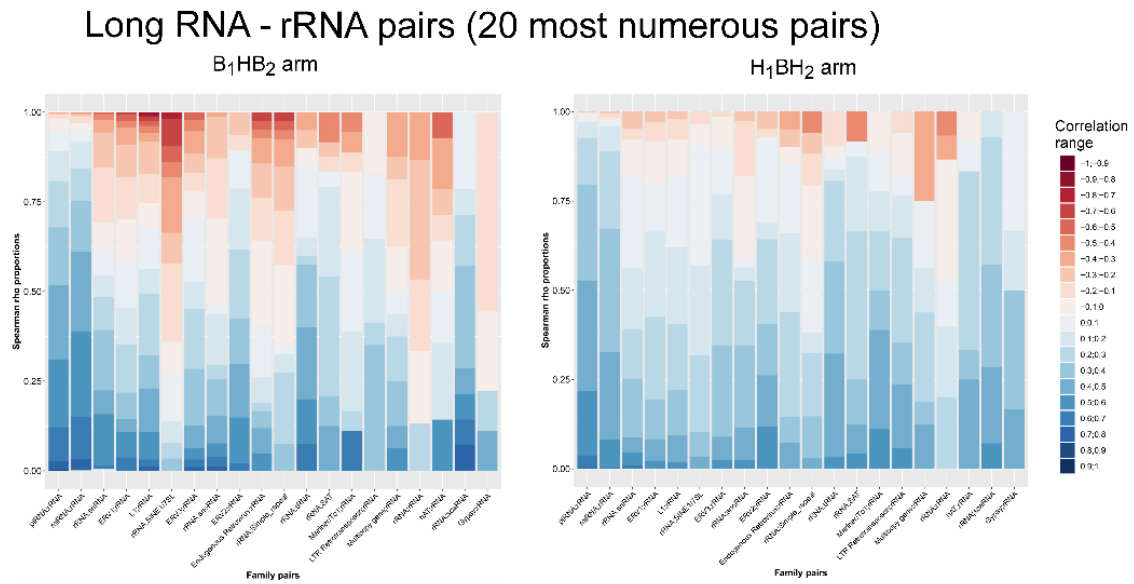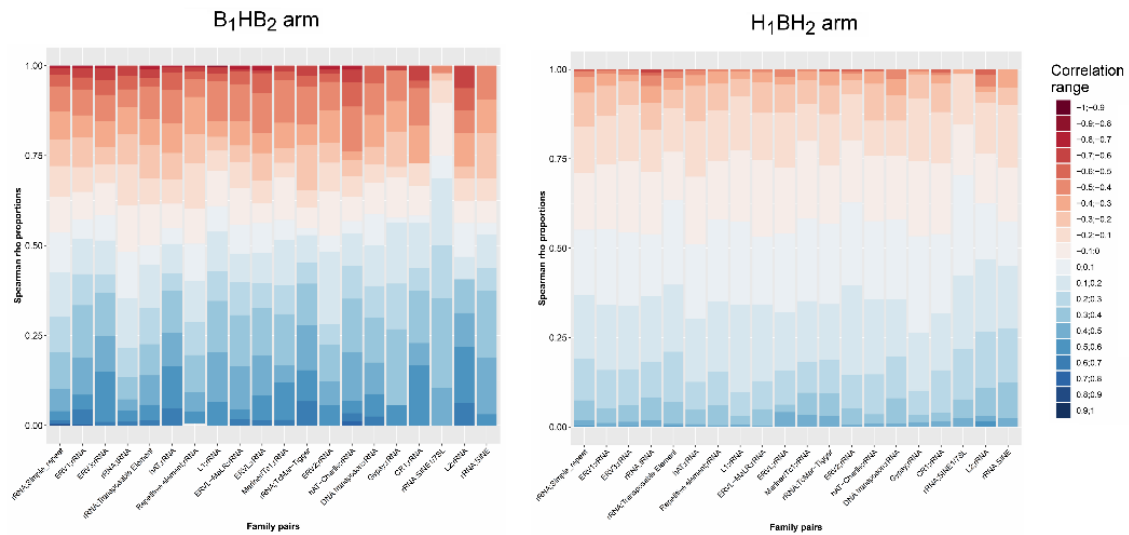
Figure 4.18. rRNA correlations to repeats and small RNAs. (A) Pairs for which an rRNA repeat (derived from the long RNA) was correlated with an expressed small RNA are shown. (B) Pairs for which an rRNA repeat (derived from the small RNAs) was correlated with a repeat from the long RNAs are shown. The top 20 most numerous pairs and their given Spearman correlations are shown, with the pairs identically graphed in the same order for both study arms. The proportion of each RNA pair belonging to given correlation range (e.g. a bin of 0.8;0.9 contains correlations with a rho greater than or equal to 0.8 and less than 0.9.) are stacked along the Y-axis and colored according to the correlation range.

Concordant with piRNA and miRNA pairings, the majority of tRNA (from small RNA

libraries) - long RNA genomic repeat pairs were positive, with a small number of each pair

exhibiting negative correlations (**Figure 4.17**). Additionally, piRNA, miRNA, tRNA, and scaRNA, while largely positively correlated with three main ERV family members (ERV1, ERV2, and ERV3) [322], also exhibit a series of strong negative correlations (**Figure 4.19**). Interestingly, all ERVs have strong negative correlations with short rRNA fragments and short RNA simple repeats. Together, this data supports that view that tRNAs, miRNAs, and piRNAs are negative regulators of the observed repeat partners, and they likely act in a target-specific manner, rather than targeting an entire class.
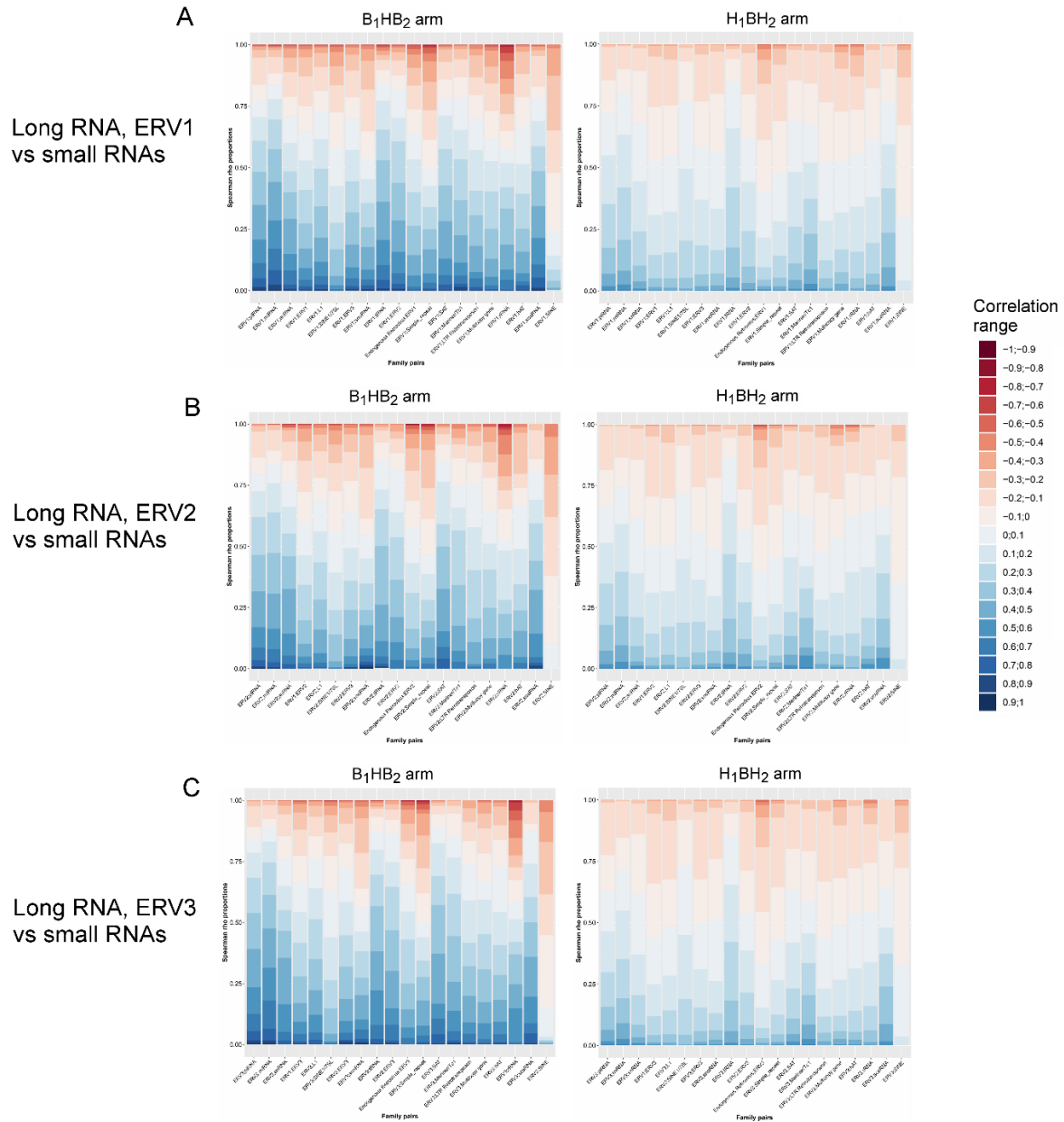
Figure 4.19. ERV correlations to small RNAs. (A-C) Pairs for which an ERV repeat (derived from the long RNA) was correlated with an expressed small RNA are shown. The top 20 most numerous pairs and their given Spearman correlations are shown, with the pairs identically graphed in the same order for both study arms. The proportion of each RNA pair belonging to given correlation range (e.g. a bin of 0.8;0.9 contains correlations with a rho greater than or equal to 0.8 and less than 0.9.) are stacked along the Y-axis and colored according to the correlation range.

Dynamic changes in sperm RNAs occur during spermiogenesis

Detection of consistent associations between small RNAs and long RNA with sperm repetitive elements suggests a dynamic relationship. As shown in **Figure 4.20,** small rRNA fragments and large rRNA fragments exhibit opposite correlations with piRNA, miRNA, tRNAs and scaRNA (**Figure 4.20,** center). In addition, these small RNAs were primarily positively correlated (**Figure 4.20,** solid line) with simple repeats and ERVs. However, select ERV1 repeats (HERV17-int and HERVIP10FH-int) and ERV3 repeats (MER68C, LTR47A2, LTR86B2, and LTR86A2) were negatively correlated (**Figure 4.20,** dotted line) with small RNAs. This suggested an opposite relationship between the select ERVs and piRNA, miRNA, tRNAs and scaRNAs. Similarly, simple repeats, while positively correlated with most small RNAs, had a series of strong negative correlations to small RNAs (**Figure 4.20,** dotted line with blue star). Specific miRNAs, highlighted by hsa-miR-4516 and hsa-miR-30a-3p and scaRNAs, such as scaRNA15, scaRNA 13, and scaRNA8, comprised a large proportion of the negative correlations to simple repeats (**Figure 4.20,** dotted line with blue star). Among the tRNAs, which are assumed to be tRNA fragments, tRNA-GLY, tRNA-VAL, tRNA-ILE, and tRNA-MET accounted for the greatest number of negative correlations.
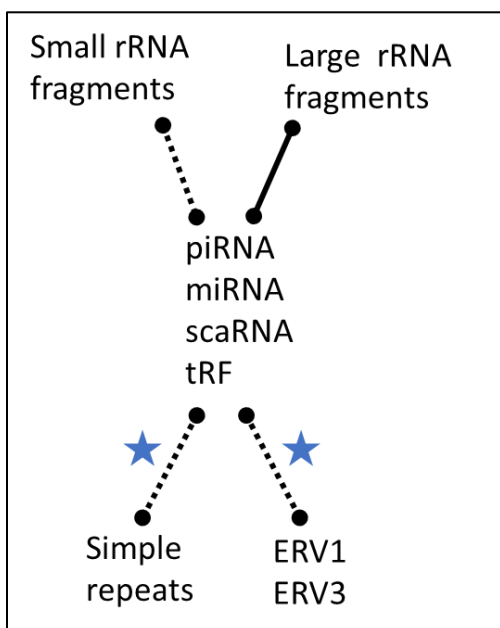
Figure 4.20. Relationships of sperm RNA repeats and small RNAs. Negative and positive correlations are represented as dashed lines and solid lines, respectively. Negative correlations that only occur in a specific manner are marked with a star. tRF=tRNA fragment; scaRNA=Small Cajal body-specific RNA; ERV=Endogenous RetroVirus

The presented interactions in **Figure 4.20** are not known to be causal (either indirectly or directly) but are intended to illuminate the possible dynamic processes during spermiogenesis. Creation of a functional spermatozoa capable of fertilization requires a complex re-arrangement of organelles during the post-meiotic phase of spermatogenesis, accompanied by removal of the cytoplasmic droplet. Additionally, the spermatid nucleus undergoes extensive chromatin compaction of the haploid genome, replacing the majority of histones with protamine toroids. Due to these changes and extensive degradation of RNAs, including ribosomal RNAs, the ejaculated spermatid is transcriptionally and translationally inert. The rRNA locus exhibited a similar read coverage of the by long and short RNA libraries. However, given the opposite associations of long and short rRNA fragments to small RNA species, it is tempting to speculate that the two rRNA populations represent decreasing and increasing rRNA fragmentation, respectively. The paradigm in mammalian spermatozoa is that sperm do not contain intact rRNA, having undergone extensive fragmentation during the post-meiotic phase of spermatogenesis. The current analysis suggests that there is a spectrum of rRNA fragmentation, ranging from long (>200 bp) fragments to short (<50 bp) fragments. These long and short rRNA fragments were not strongly correlated (either negatively or positively), and so are not connected in **Figure 4.20**. The biological phenotype associated with the network is unknown. However, the dynamics are suggestive of a choice between destructive and non-destructive pathways during spermatogenesis, and possibly the quality of the ejaculated spermatozoa. This dynamic spermatozoal network may also play a role in the considerable transcriptomic intra- and inter-individual heterogeneity of human sperm.

v. Discussion

In both rural and urban environments, humans are exposed to cocktails of endocrine disruptors [323]. While environmental regulations designate maximum allowable levels of only a subset of the numerous xenobiotics, this level is primarily determined through animal

models, which may not accurately reflect the human condition. The human male is known to mediate some intergenerational effects in offspring [17], yet the intergenerational effect of adult paternal exposures to common xenobiotics and endocrine disruptors, particularly in humans, is poorly characterized. The MARS study showed that exposure of human males to high levels of a single endocrine disruptor, di-butyl phthalate (DBP), was capable of reducing sperm motility in DBP-naïve subjects [69].

Applying RNA-seq to the MARS samples (**Figure 4.1**) showed that DBP-induced alterations in spermatozoal RNAs were largely unique to a single study arm (either the acute B_1HB_2 or chronic H_1BH_2 study arm). Each biological response to increasing or decreasing DBP levels yielded a different altered RNA profile (**Table 4.4**). Interestingly, novel RE's comprise a significant portion of altered REs, indicating that DBP exposure(s) affects far more than the previously known transcripts. The RNA profiles observed in the ejaculated spermatozoa reflect the final outcome of spermatogenesis, which includes both RNAs generated in preparation for differentiation and those acquired during epididymal maturation for transmission to the future embryo. Within the immunoprivileged state of the testis (reviewed in [324]), the spermatogenic effect of high-DBP mesalamine is likely communicated through Sertoli cells that support the germline during differentiation, or through the epididymis during transit, when the sperm first become exposed to other fluids.

Mammalian models have previously indicated that phthalate exposure may be acting on reproductive tissues through processes that include PPAR-dependent mechanisms [58-60], inducing oxidative damage [67, 68]. In the current study, several spermatogenesis-related pathways, including activation of EIF2 signaling and the PPAR-alpha/RXR-alpha signaling, were strongly associated with the transition from baseline (B_1 visit) to crossover (H visit) (**Figure 4.8B**). This suggests that a concerted shift in the Peroxisome Proliferator-Activated Receptor alpha (PPAR-alpha) and Eukaryotic Initiation Factor 2 (EIF2) pathways occurs at the initial high-DBP exposure in DBP naïve males. Conversely, within the H_1BH_2 arm, after a

temporary respite of a single spermatogenic cycle on non-DBP medication, the return to high-DBP mesalamine activated oxidative stress and DNA damage response pathways, perhaps signaling the beginning of the repair process (**Figure 4.8A**). The ontological association of androgen receptor coregulation in the H_1BH_2 arm of the MARS study also suggests that responses to *in vivo* adult exposures elicits androgen disruption, which has previously been implicated in phthalate-induced testicular dysgenesis syndrome. Notably, subjects in the H_1BH_2 arm have been chronically exposed to high-DBP mesalamine, some for several years, prior to the MARS study and are assumed to have reached a phthalate-induced expression plateau in response to the high-DBP levels. The temporary (1 spermatogenic cycle) withdrawal from high-DBP mesalamine then precedes the additional and significant stress imparted onto the germline upon re-introduction of high-DBP mesalamine.

Among the genomic repetitive elements shown to be enriched in spermatozoa, TC-rich tetramers form a larger contribution to the sperm RNA when an individual has experienced a high-DBP exposure in the previous spermatogenic cycle (**Figure 4.10**). The time required to fully recover is not known and may well be far longer than the single crossback cycle observed in the B_1HB_2 arm of the current study. Nevertheless, the study indicates that any recent high-DBP exposure increases the abundance of simple repeats in human spermatozoa. The biological function of these recurring simple repeats in spermatogenesis, fertilization and early embryo development has yet to be defined [245]. However, they are compartmentalized in sperm [156], perhaps reflective of a role in sperm chromatin organization. Phthalate's noted ability to increase DNA nicking [310] and thus spermatozoal DNA fragmentation [67, 68] in a specific manner [311] may alter the specialized compact chromatin environment in spermatozoa. The physiological effect(s) of the cocktail of background exposure to endocrine disruptors in humans, particularly in somatic tissue, is unknown. However, given the potential intergenerational effects of the paternal germline

(mediated by sperm RNA content and sperm epigenome), ubiquitous human exposure to phthalates and other known endocrine disruptors remains a concern.

The MARS study also provided the opportunity to assess the spermatozoal impact of mild Inflammatory Bowel Disease (IBD). IBD, defined as Ulcerative colitis and Crohn's disease, is a common condition, with a prevalence of approximately 1 per 500 people [325]. The current study compared males with non-flaring, mild IBD treated daily with mesalamine to a control cohort of fertile males from idiopathic infertile couples. As expected, mild IBD, or chronic mesalamine use, had minimal impact on spermatozoal contents (**Figure 4.6**).

The transcriptomic dynamics across human spermatogenesis is complex [175, 245, 326]. Mechanisms driving transitions across sperm differentiation have been largely inferred from mammalian model systems [327, 328]. The final stages of sperm differentiation, spermiogenesis, is noted for extensive modification of chromatin structure, organelle distribution, and changes in RNA composition [35, 36, 245]. However, the role of RNAs in promoting these modifications has only recently begun to be addressed [156, 329]. Previous studies on chromatin structure have suggested a role for genomic repeats in establishing and maintaining chromatin. The MARS dataset demonstrated an spermatozoal enrichment of transcribed repetitive elements, including simple repeats, ERV1, tRNA, and ERVL, which may play a role in establishing the human spermatid. The relationship between transcribed repeats (such as simple repeats and endogenous retroviruses), ribosomal RNAs, and small non-coding RNAs in ejaculated sperm was also uncovered. The identification of small cajal body-specific RNAs (scaRNAs), involved in small nuclear ribonucleoprotein (snRNP) biogenesis, highlights the importance of RNA processing and localization in spermatogenesis.

The high-level network presented in **Figure 4.20** provides a snapshot of the networks involved in successful spermiogenesis, selective spermatid apoptosis [154, 330, 331], or maturation during epididymal transit [71, 332-334]. The morphological and transcriptomic intra- and inter-individual heterogeneity of human sperm [335-337] is likely, at least in part,

due to choices along this network. This was exemplified by the RNA dynamics in spermiogenesis of DBP-naïve subjects, adjusting their response in accord with position along the network, i.e., spermatozoa recently exposed to high-DBP are enriched in simple repeats.

At present, the MARS study shows that exposure of human males to high levels of a single endocrine disruptor, di-butyl phthalate (DBP), can alter spermatozoal RNAs and expression of genomic repeats in sperm. Furthermore, an individual's history of high-DBP exposure influences their reproductive response to changes in DBP levels. The time period required to fully recover from a high-DBP exposure, while currently undetermined, was suggested in this study to be longer than a single spermatogenic cycle (approximately 90 days). Future *in vitro* and *in vivo* experiments relevant to adult phthalate exposures are required to identify the mechanisms and pinpoint the biological processes at work in the reproductive and endocrine systems of phthalate-exposed adults. Observational studies of offspring from DBP-exposed fathers will provide a path to determine the extent of the impact that paternal DBP exposure presents as health risk to their subsequent children.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Developmental Origins of Health and Disease (DOHaD) proposes that preconceptional, prenatal and childhood exposures affect health outcomes later in life. The use of animal models has shone light on the importance of maternal health and gestational influence on offspring health. Notably, environmental exposures, such as exogenous endocrine disruptors, can be influential during key periods of susceptibility in fetal development [338-341]. However, the male influence in DOHaD has only begun to be appreciated. In this final chapter, I review the discoveries made regarding the pre-conceptional environment of human embryos, and contents of the male germline.

i. Assisted Reproductive Technologies

The circumstances surrounding a successful human birth are complex. The initial events involve the fertilization of a matured oocyte by a presumably high-quality spermatozoa, followed by embryonic cell division and finally implantation [342]. Advanced Assisted Reproductive Technologies, such as IVF, ICSI, cryopreservation, controversial nuclear transfer and Mitochondrial transfer techniques, manipulate the pre-implantation embryo [343-345]. While such infertility treatments offer hope to individuals struggling to conceive, the treatments may also have unintended consequences for the offspring due to an altered pre-conceptional and pre-implantation environment. With the exception of intentional embryonic gene editing [346], the effect(s) of ART on offspring are expected to be primarily epigenetic in nature. A variety of epigenetic mechanisms, such as DNA methylation, histone modifications, and chromatin structure, are potentially altered by ART [82, 89, 91, 97].

DNA methylation, the addition of a methyl group to DNA, is typically examined in context of 5-methylcytosine (5-mC). In Chapter 2, I examined the influence of several ART processes, including embryo cryopreservation, on the 5-mC profiles of the child soon after birth. As shown in **Figure 2.1**, the DNA methylation profiles of human newborns conceived

naturally, or through the use of intrauterine insemination (IUI), or *in vitro* fertilization (IVF) using Fresh or Cryopreserved (Frozen) embryo transfer, were compared. Except for the naturally conceived infants, all newborns used in the study were born to parents experiencing some degree of infertility, which required IUI or IVF/ICSI to resolve. Naturally conceived infants were shown to have a dramatically different methylation profile compared to those born through IUI or ICSI. Therefore, this research suggests that the underlying infertility of the parents influences the child's methylation. However, outside of ensuring that the samples were from full-term newborns, this work did not consider maternal characteristics, such as race, age, and socio-economic status, due to a lack of maternal and demographic information. In particular, the natural conception group was drawn from an urban, primarily African-American population, while the assisted conception group was drawn from patients undergoing fertility treatment in a suburban clinic. Such maternal characteristics can bias results and should be considered when modeling datasets from in future newborn cohorts. However, if the current interpretation is accurate, this would lead to the striking conclusion that successful treatment of infertile individuals, even by relatively non-invasive measures, produces a cohort of children with an epigenome different from those born to fertile individuals. Human infertility is a common condition, affecting approximately 12% of couples of childbearing age in the United States [30], which has a variety of causes. Therefore, examination of the newborn's epigenome in context of the infertility diagnosis (e.g. male factor, specific female factor) will be essential for characterizing this future population of children.

A second interesting observation was the striking similarity of IUI and IVF-Frozen embryo transfer infants (**Figure 2.2**). This suggested that epigenetic aberrations in the IVF conceptions may be abrogated using embryo cryopreservation. Notably, these results are in accord with the observed reduction in birth defects using ART protocols that employ cryopreserved embryos [89]. Taken together, these studies implicate a resetting mechanism in cryopreserved embryos. However, women undergoing implantation of cryopreserved

embryos are likely to be naturally cycling, whereas patients undergoing fresh embryo transfer (excluding the use of gestational surrogates) recently experienced unusual hormonal and physical stresses in preparation for oocyte extraction. Due to a number of social and economic factors, embryo and oocyte cryopreservation are increasingly being utilized by infertility clinics [347]. The above research indicates that this trend towards cryopreservation may be beneficial to the offspring and help ameliorate any epigenetic aberrations introduced by the use of IVF/ICSI.

Periconceptual nutrition, in the context of unassisted conception, is known to alter epigenomes of offspring at specific loci termed metastable epialleles (MEs). With the current cohorts, I tested the hypothesis that ME loci were sensitive to early nutritional exposure in the context of IVF culture conditions. IVF culture conditions and parental infertility showed consistently altered methylation at certain MEs. This was the first study to reveal an impact of ART or fertility status on MEs and suggests a lasting epigenetic effect of IVF nutrition on the developing embryo. However, it is important to note that the current study was limited to loci targeted by the 450K array, and further limited by the ChAMP approach, which did not assess methylation changes of loci with fewer than two probes. The implementation of the larger EPIC array MethylationEPIC BeadChip array, which covers over 850,000 CpG sites and provides greater coverage of distal regulatory elements, or bisulphite sequencing, which can provide whole genome coverage, would be more informative in future cohorts. Additionally, while the current cohort had limited patient data, a large matched cohort with detailed records of maternal and gestational variables, employing more expansive DNA methylation detection methods, would serve well to verify the observed sites and identify additional epimutations that may be caused by ART.

ii. Sperm RNA

The RNA profiles of human sperm, as well as how RNA profiles change after endocrine disruptor exposure, was explored in Chapters 3 and 4. The male germline is a highly

specialized cell type, with a unique cell structure [35] and transcriptome [154]. Mature spermatozoa are transcriptionally and translationally inactive, with a high degree of RNA fragmentation to ensure silence and expunge the majority of the cytoplasm [35]. For the majority of the known transcriptome, the considerable heterogeneity and transcriptome fragmentation in spermatozoa renders common gene expression approaches inaccurate. Additionally, recent studies have observed a series of RNAs arising from intronic and intergenic loci, which are not enumerated in the annotated human transcriptome [154, 163].

To fully address these intronic and intergenic RNAs, the RNA Element (RE) discovery algorithm (REDa) was developed [245]. REDa is a tool for the discovery of transcribed, unannotated sequence elements from RNA-seq libraries, and was applied to a spectrum of tissues and cells representing germline, embryonic, and somatic tissues, shown in **Figure 3.3**. In all examined tissues, previously unannotated RNAs were identified, with transcription of such RNAs throughout the autosomes and sex chromosomes. Interestingly, the post-meiotic stages of spermatogenesis (Round spermatids and ejaculated sperm) contain large numbers of novel (intergenic and intronic) REs (**Figure 3.7**). While the function of these novel RNAs in spermatogenesis is not known, murine spermatozoa have shown preferential localization of RNAs [156]. Identification of the cellular location of the human sperm RNAs, both from known transcripts and novel REs, across the post-meiotic stage of spermatogenesis, would be a first step towards understanding the potential function of sperm RNAs.

It has been proposed and shown *in vitro* that human sperm deliver a collection of RNAs upon fertilization [154, 160, 211, 212]. The series of human RNA-seq profiles from sperm, oocyte, and embryo allowed for the identification of REs that were transmitted to the human oocyte solely by sperm. As expected, the majority of zygotic RNAs were shown to be derived from the oocyte. However, up to 289 sperm REs were identified as a majority contributed by paternal transmittance, with an FDR of ~3.4%, and 75 REs essentially provided by the sperm, at an FDR of ~2.7% (**Figure 3.10 A,B**). While the role of these paternally transmitted RNAs

in the embryo was not examined in the current work, such RNAs may play a regulatory role in the embryo. Of particular interest are any full length RNAs, as these are potentially translated in the embryo [154]. Micro-injection of known paternally delivered RNAs, in the context of murine or human embryos, would serve to determine if such RNAs can tune embryo development.

Novel REs (intronic, near-exon, and orphan REs) often correspond to the same regions as genomic repeats. Given this association, the population of REs was used to determine the enrichment of genomic repeats in various cell types. This approach has the advantage of knowing the exact location and proxy expression (via RE expression) of the genomic repeat. Certain retrotransposons are of known interest in the early human embryo, such as ERVL [188, 348], SINE-VNTR-Alu (SVA) elements, and LTR12 (LTRs of HERV9) [349]. The current study also demonstrated transcriptional enrichment of ERVL (MLT2A1 and MER73) and SVA (SVA-D) during human embryogenesis. Repeat transcription during spermatogenesis was also observed, with an enrichment of simple centromeric repeats, MER1A (DNA transposon), HERVE, HSAT1 (Satellite) [350], and LTR71B (member of the ERV1 family) [322] during spermatogenesis. The classes of enriched repeats appears to be vastly different between spermatogenesis and embryogenesis, with an additional switch during embryonic genome activation, when REs associated with MLT2A1 and SVA-D are transcribed. It is important to note that the observed repeat enrichment in round spermatids and spermatozoa may be indicative of a targeted retention of certain RNAs during the RNA degradation and cytoplasmic expulsion of spermatozoa, rather than active transcription of the given repeat. This distinction cannot be determined with the current data structure, but requires measurement of the spermatozoal RNAs in context of a consistent exogenous control RNA. An alternative approach would be to perform target counts with an array platform that does not require sample amplification, such as nanoString [351]. It is currently unknown if repeat transcription contributes solely to spermatogenesis, or if such RNAs also play a role in

the early embryo. Microinjection of these RNAs into the mammalian embryo, followed by examination of the embryonic epigenome and growth characteristics, would improve understanding of these RNAs.

Humans are ubiquitously exposed to endocrine disruptors (EDs), such as phthalates. However, the influence and underlying mechanisms of the cocktail of commonplace ED exposures on reproduction in humans, particularly in adult males, is poorly understood. While such exposures can easily be examined in animal models, there are practical issues to observing the causative mechanisms of reproductive effects in humans. However, as ejaculated sperm is an accessible cell type, examination of sperm RNAs may assist in addressing the reproductive effects of EDs in human males. In addition to being potential mediators of transgenerational inheritance, sperm RNAs may act as markers of environmental perturbations or disease [163, 254]. Using the MARS study [69], changes in sperm RNAs across longitudinal DBP exposure can be assessed to identify DBP-induced reproductive changes. Additionally, the availability of MARS samples, which originate from individuals suffering from mild IBD, provided the opportunity to identify IBD-induced changes in spermatozoa. IBD-induced spermatozoal alterations were thus identified by comparing control sperm samples to the B_1HB_2 arm samples, which represent men initiating the MARS study on a DBP-naïve condition. This work showed few sperm RNA changes due to mild IBD. Future cohorts should also include individuals with moderate to severe IBD, while controlling for medication use. Overall, these results are reassuring for IBD-afflicted males.

By applying a LMEM to the MARS long RNA samples, the patterns of REs across shifting DBP levels (High or Background) were generated. Surprisingly, few REs were altered strictly with DBP levels (top two patterns of **Table 4.4A**). Instead the majority of differential REs were altered in solely the Acute phase (Baseline to Crossover) or the Recovery phase (Crossover to Crossback). This occurred regardless of the study arm, suggestive of an phenotypic shift lasting long after the initial stimulus. The gene associations of differential REs

provided the opportunity to identify ontological terms and signaling pathways altered by DBP. The DBP-naïve cohort (B_1HB_2 arm) was associated with translation, PPAR/ RXR-alpha signaling, and a strong activation of GP6 signaling. PPARs are known to act as xenobiotic sensors, and can interact with phthalates, resulting in PPAR activation [58-60]. The other enriched pathways, notably GP6 signaling, have not previously been directly implicated with DBP in literature. GP6 is a classical receptor for collagen, involved in platelet aggregation and thrombus formation, but also is known to be enriched in human testis [352]. As a gradual reduction in collagen binding by sperm during uterine transit has previously been suggested to enable sperm capacitation [294, 295], GP6 signaling may play a role in sperm capacitation. In contrast, the H_1BH_2 arm, implicates several pathways classically associated with experimental phthalate exposures and phthalate-induced testicular dysgenesis syndrome, such as oxidative stress, DNA damage response pathways, and androgen receptor regulation. This suggested that re-introduction of DBP-containing medication after a single spermatogenic cycle at background DBP level may harm the germline. The use of normalized RNA expression to predict DBP-induced activation or repression of signaling pathways provides a starting point for targeted proteomic analyses. However, verification of protein expression and the presence of known protein modifications involved in signaling pathway modulation is needed.

In addition to the gene ontologies and signaling pathways associated with the differential REs, examination of repeat-associated REs suggested that high-DBP exposure influenced repeat transcription. TC-rich simple repeats are upregulated by recent (1 spermatogenic cycle) high DBP exposure, and centromere-associated repeats appear to be downregulated by high DBP in DBP-naïve individuals. The current approach has the advantages of precisely placing expression of genomic repeats, by using RE expression as a proxy for repeat expression. However, this assumes that the overlapping RE is an accurate measure of the individual genomic loci. Therefore, the association of repeat transcription with

DBP exposure(s) should be replicated using the read coverage of the individual repeat loci. Simple repeats have previously been implicated in spermatozoal nuclear matrix-associated regions [230], while centromeric RNAs may play a role in establishing stage specific chromosomal structure and position throughout spermatogenesis [229, 230]. Altogether, this data suggests that DBP likely alters germline chromatin structure during spermatogenesis. While testing this hypothesis is outside the scope of this thesis, re-examination of the MARS cohort or future cohorts for chromatin accessibility and histone marks (e.g. ChIP-seq) would provide a further understanding of DBP-induced chromatin modifications.

Small RNAs have been implicated as likely mediators of paternal transmission. Using a subset of the MARS samples, small RNA-seq libraries were generated to identify the changes induced by human *in vivo* DBP exposure. Compared to the numerous DBP-regulated REs in the long RNAs, relatively few small RNAs were identified as being altered by DBP (**Figure 4.15**), in either study arm. However, a single piRNA, hsa_piR_01967, was downregulated in both study arms upon high-DBP exposure. Despite the limited power of the present small RNA analysis, this work suggests that DBP does impact small RNAs. The physiological importance of the altered small RNAs in testis and the spermatogenic cycle is not yet known. However, taken together, the long RNA and short RNA fractions of the MARS samples suggest that DBP may alter various cellular mechanisms and signaling pathways during post-meiotic spermatogenesis.

The male and female germline are known to have abundant expression of small regulatory RNAs, such as piRNAs. Concordantly, abundant piRNAs were observed in the MARS dataset. However, the regulatory relationships of RNAs during spermatogenesis is not well known. Previous studies imply a role of small RNAs in repeat suppression, with piRNAs acting as direct regulators of transposable elements [317] and miRNAs having an indirect role [318-321]. However, the relationship of the small RNAs to the larger coding RNAs and RNAs generated from genomic repeats during spermatogenesis is poorly defined. Using the MARS

datasets, the availability of small RNAs and proxy genomic repeat expression permitted the construction of the relationship between the two RNA categories. As shown in **Figure 4.20**, consistent associations between small RNAs and long RNA with sperm repetitive elements suggests a dynamic relationship. In particular, small rRNA fragments and large rRNA fragments exhibit opposite correlations with piRNA, miRNA, tRNAs and scaRNA. Select ERV1 and ERV3 elements were also negatively correlated (**Figure 4.20,** dotted line) with small RNAs, suggesting a contrary relationship. Similarly, simple repeats, while positively correlated with most small RNAs, had a series of strong negative correlations to select small RNAs. Notably, the network suggests opposite relationships of long and short rRNA fragments to small regulatory RNAs. Spermatozoa are known to have extensively degraded ribosomal RNA, and both long and short rRNA fragments have similar coverage of the consensus rRNA loci. It is tempting to thus hypothesize that the small regulatory RNAs may be acting to cause rRNA fragmentation. This hypothesis will require extensive work to prove (1) binding of the small RNAs to the target rRNA locus and (2) causative fragmentation of rRNA. Regardless, the network dynamic is suggestive of a choice between destructive and non-destructive pathways during spermatogenesis, and possibly the quality of the ejaculated spermatozoa.

iii. Conclusion

The conception and uterine establishment of a human embryo is influenced by a variety of factors, including the quality of the male and female germline. In Chapter 2, I outlined the importance of the peri-conceptional environment in the context of assisted reproductive technologies, namely, IVF/ICSI and cryopreservation. The sub-fertility of the parents and the use of embryo cryopreservation alters methylation in the newborn's blood cells. Additionally, while metastable epialleles (MEs) have been previously associated with methyl donor supplementation (in mouse) and seasonal diet changes (in human), this is the first work to associate ART/infertility with MEs in humans.

The involvement of the male germline in DoHAD has mostly been shown in mouse and rat models. The tools to identify sperm RNAs, which likely are mediators of generational transmission were developed and human spermatozoa were examined in the context of a common endocrine disruptor. Using ejaculates from men longitudinally exposed to DBP, both long and small RNAs were found to be altered by phthalate. Many of the differential long RNAs, examined in context of RNA elements, were novel, non-exonic RNAs. These exploratory works provide a set of RNAs and repeat classes to examine in future studies of human spermatogenesis, embryogenesis, and endrocrine disruptor exposure.

Appendix A : DNA quality scores for Newborn bloodspots. A260/A230 and A260/A280 quality scores for IUI, FH, and FZ samples are listed.

Sample name	A260/A230	A260/A280	Sample name	A260/A230	A260/A280
10006_B	0.08	2.5	10798_B	0.08	2.74
10010_A	0.13	2.58	10821_C	0.09	2.03
10046_B	0.09	2.61	10829_C	0.1	2.52
10055_C	0.1	2.65	10831_C	0.11	2.21
10058_A	0.1	3.18	10843_B	0.11	2.21
10101_B	0.15	2.07	10846_B	0.12	2.28
10117_B	0.1	2.64	10863_A	0.15	2.97
10133_B	0.08	2.84	10869_B	0.14	1.77
10136_C	0.1	2.49	10882_B	0.09	2.58
10141_C	0.09	2.83	10890_C	0.1	2.32
10156_B	0.09	1.92	10894_C	0.08	2.42
10208_C	0.07	2.53	10932_B	0.07	2.75
10225_A	0.14	2.52	10947_C	0.09	2.79
10231_B	0.11	1.85	11002_B	0.07	2.77
10255_A	0.1	2.84	11016_B	0.09	2.44
10291_A	0.14	2.55	11029_B	0.08	2.68
10292_C	0.09	2.6	11034_B	0.07	2.77
10344_B	0.1	2.33	11048_C	0.09	2.99
10346_B	0.08	2.32	11066_A	0.13	2.41
10347_C	0.17	2.32	11067_A	0.11	2.1
10350_B	0.11	2	11127_C	0.11	2.67
10359_C	0.12	2.27	11165_C	0.16	2.35
10374_A	0.09	2.92	11177_C	0.06	2.62
10389_C	0.11	2.43	11283_B	0.11	2.62
10402_B	0.13	1.91	11340_C	0.09	2.68
10432_B	0.12	2.03	11359_B	0.11	2.1
10434_B	0.11	2.09	11368_C	0.18	1.77
10443_C	0.13	2.27	11378_B	0.1	4.99
10447_C	0.08	2.9	11381_C	0.08	2.5
10453_C	0.11	2.79	11387_B	0.08	3.66
10460_C	0.07	2.94	11414_A	0.08	1.99
10511_C	0.12	2.52	11420_B	0.09	2.82
10512_A	0.11	2.79	11452_C	0.09	1.82
10555_C	0.1	2.94	11458_B	0.12	2.19
10558_B	0.12	2.49	11465_C	0.12	1.97
10587_C	0.09	2.76	11508_A	0.14	2.21
10601_B	0.13	2.32	11523_A	0.13	2.37

10612_C	0.07	3.61	11541_B	0.06	3.44
10649_C	0.12	2.61	11580_C	0.1	2.59
10653_A	0.1	2.43	11598_B	0.1	2.82
10667_C	0.09	2.93	11600_C	0.14	2.19
10695_C	0.1	2.36	11624_B	0.12	2.37
10697_B	0.09	3.39	11629_B	0.15	3.4
10706_B	0.11	2.05	11651_B	0.12	2.22
10716_A	0.13	2.35	11691_A	0.1	2.72
10729_C	0.1	2.42	11981_A	0.11	2.25
10743_C	0.1	2.44	11998_A	0.07	4.51

Appendix B: Enhancers overlapping differentially methylated regions. Chromosomal locations of enhancers are provided.

FH vs IUI	FZ vs IUI	FZ vs FH	IUI vs NAT	FH vs NAT	FZ vs NAT
chr10:80998898-81000011		chr17:75883563-75884329	chr11:2919958-2920393	chr10:14051386-14052060	chr10:14051386-14052060
chr10:8373481-8373809		chr10:134361912-134362323	chr1:201619383-201619798	chr10:21798958-21799135	chr10:21798958-21799135
chr12:117483062-117483484		chr10:29228763-29229094	chr1:247511365-247511433	chr11:2919958-2920393	chr10:72973788-72974249
chr12:14413561-14413925		chr10:2951238-2951396	chr12:95840387-95840581	chr1:201619383-201619798	chr1:12600525-12600853
chr1:42385609-42385954		chr10:49879628-49879846	chr16:57701111-57701632	chr1:247511365-247511433	chr11:2919958-2920393
chr16:86012146-86012389		chr10:80998898-81000011	chr16:87978906-87979489	chr15:75018665-75019034	chr1:201619383-201619798
chr17:75096190-75096538		chr10:8373481-8373809	chr16:89893816-89894170	chr17:79068878-79069252	chr12:53359181-53359363
chr17:79128807-79129119		chr11:32109882-32110285	chr17:76801462-76801678	chr18:11849903-11850727	chr12:7781011-7781133
chr17:80829054-80829460		chr11:78131805-78132171	chr17:79068878-79069252	chr19:2332363-2332647	chr15:75018665-75019349
chr18:74114511-74114735		chr12:53612468-53612778	chr19:2332363-2332647	chr19:42440101-42440390	chr1:59280364-59280505
chr21:44573696-44574026		chr13:110521819-110522258	chr2:200468498-200468920	chr19:57351144-57351414	chr16:85785568-85785942
chr2:200468498-200468920		chr13:28554929-28555135	chr2:8597253-8597584	chr21:44104688-44105340	chr17:38170793-38171091
chr2:240362327-240362618		chr14:107252931-107253417	chr7:101361034-101362207	chr21:44573696-44574026	chr17:56744054-56744558
chr2:242954051-242954322		chr1:43250353-43250859	chr7:2150639-2150902	chr2:1608919-1609202	chr17:79068878-79069252
chr22:43165805-43166293		chr14:95239496-95239653	chr7:4848839-4849223	chr2:1609558-1609866	chr18:11849903-11850727
chr5:1103805-1104368		chr15:31372759-31373034		chr2:200468498-200468920	chr18:72916177-72916455
chr5:1107341-1108073		chr16:1585533-1585949		chr2:241585712-241586190	chr19:2332363-2332647
chr5:669501-669935		chr16:86016233-86016516		chr2:3583560-3583858	chr20:22567524-22567732
chr7:2124507-2124641		chr16:86795224-86795595		chr2:8597253-8597584	chr21:44573696-44574026
chr8:142237099-142237665		chr16:88700249-88701257		chr4:3374732-3375130	chr2:1609558-1609866
chr8:49426952-49427328		chr16:89184955-89185230		chr5:669501-669935	chr2:240362327-240362618
chr8:49427546-49427722		chr17:14206970-14207459		chr7:2548054-2548433	chr2:241585712-241586190
		chr17:19627830-19628049		chr7:4764796-4765177	chr22:43165805-43166293
		chr17:32253-32376		chr8:144948428-144948782	chr2:8597253-8597584
		chr17:76875667-76876103			chr2:87036674-87037618
		chr17:78793273-78793673			chr4:3374732-3375130
		chr18:11849903-11850727			chr5:1103805-1104368
		chr18:77552344-77552827			chr5:669501-669935
		chr18:77723013-77723575			chr7:4848839-4849223
		chr19:3369394-3369884			chr7:73157180-73157738
		chr19:58715620-58715769			chr8:11560575-11560973
		chr2:121624920-121625080			chr8:144948428-144948782
		chr21:38630405-38630786			

	chr2:240868441-240868868			
	chr2:394418-394640			
	chr4:3374732-3375130			
	chr4:640356-640557			
	chr5:1103805-1104368			
	chr5:1107341-1108073			
	chr5:157001390-157002014			
	chr5:178692524-178692976			
	chr5:669501-669935			
	chr5:71852712-71853181			
	chr6:170589556-170589600			
	chr6:28885422-28885544			
	chr6:30720004-30720713			
	chr7:25992296-25992472			
	chr7:47579966-47580265			
	chr7:98990795-98991133			
	chr8:142237099-142237665			
	chr8:1847894-1848163			
	chr8:49426952-49427328			
	chr8:49427546-49427722			
	chr8:49835974-49836286			
	chr9:140023271-140023604			

Appendix C. Regulators with consistent methylation changes between genders of multiple conception groups. Regulators with consistent methylation changes between genders of two or more conception groups are shown. Methylation changes for each regulator in the supporting conception groups are indicated as beta-value changes. Genes associated with the given regulator are named and described in the columns titled "Associated Gene" and "Protein function".

	Regulator	Methylation changes	Count of supporting conception groups	Associated Gene	Protein function
Hyper-methylated	chr15:75018665-75019349	0.06;0.07;0.07;0.05	4	CYP1A1	A member of the cytochrome P450 superfamily, monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids [353].
Hyper-methylated	chr1:39547903-39548129	0.05;0.05;0.04	3	Homo sapiens microtubule-actin crosslinking factor 1 (MACF1)	A protein that forms bridges between cytoskeletal elements, MACF1 impacts microtubule dynamics and associated cellular processes [354].
Hyper-methylated	chr12:95840387-95840581	0.05;0.06;0.04	3	Novel lincRNA (RP11-167N24.3) and RNU6-735P	RNU6-735P is a pseudogene of unknown function.
Hyper-methylated	chr1:75590785-75591155	0.15;0.12;0.11	3	IHX8	A LIM homeobox transcription factor that plays a role in tooth morphogenesis, oogenesis and in neuronal differentiation [355].
Hyper-methylated	chr5:71852712-71853181	0.03;0.04;0.07	3	LOC102503427	ncRNA, validated in RefSeq [356].
Hyper-methylated	chr16:84869763-84870234	0.05;0.03	2	Cysteine-rich secretory protein LCCL domain containing 2 (CRISPLD2)	A secretory protein which promotes matrix assembly and may play a role in non-syndromic cleft palate [357, 358].
Hyper-methylated	chr7:1062588-1063089	0.03;0.03	2	Chromosome 7 open reading frame 50 (C7orf50)	A predicted intracellular protein with unknown function [359].
Hyper-methylated	chr8:49426952-49427328	0.09;0.08	2	Novel lincRNA (RP11-567J20.3) and uncharacterized LOC101929268	Non-coding RNAs with unknown function.
Hyper-methylated	chr8:49427546-49427722	0.09;0.08	2	Novel lincRNA (RP11-567J20.3) and uncharacterized LOC101929268	Non-coding RNAs with unknown function.
Hyper-methylated	chr10:77165098-77165501	0.05;0.07	2	ZNF503 antisense RNA 2 (ZNF503-AS2)	A noncoding antisense RNA [360]
Hyper-methylated	chr17:37896808-37897181	0.03;0.05	2	Growth factor receptor-bound protein 7 (GRB7)	A growth factor receptor-binding protein with roles in integrin signaling and cell migration [361].
Hyper-methylated	chr17:56744054-56744558	0.05;0.09	2	Testis expressed 14 (TEX14)	A protein highly expressed in testis and required for the formation of intercellular bridges during spermatogenesis [362].
Hyper-methylated	chr19:13113447-13113871	0.03;0.05	2	Nuclear factor I/X (CCAAT-binding transcription factor) (NFIX)	A transcription factor that binds viral and cellular promoters [363]
Hyper-methylated	chr7:128094514-128094949	0.04;0.07	2	Hypoxia inducible lipid droplet-	Protein involved in intracellular lipid accumulation [364]

				associated (HILPDA)	
Hyper-methylated	chr10:29228763-29229094	0.09;0.07	2		
Hyper-methylated	chr19:384369-384669	0.04; 0.06	2	Theg spermatid protein (THEG)	A protein expressed in the haploid germ cell nucleus with potential roles in protein assembly [365].
Hyper-methylated	chr2:240362327-240362618	0.08;0.05	2	Histone deacetylase 4 (HDAC4)	A histone deacetylase involved in transcriptional regulation through MEF2C and MEF2D binding [366]
Hypo-methylated	chr21:44573696-44574026	-0.07; -0.10; -0.07	3	Novel lincRNA (AP001631.10) and Homo sapiens crystallin, alpha A (CRYAA)	CRYAA is a (HSP20) family molecular chaperone with autokinase activity, which may be involved in intracellular architecture. This protein is preferentially expressed in the lens. [367]
Hypo-methylated	chr7:157405982-157406183	-0.06; -0.10; -0.07	3	PTPRN2	Protein tyrosine phosphatase (PTP), of the receptor-type PTPs. PTPRN is a major autoantigen associated with insulin-dependent diabetes mellitus [368].
Hypo-methylated	chr12:133021534-133022956	-0.12; -0.07	2	Novel lincRNA (RP11-503G7.1) and Novel lincRNA (RP11-503G7.2) and mucin 8 (MUC8)	MUC8 is expressed in epithelial cells of the human airway, although the function of this gene is still being elucidated [369].
Hypo-methylated	chr12:132903795-132904258	-0.03; -0.05	2	Novel antisense gene (RP13-895J2.7) and Homo sapiens UDP-N-acetyl-alpha-D-galactosamine:poly peptide N-acetylgalactosaminyltransferase 9 (GalNAc-T9) (GALNT9)	GALNT9 and related enzymes initiate O-linked oligosaccharide biosynthesis through transfer of N-acetyl-D-galactosamine residues [370].

Appendix D. Enhancers with consistent bloodspot methylation changes. Enhancers highlighted in blue are consistently altered between all three assisted conditions and natural conception. Statistically significant methylation changes observed in conception comparisons between NAT and IUI, FH, or FZ, are given in units of beta-values, with changes in male and female comparisons shown in blue and red, respectively. Genes associated with the given regulator are named and described in the columns titled "Associated Gene(s)" and "Protein function".

Observed methylation change	Enhancer	Methylation Change	Associated Gene(s)	Protein function	Notes
Hypo-methylated	chr10:14051386-14052060	-0.06;-0.08/-0.10;-0.07;-0.07	FERM domain containing 4A (FRMD4A)	May play roles in cytoskeleton structure and cell polarity through protein binding [371].	Evolutionarily conserved region is located in the intron of FRMD4A and promoter region of an alternative FRMD4A isoform. Region contains many transcription factor binding sites. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr10:21798958-21799135	-0.04;-0.04/-0.09;-0.06;-0.09	Cancer Susceptibility Candidate 10 (CASC10) and SKI/DACH domain containing 1 (SKIDA1)	SKIDA1 exhibits nucleotide binding and may play a role in cancer risk [372].	Site is located approximately 13 kb and 3 kb from start site of C10orf114 and stop site of SKIDA1, respectively. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr11:2919958-2920393	-0.06;-0.05;-0.09/-0.06;-0.05;-0.06	Cyclin-dependent kinase inhibitor 1C (p57, Kip2) (CDKN1C), Solute carrier family 22, member 18 (SLC22A18), and Solute carrier family 22 (organic cation transporter), member 18 antisense (SLC22A18AS)	CDKN1C inhibits several G1 cyclin/Cdk complexes and may be a tumor suppressor. SLC22A18 acts as an organic cation transporter and plays a role in drug transport [373, 374].	Site is located in the intron of SLC22A18AS, is directly upstream of the promoter for SLC22A18, and is approximately 13 kb upstream of the transcriptional start site for CDKN1C. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr17:79068878-79069252	-0.07;-0.06;-0.10/-0.10;-0.08;-0.06	BAI1-associated protein 2 (BAIAP2)	An adaptor protein which aids in the conduction of signals from membrane bound G-proteins to cytoplasmic effector proteins. BAIAP2 may play a role in axonigenesis and insulin signaling in the nervous system [375].	Enhancer is located in the intron of BAIAP2, adjacent to many transcription factor binding sites. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr18:11849903-11850727	-0.08;-0.06;-0.06/-0.05;-0.04	Guanine nucleotide binding protein (G protein), alpha activating activity polypeptide, olfactory type (GNAL), Charged multivesicular body protein 1B (CHMP1B)	GNAL, a stimulatory G protein alpha subunit, modulates signal transduction within the olfactory neuroepithelium and basal ganglia. CHMP1B , a component of the ESCRT-III complex, is likely involved in multivesicular bodies (MVBs) formation [376, 377].	Site is located in the intron of GNAL and is in the promoter region of CHMP1B. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr19:2332363-2332647	-0.11;-0.09;-0.10/-0.07;-0.07;-0.05	Signal peptide peptidase like 2B (SPPL2B), LSM7 homolog, U6 small nuclear RNA associated (S. cerevisiae) (LSM7)	SPPL2B encodes an intramembrane-cleaving aspartic protease (I-CLiP) involved in ITM2B and TNF processing. LSM7 is an Sm-like protein which likely participates in splicing through interaction with U6 snRNA [378, 379].	Site is located in the intron of SPPL2B and is approximately 4 kb upstream of the LSM7 transcriptional start site. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.

Hypo-methylated	chr2:1609558-1609866	-0.09;-0.10/-0.07;-0.08;-0.07	LincRNAs (AC144450.2, AC141930.2), putative antisense gene AC144450.1, Peroxidasin homolog (PXDN)	Secreted heme-containing peroxidase is involved in extracellular matrix formation. Involved in ocular development and may have a systemic role in peroxidase metabolism. [380, 381]	Site is located in the intron of AC144450.1, and is approximately 25 kb, 15 kb, and 26 kb away from AC141930.2, AC144450.2, and PXDN, respectively. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr2:241585712-241586190	-0.10;-0.07/-0.10;-0.06;-0.11	G protein-coupled receptor 35 (GPR35) and Aquaporin 12B (AQP12B)	A receptor for kynurenic acid, GPR35 is involved in intracellular signaling. AQP12B, as a member of the aquaporin family, forms a pore to facilitate transport of water and solutes across cell membranes [382, 383].	Site is located approximately 15 kb and 30 kb from GPR35 and AQP12B, respectively. Site exhibits moderate chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr2:8597253-8597584	-0.07;-0.09;-0.08/-0.08;-0.06;-0.08			Site is located at least 85 kb away from any annotated gene. Site exhibits moderate chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr21:44573696-44574026	-0.06;-0.08/-0.10;-0.11;-0.10	Novel lincRNA (AP001631.10) and Crystallin, alpha A (CRYAA)	CRYAA is a (HSP20) family molecular chaperone with autokinase activity, which may be involved in intracellular architecture. This protein is preferentially expressed in the lens. [367]	Site is located ~ 5kb and 15 kb away from AP001631.10 and CRYAA, respectively. Site exhibits high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr4:3374732-3375130	-0.07;-0.08/-0.06;-0.05;-0.08	Regulator of G-protein signaling 12 (RGS12)	Involved in signal transduction, RGS12 acts inhibits signal transduction and thus acts as a transcriptional repressor [384].	Site is located in intron of RGS12. Site exhibits a high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr8:144948428-144948782	-0.04;-0.04;-0.07/-0.08;-0.11	Epiplakin 1 (EPPK1)	A member of the plakin family, EPPK1 may regulate and maintain keratin intermediate filament networks [385].	Site is located in the unspliced 5' UTR of EPPK1. Site exhibits a high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr1:201619383-201619798	-0.05;-0.05;-0.06/-0.06;-0.05;-0.07	Neuron navigator 1 (NAV1)	A member of the neuron navigator family which may play a role in axon guidance [386]	Site is located in the first intron of NAV1. Site exhibits a high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr15:75018665-75019349	-0.05;-0.05/-0.04;-0.05	CYP1A1	A member of the cytochrome P450 superfamily, monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids [353].	Site is located in the promoter of CYP1A1. Site exhibits a high degree of chromatin accessibility, as determined by DNase Hypersensitivity.
Hypo-methylated	chr5:669501-669935	-0.05;-0.05/-0.05;-0.04;-0.06	Tubulin polymerization promoting protein (TPPP)	Protein involved in polymerization of tubulin and maintenance of the microtubule network [387].	Site is located in the first intron of TPPP. Site exhibits a moderate degree of chromatin accessibility, as determined by DNase Hypersensitivity.

Appendix E: RE discovery computational methods (REDa)

<u>Processing bam files</u>

The RE discovery algorithm is designed to be run entirely in R, with the user providing aligned reads in BAM file format. To conserve memory, the BAM files are first converted to bigWig format. The user is expected to use the helper function "generatebw" to generate both the required bigWigs as well as the text file containing the number of aligned reads per sample. However, the required bigWigs can also be generated by converting BAM files to bedgraph format, using the bedtools tool genomeCoverageBed, with the parameters "-split –bg", and subsequently bigwig format, using the bedGraphToBigWig program (available from the UCSC Genome Browser utilities).

<u>Collapsing redundant exons</u>

Gene isoforms and transcripts often share overlapping exons and UTRs. In the RE discovery tool, overlapping annotated regions (here described as exons, regardless of the coding potential) on the same strand are collapsed into a single loci, designated as "exonic". This singular representation of multi-exonic transcripts is not conducive to isoform discovery, but works well for fragmented RNAs, as is seen in spermatozoa or formalin-fixed paraffin embedded (FFPE) tissues. The "prepareExons" function collapses exons using the "bedR" package.

<u>Discovering expressed regions</u>

The initial step in discovering expressed regions of the genome is performed in the "findRE" function. The user needs to provide their genome of interest in BSgenome format, and a gene annotation file in GTF format. Although the current study performed RE discovery on the hg38 build of the human genome, the algorithm is adaptable for different species and genome builds. There is no restriction on the format of chromosome names used in the genome. However, later annotation steps are coded to be used with ensembl gene annotations available through the R package 'biomaRt', and thus it is recommended that the GTF file uses ensembl gene IDs (please see https://www.gencodegenes.org/releases/current.html for examples). The genome is first processed by binning into 10 bp regions. Coverage across each 10 bp bin is calculated and bins overlapping an annotated region in the GTF file (here described as exons, regardless of the coding potential) are removed from consideration. The remaining genomic bins are compared to the designated library size-normalized threshold mu (μ), with those bins equal to or exceeding the coverage required in μ retained as a novel RNA element (RE). This threshold μ, for a theoretical RNA-seq library with 3 million reads, would require a mean coverage of 7.5 reads for a given 10 bp bin in order to label the bin as expressed. REs within 100 bp of one another are merged into a single new RE. It is important to note that the default parameters used by the "findRE" function retain expressed regions if they are present in at least one sample. REs defined across each sample are subsequently merged with the "combineRE" function, in which the REs from each iteration are concatenated into a single GRanges object. The concatenated object is then reduced, merging overlapping and adjacent (within 50 bp) REs.

<u>Annotating REs</u>

Once the newly discovered REs are generated, they need to be annotated according to their genomic position, which is achieved in the "annotateRE" function. Intronic regions are defined according to the user-supplied GTF file, followed by identifying novel REs that overlap an intron. The distance between any non-intronic novel REs and exons is then calculated and used to define non-intronic novel REs as near-exon (less than 10kb from an exon) or orphan (greater than or equal to 10 kb from an exon).

In the human genome, poorly annotated genes have occasionally been observed to have transcription beyond their designated borders. In an optional step, performed with the "extendExon" function, near-exon REs within a default 20 bp of an exon are examined. The read coverage across near-exon REs, in 10 bp increments, are compared to the average read coverage of the adjacent exon. If the read coverage in the near-exon RE is increased or decreased by less than 50% of the average read coverage of the adjacent exon, the novel RE is merged with the adjacent exon. This process is repeated until no more 10 bp bins of the near-exon RE remain, or the near-exon RE coverage changes by more than 50% of the average read coverage of the adjacent exon.

In order to provide the user with the common gene symbol of each ensembl gene ID, the function "annotateFinal" is provided. The "biomaRt" package is used to transform the ensembl gene ID into the common gene symbol. The output is a complete bed file of exonic and novel REs, along with the common gene symbol(s) and ensembl gene id(s) of the REs. This bed file can subsequently be used in expression analysis.

Appendix F: RNA-seq samples applied to REDa (RE discovery). This table provides the unique sample name (Sample.name) for each RNA-seq library, as well as the corresponding tissue type and RNA-preparation type (Tissue, Tissue.simple, RNA.type and Pool). The publication or GEO dataset from which the sample was drawn is provided in "Publication" column.

Sample.name	Group	Tissue	Tissue.simple	RNA.type	Date	Pool	Publication
Test.Ambion	Krawetz	Testis	Testis	Aplus	2009	pool	doi: 10.1007/s00441-015-2237-1
Test.Clone	Krawetz	Testis	Testis	Aplus	2009	pool	doi: 10.1007/s00441-015-2237-1
testes1	Krawetz	Testis	Testis	Total RNA	2013	pool	doi: 10.1007/s00441-015-2237-1
testes2	Krawetz	Testis	Testis	Total RNA	2013	pool	doi: 10.1007/s00441-015-2237-1
SRR893048	Xue	oocyte	Oocyte	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893049	Xue	pronucleus	Embryo_pronucleus	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893050	Xue	pronucleus	Embryo_pronucleus	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893051	Xue	pronucleus	Embryo_pronucleus	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893052	Xue	zygote	Embryo_zygote	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893053	Xue	zygote	Embryo_zygote	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893054	Xue	2-cell_blastomere	Embryo_2cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893055	Xue	2-cell_blastomere	Embryo_2cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893056	Xue	2-cell_blastomere	Embryo_2cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893057	Xue	4-cell_blastomere	Embryo_4cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893058	Xue	4-cell_blastomere	Embryo_4cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893059	Xue	4-cell_blastomere	Embryo_4cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893060	Xue	4-cell_blastomere	Embryo_4cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893061	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893062	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893063	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893064	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893065	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893066	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893067	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893068	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893069	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893070	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893071	Xue	8-cell_blastomere	Embryo_8cell	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893072	Xue	morula	Embryo_morula	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893073	Xue	morula	Embryo_morula	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893074	Xue	morula	Embryo_morula	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893046	Xue	oocyte	Oocyte	Aplus	2013	single cell	doi:10.1038/nature12364
SRR893047	Xue	oocyte	Oocyte	Aplus	2013	single cell	doi:10.1038/nature12364

SRR2130093	Dang	zygote	Embryo_zygote	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130125	Dang	TE	Embryo_trophoblast	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130124	Dang	TE	Embryo_trophoblast	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130123	Dang	TE	Embryo_trophoblast	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130122	Dang	ICM	Embryo_ICM	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130121	Dang	ICM	Embryo_ICM	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130120	Dang	ICM	Embryo_ICM	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130119	Dang	late_blastocyst	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130118	Dang	late_blastoycst	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130117	Dang	late_blastoycst	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130116	Dang	middle_blastocyst	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130115	Dang	middle_blastocyst	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130114	Dang	middle_blastocyst	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130113	Dang	early_blastocsyt	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130112	Dang	early_blastocsyt	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130111	Dang	early_blastocsyt	Embryo_blastocyst	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130110	Dang	morula	Embryo_morula	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130109	Dang	morula	Embryo_morula	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130108	Dang	morula	Embryo_morula	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130107	Dang	morula	Embryo_morula	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130106	Dang	8_cell	Embryo_8cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130105	Dang	8_cell	Embryo_8cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130104	Dang	8_cell	Embryo_8cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130103	Dang	8_cell	Embryo_8cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130102	Dang	8_cell	Embryo_8cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130101	Dang	4_cell	Embryo_4cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130100	Dang	4_cell	Embryo_4cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130099	Dang	4_cell	Embryo_4cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130098	Dang	4_cell	Embryo_4cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130097	Dang	2_cell	Embryo_2cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130096	Dang	2_cell	Embryo_2cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130095	Dang	2_cell	Embryo_2cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130094	Dang	2_cell	Embryo_2cell	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130092	Dang	zygote	Embryo_zygote	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130091	Dang	zygote	Embryo_zygote	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130090	Dang	zygote	Embryo_zygote	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130089	Dang	zygote	Embryo_zygote	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130088	Dang	MII_oocyte	Oocyte	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130087	Dang	MII_oocyte	Oocyte	Total RNA	2016	single cell	https://doi.org/10.1186/s13059-016-0991-3

SRR2130086	Dang	MII_oocyte	Oocyte	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR2130085	Dang	MII_oocyte	Oocyte	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3062174	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3062173	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3062172	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3062171	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3501436	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3501437	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
SRR3501438	Dang	human_stem_cells	ESC	Total RNA	20 16	single cell	https://doi.org/10.1186/s13059-016-0991-3
sperm_A	Gisselmann	Sperm	Sperm	Aplus	20 16	pool	doi:10.1038/srep32255
sperm_B	Gisselmann	Sperm	Sperm	Aplus	20 16	pool	doi:10.1038/srep32255
sperm_C	Gisselmann	Sperm	Sperm	Aplus	20 16	pool	doi:10.1038/srep32255
sperm_stranded_pooled	Gisselmann	Sperm	Sperm	Total RNA	20 16	pool	doi:10.1038/srep32255
SRR3192419	Gingeras	liver.female.6yo	Liver	Aplus	20 16	pool	GEO Accession: GSM2072373
SRR3192418	Gingeras	liver.male.32yo	Liver	Aplus	20 16	pool	GEO Accession: GSM2072372
SRR3192440	Gingeras	liver.fetal	Liver	Aplus	20 16	pool	GEO Accession: GSM2072387
SRR4422587	Gingeras	Testis	Testis	Aplus	20 16	pool	GEO Accession: GSM2343115
SRR4422588	Gingeras	Testis	Testis	Aplus	20 16	pool	GEO Accession: GSM2343115
SRR4421667	Gingeras	Testis	Testis	Aplus	20 16	pool	GEO Accession: GSM2343589
SRR4421668	Gingeras	Testis	Testis	Aplus	20 16	pool	GEO Accession: GSM2343589
SRR3192439	Gingeras	liver.fetal	Liver	Aplus	20 16	pool	GEO Accession: GSM2072386
SRR1791304	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR1791305	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR1791306	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR1791307	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR1791308	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR1791309	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR1791310	Jodar	Sperm	Sperm	Total RNA	20 13	pool	DOI: 10.1126/scitranslmed.aab1287
SRR3146452	Jan	A.dark.spermatogonia	Adark_sperm	Total RNA	20 16	pool	doi: 10.1242/dev.152413
SRR3146494	Jan	A.pale.spermatogonia	Apale_sperm	Total RNA	20 16	pool	doi: 10.1242/dev.152413
SRR3146497	Jan	early.pachytene.spermatocytes	Earlypach_sperm	Total RNA	20 16	pool	doi: 10.1242/dev.152413
SRR3146504	Jan	leptotene.zygotene.spermatocytes	Lept.zygo_sperm	Total RNA	20 16	pool	doi: 10.1242/dev.152413
SRR3146507	Jan	late.pachytene.spermatocytes	Latepach_sperm	Total RNA	20 16	pool	doi: 10.1242/dev.152413
SRR3146508	Jan	round.spermatids	Round_sperm	Total RNA	20 16	pool	doi: 10.1242/dev.152413

Appendix G: Location of novel REs exceeding 1 kb in length. A total of 138 novel REs exceed one kilobase in length. The chromosomal location (in hg38 coordinates) of the RE is provided in the first three columns, "Chromosome", "Start position", and "End position". The length of the RE, in base pairs, is provided in column "RE length". For intronic and near-exon REs, the associated gene names are provided in column "Gene symbol". The designation of an RE as near-exon, intronic, or novel is given in column "RE class".

Chromosome	Start position	End position	RE length	Gene symbol	RE class
chr9	133019427	133021126	1700	EEF1A1P5	NOVEL_10KB_EXON
chr5	4867045	4868064	1020	AC026415.1	NOVEL_INTRONIC
chr1	1595371	1596770	1400	FNDC10	NOVEL_10KB_EXON
chr1	25907041	25908250	1210	STMN1	NOVEL_10KB_EXON
chr1	37859741	37862430	2690	INPP5B	NOVEL_10KB_EXON
chr1	40860961	40862366	1406	CITED4	NOVEL_10KB_EXON
chr1	180190981	180192010	1030	QSOX1	NOVEL_INTRONIC
chr2	19351349	19352410	1062	OSR1	NOVEL_10KB_EXON
chr2	44167539	44168859	1321	AC019129.2	NOVEL_10KB_EXON
chr2	132270379	132271658	1280	CDC27P1	NOVEL_10KB_EXON
chr3	50258368	50259819	1452	GNAI2	NOVEL_10KB_EXON
chr4	15002500	15004335	1836	CPEB2	NOVEL_10KB_EXON
chr4	118630440	118634189	3750	AC110079.1	NOVEL_10KB_EXON
chr5	177456545	177458707	2163	DBN1	NOVEL_10KB_EXON
chr6	2966400	2972205	5806	SERPINB6	NOVEL_10KB_EXON
chr7	905617	906626	1010	ADAP1	NOVEL_INTRONIC
chr7	65838216	65839335	1120	AC093582.1	NOVEL_10KB_EXON
chr7	65840796	65842065	1270	AC093582.1	NOVEL_10KB_EXON
chr9	136107687	136109081	1395	AL138781.1	NOVEL_10KB_EXON
chr10	26645650	26646779	1130	AL390961.1	NOVEL_10KB_EXON
chr14	70230224	70232587	2364	AL160191.3	NOVEL_10KB_EXON
chr14	106286318	106287657	1340	HOMER2P2	NOVEL_10KB_EXON
chr17	7881186	7882765	1580	NAA38	NOVEL_INTRONIC
chr19	4227969	4229118	1150	EBI3	NOVEL_10KB_EXON
chr22	23892523	23893922	1400	AP000350.4	NOVEL_INTRONIC
chr22	30967066	30968552	1487	MORC2	NOVEL_10KB_EXON
chr22	41994621	41998652	4032	SEPT3	NOVEL_10KB_EXON
chr1	38129381	38130522	1142	AL139158.3	NOVEL_10KB_EXON
chr1	149924351	149925820	1470	SF3B4	NOVEL_INTRONIC
chr1	154952201	154953600	1400	PBXIP1	NOVEL_INTRONIC
chr2	112062629	112063728	1100	TMEM87B	NOVEL_INTRONIC
chr3	168845049	168846468	1420	NA	NOVEL_ORPHAN
chr6	36863796	36865095	1300	PPIL1	NOVEL_INTRONIC
chr7	30157438	30162876	5439	MTURN,AC007036.3	NOVEL_10KB_EXON
chr8	99944163	99946062	1900	NA	NOVEL_ORPHAN
chr9	34667387	34668426	1040	AL162231.2	NOVEL_INTRONIC
chr9	35103667	35105374	1708	FAM214B	NOVEL_10KB_EXON
chr11	70121258	70122757	1500	ANO1	NOVEL_INTRONIC
chr12	2855495	2856944	1450	ITFG2	NOVEL_INTRONIC
chr15	22767380	22770269	2890	AC138649.1	NOVEL_INTRONIC
chr16	22115491	22116500	1010	VWA3A	NOVEL_INTRONIC
chr19	4618609	4619678	1070	AC011498.5	NOVEL_10KB_EXON
chr19	50497439	50500158	2720	EMC10	NOVEL_10KB_EXON
chr20	48501613	48516852	15240	RNU7-144P	NOVEL_10KB_EXON
chr1	2604371	2605540	1170	MMEL1	NOVEL_INTRONIC
chr1	178514771	178515820	1050	TEX35	NOVEL_INTRONIC

chr1	222471921	222473380	1460	CICP13	NOVEL_10KB_EXON
chr1	222473491	222475320	1830	AL513314.2	NOVEL_10KB_EXON
chr1	222476031	222477304	1274	AL513314.2	NOVEL_10KB_EXON
chr6	23854156	23855645	1490	AL139093.1	NOVEL_10KB_EXON
chr7	149587	152547	2961	AC093627.4	NOVEL_10KB_EXON
chr7	942327	943466	1140	ADAP1	NOVEL_INTRONIC
chr7	51390017	51391366	1350	AC012441.1	NOVEL_10KB_EXON
chr7	51391797	51393186	1390	AC012441.1	NOVEL_10KB_EXON
chr7	63930996	63932015	1020	SLC25A1P3	NOVEL_10KB_EXON
chr7	112618816	112620645	1830	AC002463.1	NOVEL_10KB_EXON
chr10	132887000	132888069	1070	CFAP46	NOVEL_INTRONIC
chr16	1672491	1673510	1020	CRAMP1	NOVEL_INTRONIC
chr16	2528236	2531510	3275	AMDHD2	NOVEL_10KB_EXON
chr17	3385930	3386955	1026	AC087498.1	NOVEL_10KB_EXON
chr20	50958623	50963931	5309	MOCS3	NOVEL_10KB_EXON
chr1	24086831	24087900	1070	MYOM3	NOVEL_INTRONIC
chr1	154206511	154207648	1138	C1orf43	NOVEL_10KB_EXON
chr2	45309399	45311588	2190	LINC01121	NOVEL_INTRONIC
chr2	165665809	165667138	1330	CSRNP3	NOVEL_INTRONIC
chr2	216857249	216858478	1230	AC007563.2	NOVEL_INTRONIC
chr3	169821889	169822941	1053	LRRIQ4	NOVEL_10KB_EXON
chr3	195941779	195945588	3810	AC124944.3	NOVEL_INTRONIC
chr3	196001229	196005068	3840	AC024937.2	NOVEL_10KB_EXON
chr3	197640019	197642188	2170	AC024560.3	NOVEL_10KB_EXON
chr4	10083250	10084379	1130	WDR1	NOVEL_INTRONIC
chr4	119405650	119407129	1480	GTF2IP12	NOVEL_10KB_EXON
chr4	119407310	119409332	2023	GTF2IP12	NOVEL_10KB_EXON
chr4	183850820	183852569	1750	STOX2	NOVEL_INTRONIC
chr4	189846080	189847239	1160	FRG1-DT	NOVEL_INTRONIC
chr5	1620305	1624854	4550	AC026412.1	NOVEL_INTRONIC
chr5	7299375	7301459	2085	AC091951.1	NOVEL_10KB_EXON
chr5	70487825	70491084	3260	AC146944.3	NOVEL_10KB_EXON
chr5	70509795	70511814	2020	AC146944.1	NOVEL_10KB_EXON
chr5	70774825	70778714	3890	AC139272.1	NOVEL_INTRONIC
chr6	166899956	166902745	2790	RPS6KA2	NOVEL_INTRONIC
chr7	32591137	32592176	1040	DPY19L1P1	NOVEL_INTRONIC
chr7	37847147	37848366	1220	EPDR1	NOVEL_INTRONIC
chr7	57216146	57218355	2210	AC099654.1	NOVEL_INTRONIC
chr7	57219716	57221165	1450	AC099654.1	NOVEL_INTRONIC
chr7	132561286	132564405	3120	PLXNA4	NOVEL_INTRONIC
chr7	143579446	143580725	1280	AC073264.2	NOVEL_10KB_EXON
chr7	143748736	143750055	1320	RNU6-267P	NOVEL_10KB_EXON
chr8	27011147	27012422	1276	AC067904.1	NOVEL_10KB_EXON
chr8	42451183	42452682	1500	SLC20A2	NOVEL_INTRONIC
chr8	54049383	54050712	1330	LYPLA1	NOVEL_INTRONIC
chr9	68971447	68972466	1020	PIP5K1B	NOVEL_INTRONIC
chr9	88129307	88132710	3404	SPATA31C2	NOVEL_10KB_EXON
chr10	671190	672949	1760	DIP2C	NOVEL_INTRONIC
chr10	46129883	46131429	1547	AGAP7P	NOVEL_10KB_EXON
chr10	46357239	46358769	1531	AGAP14P	NOVEL_10KB_EXON
chr10	131385470	131386809	1340	NA	NOVEL_ORPHAN
chr11	119397667	119398776	1110	USP2-AS1	NOVEL_INTRONIC
chr12	97988365	97989664	1300	MIR4303	NOVEL_10KB_EXON
chr13	40481146	40483205	2060	LINC00598	NOVEL_INTRONIC
chr13	111317126	111318245	1120	TEX29	NOVEL_INTRONIC
chr15	23439460	23441682	2223	GOLGA6L2	NOVEL_10KB_EXON

chr16	90166011	90167060	1050	FAM157C	NOVEL_INTRONIC
chr17	68148996	68150055	1060	LRRC37A16P	NOVEL_10KB_EXON
chr17	81580005	81581034	1030	NPLOC4	NOVEL_INTRONIC
chr18	73325334	73326733	1400	LINC02582	NOVEL_INTRONIC
chr19	23322559	23323958	1400	ZNF91	NOVEL_INTRONIC
chr19	23465429	23469978	4550	AC074140.1	NOVEL_10KB_EXON
chr19	23878309	23882468	4160	AC139769.2	NOVEL_10KB_EXON
chr19	24179389	24181548	2160	NA	NOVEL_ORPHAN
chr19	29015399	29016728	1330	LINC01532	NOVEL_10KB_EXON
chr19	46918489	46922356	3868	ARHGAP35	NOVEL_10KB_EXON
chr19	50039299	50040528	1230	ZNF473	NOVEL_INTRONIC
chr20	1817803	1819042	1240	AL121760.1	NOVEL_10KB_EXON
chr20	5483953	5485022	1070	AL121757.1	NOVEL_10KB_EXON
chr20	47833573	47834972	1400	RNU7-173P	NOVEL_10KB_EXON
chr20	47894373	47897142	2770	RNU7-92P	NOVEL_10KB_EXON
chr20	47897423	47898942	1520	RNU7-92P	NOVEL_10KB_EXON
chr20	48122623	48126802	4180	NA	NOVEL_ORPHAN
chr20	48473243	48475192	1950	AL049541.1	NOVEL_10KB_EXON
chr20	48499743	48501212	1470	RNU7-144P	NOVEL_10KB_EXON
chr21	28993116	28994175	1060	LTN1	NOVEL_10KB_EXON
chrX	527285	528594	1310	FABP5P13	NOVEL_10KB_EXON
chrX	622205	623384	1180	SHOX	NOVEL_10KB_EXON
chrX	625535	627004	1470	SHOX	NOVEL_INTRONIC
chrX	942355	943544	1190	NA	NOVEL_ORPHAN
chrX	1910245	1911684	1440	NA	NOVEL_ORPHAN
chr2	241808169	241809368	1200	AC131097.3	NOVEL_10KB_EXON
chr15	64951720	64952949	1230	ANKDD1A	NOVEL_INTRONIC
chr3	195644509	195646888	2380	AC233280.2	NOVEL_10KB_EXON
chr21	8228686	8230745	2060	FP671120.1	NOVEL_10KB_EXON
chr21	8455846	8457845	2000	NA	NOVEL_ORPHAN
chrY	527180	528589	1410	FABP5P13	NOVEL_10KB_EXON
chrY	622200	623389	1190	SHOX	NOVEL_10KB_EXON
chrY	625540	627009	1470	SHOX	NOVEL_INTRONIC
chrY	942350	943549	1200	NA	NOVEL_ORPHAN
chrY	1910130	1911589	1460	NA	NOVEL_ORPHAN
chrY	2213400	2214439	1040	DHRSX	NOVEL_10KB_EXON

Appendix H: Specifications for quantitative quality control of MARS samples.

Step 1: If the sample has less than 20% of the aligned reads belonging to the human genome and less than 2 million autosomal reads (duplicates included), then the sample is assigned as "Category 1: Low Genomic". Otherwise, the sample continues to step 2.

Step 2: If the sample has greater than 75% of the aligned reads belonging to the human genome and one of the two following conditions: Read duplication rate less than 10% or less than 5% of aligned reads belonging to ribosomal RNA, then the sample is assigned as "Category 2: High intergenic reads". Otherwise, the sample continues to step 3.

Step 3: If the sample has a % of aligned reads belonging to the human genome between 20% and 40%, and greater than 4% of the aligned reads belonging to the bacteria or viruses, then the sample is assigned as "Category 3: High bacterial and viral reads". Otherwise, the sample continues to step 4.

Step 4: If the sample has a % of aligned reads belonging to the human genome between 20% and 40%, and greater than 50% of the aligned reads belonging to the ERCC spike-in, then the sample is assigned as "Category 4: High spike-in reads". Otherwise, the sample continues to step 5.

Step 5: If the sample has a % of aligned reads belonging to the human genome between 20% and 40%, and greater than 40% of the reads remaining unmapped, then the sample is assigned as "Category 5: High unmapped reads". Otherwise, the sample has passed quality control and is determined to be of satisfactory quality.

The R function for the above process is shown below:

```
makelab <- function(x){ ##where x is a row number, corresponding to a particular sample
  a <- ar{ar$sample==x,}
  if(a$genomic_prop<0.2 & a$autosome<2000000){
    b <- "Cond1_lowgenomic"}
  else if(a$genomic_prop>0.75 & (a$duplicates_prop < 0.1 | a$rRNA_prop < 0.05)){
    b <- "Cond2_grassy"}
  else if(a$genomic_prop>=0.2 & a$genomic_prop<=0.4 & a$bactviral_prop>0.04){
    b <- "Cond3_highbactviral"}
  else if(a$genomic_prop>=0.2 & a$genomic_prop<=0.4 & a$ercc_prop>0.5){
    b <- "Cond4_highercc"}
  else if(a$genomic_prop>=0.2 & a$genomic_prop<=0.4 & a$unmapped_prop>0.4){
    b <- "Cond5_highunmapped"}
  else {
    b <- "OK"
    }
    return(b)
}
```

Appendix I: Differential expression of sperm-enriched genomic repeats. The number of differential REs that overlap a given repeat are indicated for (A) Simple repeats, (B) Complex repeats, and (C) Centromeric repeats. X-axis provides the repeat name, while the Y-axis indicates the number of differential REs for each significant expression change.
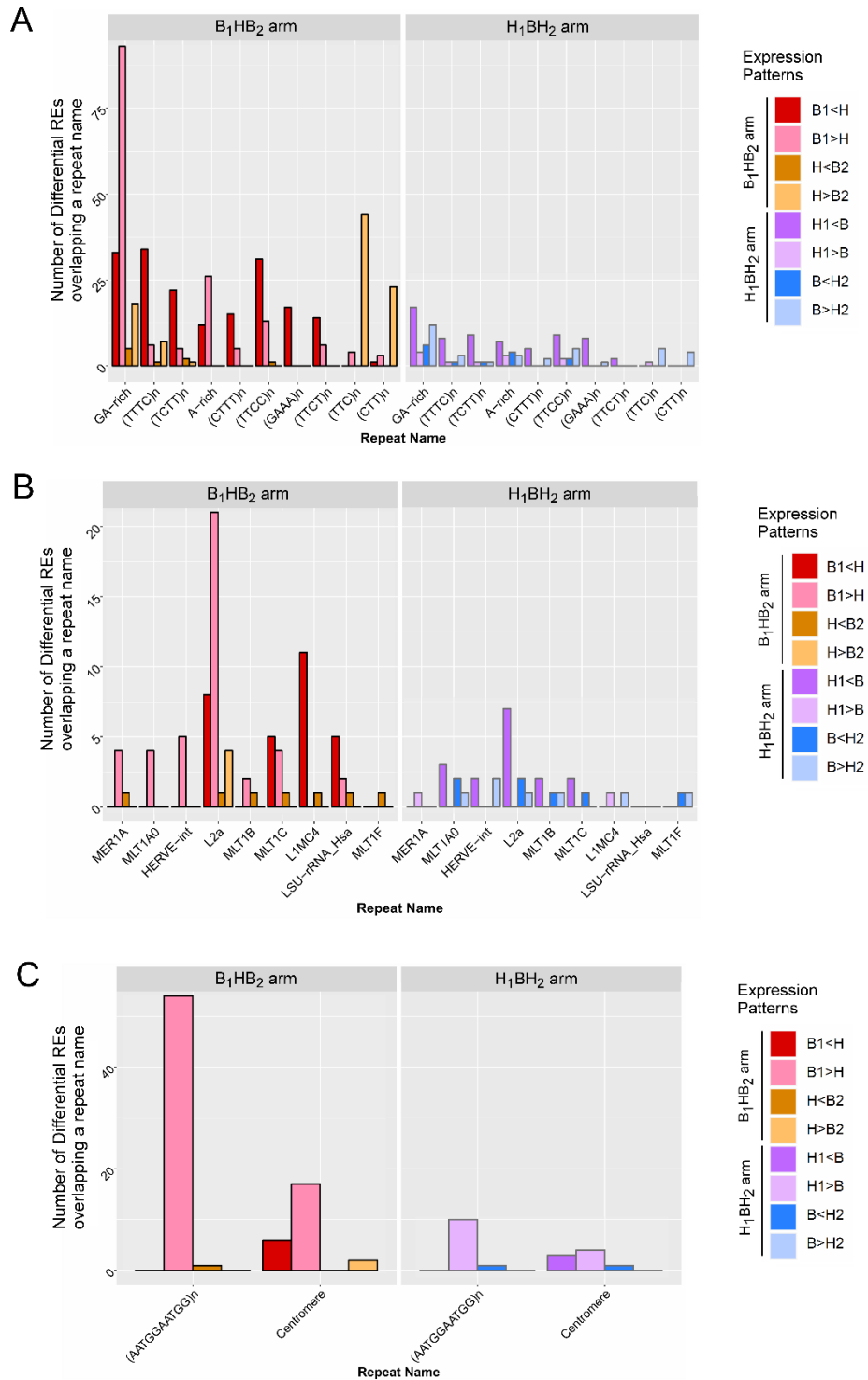
Appendix J. piRNA clusters with multiple piRNAs expressed in human sperm. The cluster location and name are indicated in the columns "Chromosome", "Start", "End", "Cluster width", and "Cluster_name". "Cluster expression in Testis" provides the expression details from the pooled generic testis dataset, from which the clusters were generated. The column "Count of expressed piRNAs" indicates the number of piRNAs present in human sperm (median RPM > 1 RPM) belonging to the given cluster.

Chromosome	Start	End	Cluster Width	Cluster name	Cluster expression in Testis	Count of expressed piRNAs
chr19	16011114	16044967	33854	cluster_168	normal_testis_generic|10711.0258613575|mono:plus|168	6
chr3	12848046	12879013	30968	cluster_234	normal_testis_generic|1513.99429755676|mono:minus|234	6
chr1	179582014	179593814	11801	cluster_21	normal_testis_generic|1903.72089572736|mono:minus|21	3
chr10	28786000	28797031	11032	cluster_27	normal_testis_generic|7153.61881881805|mono:minus|27	3
chr15	51246002	51308028	62027	cluster_99	normal_testis_generic|14407.8869660576|mono:minus|99	3
chr15	96761025	96790009	28985	cluster_111	normal_testis_generic|3644.61909348552|mono:minus|111	3
chr4	119403036	119414007	10972	cluster_248	normal_testis_generic|218.559754300489|mono:minus|248	3
chr9	81910020	81933023	23004	cluster_330	normal_testis_generic|1269.70452451239|mono:minus|330	3
chr10	79680067	79719007	38941	cluster_40	normal_testis_generic|535.912720047788|mono:minus|40	2
chr11	45666000	45730027	64028	cluster_51	normal_testis_generic|23143.2207101714|mono:minus|51	2
chr11	75003029	75014999	11971	cluster_56	normal_testis_generic|1048.76550357495|mono:plus|56	2
chr6	33863011	33925970	62960	cluster_271	normal_testis_generic|34453.9318019432|bi:minus-plus (split between 33892867 and 33892867)|271	2
chr9	113101011	113122392	21382	cluster_340	normal_testis_generic|11059.4403415613|bi:minus-plus (split between 113112111 and 113112118)|340	2
chrX	9400016	9419006	18991	cluster_351	normal_testis_generic|1353.62783508312|mono:minus|351	2

Appendix K. Small RNAs altered by DBP. Cells highlighted in lime green exhibit the same expression trend in both study arms, while those highlighted in light orange exhibit opposite expression trends.

B₁HB₂ arm		H₁BH₂ arm		
High-DBP > Non-DBP	High-DBP < Non-DBP	High-DBP > Non-DBP	High-DBP < Non-DBP	
L1M2_5	hsa_piR_016677	7SL	ALRb	hsa_piR_008397
AluYg6	hsa_piR_016742	GOLEM	AluSq	hsa_piR_009228
CHARLIE3	hsa-miR-192-5p	hsa-miR-339-5p	AluSq10	hsa_piR_013247
	hsa_piR_019675	hsa_piR_005076	AluYf1	hsa_piR_016582
	hsa-miR-200a-3p	hsa_piR_010010	AluYk12	hsa_piR_016584
	tRNA-Gln-CAA	hsa_piR_011482	CHARLIE3	hsa_piR_016804
	hsa-miR-186-5p	hsa_piR_017724	ENSG00000199313|ENST00000362443	hsa_piR_017550
	hsa-miR-27b-3p	MER28	ENSG00000222094|ENST00000410162	hsa_piR_017591
	CHARLIE10	TRNA_GLU	ENSG00000252461|ENST00000516652	hsa_piR_017781
	MER54A		ENSG00000252677|ENST00000516868	hsa_piR_017990
	hsa-miR-499a-5p		HERV16	hsa_piR_017996
	hsa_piR_016735		HERV-Fc1	hsa_piR_018717
			HERVK3I	hsa_piR_018790
			hsa-let-7d-3p	hsa_piR_019675
			hsa-miR-142-5p	hsa_piR_020497
			hsa-miR-151b	hsa_piR_021041
			hsa-miR-28-3p	hsa_piR_021722
			hsa-miR-320a	hsa_piR_022107
			hsa-miR-3656	hsa_piR_023221
			hsa-miR-423-5p	hsa_piR_023224
			hsa-miR-486-5p	hsa_piR_023415
			hsa-miR-508-5p	L1M3C_5
			hsa-miR-574-5p	L1MB7
			hsa-miR-891a-5p	L1MC3
			hsa_piR_000753	L1P4d_5end
			hsa_piR_001809	LTR10C
			hsa_piR_003116	LTR16A
			hsa_piR_003180	LTR1F1
			hsa_piR_003220	LTR30
			hsa_piR_003257	LTR66
			hsa_piR_004427	MER54
			hsa_piR_004880	MER66_I
			hsa_piR_005278	MER72
			hsa_piR_005371	MLT1D
			hsa_piR_005675	MST_I
			hsa_piR_005767	SVA_A
			hsa_piR_006426	THE1D
			hsa_piR_006434	tRNA-Ala-GCA
			hsa_piR_008114	

REFERENCES

1. Almond, D. and J. Currie, *Killing Me Softly: The Fetal Origins Hypothesis.* The journal of economic perspectives : a journal of the American Economic Association, 2011. **25**(3): p. 153-172.

2. Mandy, M. and M. Nyirenda, *Developmental Origins of Health and Disease: the relevance to developing nations.* International health, 2018. **10**(2): p. 66-70.

3. Oestreich, A.K. and K.H. Moley, *Developmental and Transmittable Origins of Obesity-Associated Health Disorders.* Trends in Genetics, 2017. **33**(6): p. 399-407.

4. Varcin, K.J., et al., *Prenatal maternal stress events and phenotypic outcomes in Autism Spectrum Disorder.* Autism Research, 2017. **10**(11): p. 1866-1877.

5. Lin, Y., et al., *Effects of prenatal and postnatal maternal emotional stress on toddlers' cognitive and temperamental development.* Journal of Affective Disorders, 2017. **207**: p. 9-17.

6. Korja, R., et al., *The Relations Between Maternal Prenatal Anxiety or Stress and Child's Early Negative Reactivity or Self-Regulation: A Systematic Review.* Child Psychiatry & Human Development, 2017. **48**(6): p. 851-869.

7. Lowensohn, R.I., D.D. Stadler, and C.R. Naze, *Current Concepts of Maternal Nutrition.* Obstetrical & Gynecological Survey, 2016. **71**(7): p. 413–426.

8. Macpherson, A.J., M.G. de Agüero, and S.C. Ganal-Vonarburg, *How nutrition and the maternal microbiota shape the neonatal immune system.* Nature Reviews Immunology, 2017. **17**: p. 508.

9. Kim, J.H. and A.R. Scialli, *Thalidomide: The Tragedy of Birth Defects and the Effective Treatment of Disease.* Toxicological Sciences, 2011. **122**(1): p. 1-6.

10.   Newbold, R.R., *Lessons learned from perinatal exposure to diethylstilbestrol.* Toxicology and Applied Pharmacology, 2004. **199**(2): p. 142-150.

11.   Grandjean, V., et al., *RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders.* Scientific Reports, 2015. **5**: p. 18193.

12.   Öst, A., et al., *Paternal Diet Defines Offspring Chromatin State and Intergenerational Obesity.* Cell, 2014. **159**(6): p. 1352-1364.

13.   Short, A.K., et al., *Exercise alters mouse sperm small noncoding RNAs and induces a transgenerational modification of male offspring conditioned fear and anxiety.* Translational Psychiatry, 2017. **7**: p. e1114.

14.   Morgan, C.P. and T.L. Bale, *Early Prenatal Stress Epigenetically Programs Dysmasculinization in Second-Generation Offspring via the Paternal Lineage.* The Journal of Neuroscience, 2011. **31**(33): p. 11748.

15.   Short, A.K., et al., *Elevated paternal glucocorticoid exposure alters the small noncoding RNA profile in sperm and modifies anxiety and depressive phenotypes in the offspring.* Translational Psychiatry, 2016. **6**: p. e837.

16.   Bygren, L.O., et al., *Change in paternal grandmothers´ early food supply influenced cardiovascular mortality of the female grandchildren.* BMC Genetics, 2014. **15**(1): p. 12.

17.   Vågerö, D., et al., *Paternal grandfather's access to food predicts all-cause and cancer mortality in grandsons.* Nature Communications, 2018. **9**(1): p. 5124.

18.   Veenendaal, M., et al., *Transgenerational effects of prenatal exposure to the 1944–45 Dutch famine.* BJOG: An International Journal of Obstetrics & Gynaecology, 2013. **120**(5): p. 548-554.

19.   Estill, M.S. and S.A. Krawetz, *The Epigenetic Consequences of Paternal Exposure to Environmental Contaminants and Reproductive Toxicants.* Current Environmental Health Reports, 2016. **3**(3): p. 202-213.

20.   Mima, M., D. Greenwald, and S. Ohlander, *Environmental Toxins and Male Fertility.* Current Urology Reports, 2018. **19**(7): p. 50.

21.   Patel, A.S., J.Y. Leong, and R. Ramasamy, *Prediction of male infertility by the World Health Organization laboratory manual for assessment of semen analysis: A systematic review.* Arab Journal of Urology, 2018. **16**(1): p. 96-102.

22.   Malić Vončina, S., et al., *Sperm DNA fragmentation and mitochondrial membrane potential combined are better for predicting natural conception than standard sperm parameters.* Fertility and Sterility, 2016. **105**(3): p. 637-644.e1.

23.   Luo, S., et al., *Gestational and lactational exposure to low-dose bisphenol A increases Th17 cells in mice offspring.* Environmental Toxicology and Pharmacology, 2016. **47**: p. 149-158.

24.   Brehm, E., et al., *Prenatal Exposure to Di(2-Ethylhexyl) Phthalate Causes Long-Term Transgenerational Effects on Female Reproduction in Mice.* Endocrinology, 2017. **159**(2): p. 795-809.

25.   Chamorro-Garcia, R., et al., *Ancestral perinatal obesogen exposure results in a transgenerational thrifty phenotype in mice.* Nature Communications, 2017. **8**(1): p. 2012.

26.   Wang, Y., et al., *Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins.* Epigenetics, 2018. **13**(9): p. 975-987.

27.   Kochmanski, J., et al., *Age-related epigenome-wide DNA methylation and hydroxymethylation in longitudinal mouse blood.* Epigenetics, 2018. **13**(7): p. 779-792.

28.   Wang, Y., N.L. Pedersen, and S. Hägg, *Implementing a method for studying longitudinal DNA methylation variability in association with age.* Epigenetics, 2018. **13**(8): p. 866-874.

29. Schübeler, D., *Function and information content of DNA methylation.* Nature, 2015. **517**: p. 321.

30. Louis, J.F., et al., *The prevalence of couple infertility in the United States from a male perspective: evidence from a nationally representative sample.* Andrology, 2013. **1**(5): p. 741-748.

31. Edgell, T.A., et al., *Fresh versus frozen embryo transfer: backing clinical decisions with scientific and clinical evidence.* Human Reproduction Update, 2014. **20**(6): p. 808-821.

32. Reed, B.G. and B.R. Carr, *The Normal Menstrual Cycle and the Control of Ovulation.* Endotext [Internet], ed. K. Feingold, B. Anawalt, and A. Boyce. 2018, South Dartmouth (MA): MDText.com, Inc.

33. Edwards, R.G., *IVF, IVM, natural cycle IVF, minimal stimulation IVF – time for a rethink.* Reproductive BioMedicine Online, 2007. **15**(1): p. 106-119.

34. Haemmerli Keller, K., et al., *Treatment-related psychological stress in different in vitro fertilization therapies with and without gonadotropin stimulation.* Acta Obstetricia et Gynecologica Scandinavica, 2018. **97**(3): p. 269-276.

35. Krawetz, S.A., *Paternal contribution: new insights and future challenges.* Nature Reviews Genetics, 2005. **6**: p. 633.

36. Amann, R.P., *The Cycle of the Seminiferous Epithelium in Humans: A Need to Revisit?* Journal of Andrology, 2008. **29**(5): p. 469-487.

37. Rowley, M.J., F. Teshima, and C.G. Heller, *Duration of Transit of Spermatozoa through the Human Male Ductular System**This investigation was made possible by grants-in-aid from the Atomic Energy Commission, AT(45-1)1780, and the National Institutes of Health, (HD 00804).* Fertility and Sterility, 1970. **21**(5): p. 390-396.

38.   Gervasi, M.G. and P.E. Visconti, *Molecular changes and signaling events occurring in spermatozoa during epididymal maturation.* Andrology, 2017. **5**(2): p. 204-218.

39.   Hammoud, S.S., et al., *Distinctive chromatin in human sperm packages genes for embryo development.* Nature, 2009. **460**: p. 473.

40.   Wykes, S.M. and S.A. Krawetz, *The Structural Organization of Sperm Chromatin.* Journal of Biological Chemistry, 2003. **278**(32): p. 29471-29477.

41.   Martins, R.P. and S.A. Krawetz, *Nuclear organization of the protamine locus.* Soc Reprod Fertil Suppl, 2007. **64**: p. 1-12.

42.   Yoshida, K., et al., *Mapping of histone-binding sites in histone replacement-completed spermatozoa.* Nature Communications, 2018. **9**(1): p. 3885.

43.   Castillo, J., M. Jodar, and R. Oliva, *The contribution of human sperm proteins to the development and epigenome of the preimplantation embryo.* Human Reproduction Update, 2018. **24**(5): p. 535-555.

44.   Spadafora, C., *Sperm-Mediated Transgenerational Inheritance.* Frontiers in Microbiology, 2017. **8**(2401).

45.   Nixon, B., et al., *Proteomic profiling of mouse epididymosomes reveals their contributions to post-testicular sperm maturation.* Molecular &amp; Cellular Proteomics, 2018: p. mcp.RA118.000946.

46.   Donkin, I., et al., *Obesity and Bariatric Surgery Drive Epigenetic Variation of Spermatozoa in Humans.* Cell Metabolism, 2016. **23**(2): p. 369-378.

47.   Ingerslev, L.R., et al., *Endurance training remodels sperm-borne small RNA expression and methylation at neurological gene hotspots.* Clinical Epigenetics, 2018. **10**(1): p. 12.

48.   Kaprara, A. and I.T. Huhtaniemi, *The hypothalamus-pituitary-gonad axis: Tales of mice and men.* Metabolism, 2018. **86**: p. 3-17.

49.   O'Shaughnessy, P.J., *Hormonal control of germ cell development and spermatogenesis.* Seminars in Cell & Developmental Biology, 2014. **29**: p. 55-65.

50.   Gold, E., F.E. Marino, and G. Risbridger, *The inhibin/activin signalling pathway in human gonadal and adrenal cancers.* MHR: Basic science of reproductive medicine, 2014. **20**(12): p. 1223-1237.

51.   Atlas, T.H.P. *INHBA*. 2019 February 1, 2019]; Available from: https://www.proteinatlas.org/ENSG00000122641-INHBA/tissue.

52.   Dwyer, A.A., T. Raivio, and N. Pitteloud, *Gonadotrophin replacement for induction of fertility in hypogonadal men.* Best Practice & Research Clinical Endocrinology & Metabolism, 2015. **29**(1): p. 91-103.

53.   Isidori, A.M., E. Giannetta, and A. Lenzi, *Male hypogonadism.* Pituitary, 2008. **11**(2): p. 171.

54.   Hauser, R., et al., *Male Reproductive Disorders, Diseases, and Costs of Exposure to Endocrine-Disrupting Chemicals in the European Union.* The Journal of Clinical Endocrinology & Metabolism, 2015. **100**(4): p. 1267-1277.

55.   Gore, A.C., et al., *Executive Summary to EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals.* Endocrine Reviews, 2015. **36**(6): p. 593-602.

56.   Hait, E.J., A.M. Calafat, and R. Hauser, *Urinary phthalate metabolite concentrations among men with inflammatory bowel disease on mesalamine therapy.* Endocr Disruptors (Austin), 2014. **1**(1).

57.   Hauser, R., et al., *Medications as a source of human exposure to phthalates.* Environ Health Perspect, 2004. **112**(6): p. 751-3.

58.   Kratochvil, I., et al., *Mono(2-ethylhexyl) phthalate (MEHP) and mono(2-ethyl-5-oxohexyl) phthalate (MEOHP) but not di(2-ethylhexyl) phthalate (DEHP) bind productively to the*

peroxisome proliferator-activated receptor γ. Rapid Communications in Mass Spectrometry, 2018. **0**(0).

59.   Laurenzana, E.M., et al., *Activation of the Constitutive Androstane Receptor by Monophthalates.* Chemical Research in Toxicology, 2016. **29**(10): p. 1651-1661.

60.   Hurst, C.H. and D.J. Waxman, *Activation of PPARα and PPARγ by Environmental Phthalate Monoesters.* Toxicological Sciences, 2003. **74**(2): p. 297-308.

61.   Veeramachaneni, D.N.R. and R.K. Gary, *Phthalate-induced pathology in the foetal testis involves more than decreased testosterone production.* REPRODUCTION, 2014. **147**(4): p. 435-442.

62.   Vandenberg, L.N., et al., *Human exposure to bisphenol A (BPA).* Reproductive Toxicology, 2007. **24**(2): p. 139-177.

63.   Chen, Z., et al., *Long-term exposure to a 'safe' dose of bisphenol A reduced protein acetylation in adult rat testes.* Scientific Reports, 2017. **7**: p. 40337.

64.   Wang, C., et al., *The classic EDCs, phthalate esters and organochlorines, in relation to abnormal sperm quality: a systematic review with meta-analysis.* Scientific Reports, 2016. **6**: p. 19982.

65.   Duty, S.M., et al., *Phthalate Exposure and Human Semen Parameters.* Epidemiology, 2003. **14**(3): p. 269-277.

66.   Hallas, J., et al., *Association between use of phthalate-containing medication and semen quality among men in couples referred for assisted reproduction.* Human Reproduction, 2018. **33**(3): p. 503-511.

67.   Wu, H., et al., *Preconception urinary phthalate concentrations and sperm DNA methylation profiles among men undergoing IVF treatment: a cross-sectional study.* Human Reproduction, 2017. **32**(11): p. 2159-2169.

68.   Hauser, R., et al., *DNA damage in human sperm is related to urinary levels of phthalate monoester and oxidative metabolites.* Human Reproduction, 2007. **22**(3): p. 688-695.

69.   Nassan, F.L., et al., *A crossover–crossback prospective study of dibutyl-phthalate exposure from mesalamine medications and semen quality in men with inflammatory bowel disease.* Environment International, 2016. **95**: p. 120-130.

70.   Nassan, F.L., et al., *A crossover–crossback prospective study of dibutyl-phthalate exposure from mesalamine medications and serum reproductive hormones in men.* Environmental Research, 2018. **160**: p. 121-131.

71.   Sharma, U., et al., *Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals.* Science, 2016. **351**(6271): p. 391.

72.   Chen, Q., et al., *Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder.* Science, 2016. **351**(6271): p. 397.

73.   Rodgers, A.B., et al., *Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress.* Proceedings of the National Academy of Sciences, 2015. **112**(44): p. 13699.

74.   Brison, D.R., S.A. Roberts, and S.J. Kimber, *How should we assess the safety of IVF technologies?* Reproductive BioMedicine Online, 2013. **27**(6): p. 710-721.

75.   Lazaraviciute, G., et al., *A systematic review and meta-analysis of DNA methylation levels and imprinting disorders in children conceived by IVF/ICSI compared with children conceived spontaneously.* Human Reproduction Update, 2014. **20**(6): p. 840-852.

76.   Estill, M.S., et al., *Assisted reproductive technology alters deoxyribonucleic acid methylation profiles in bloodspots of newborn infants.* Fertility and Sterility, 2016. **106**(3): p. 629-639.e10.

77.   Estill, M.S., R. Hauser, and S.A. Krawetz, *RNA element discovery from germ cell to blastocyst.* Nucleic Acids Research, 2018. **47**(5): p. 2263-2275.

78.   Wadhwa, P.D., et al., *Developmental Origins of Health and Disease: Brief History of the Approach and Current Focus on Epigenetic Mechanisms.* Semin Reprod Med, 2009. **27**(05): p. 358-368.

79.   Dominguez-Salas, P., et al., *Maternal nutrition at conception modulates DNA methylation of human metastable epialleles.* Nature Communications, 2014. **5**: p. 3746.

80.   Waterland, R.A., et al., *Season of Conception in Rural Gambia Affects DNA Methylation at Putative Human Metastable Epialleles.* PLOS Genetics, 2010. **6**(12): p. e1001252.

81.   Donjacour, A., et al., *Effect of ICSI on gene expression and development of mouse preimplantation embryos.* Human Reproduction, 2010. **25**(12): p. 3012-3024.

82.   Giritharan, G., et al., *In Vitro Culture of Mouse Embryos Reduces Differential Gene Expression Between Inner Cell Mass and Trophectoderm.* Reproductive Sciences, 2012. **19**(3): p. 243-252.

83.   Le, F., et al., *In Vitro Fertilization Alters Growth and Expression of Igf2/H19 and Their Epigenetic Mechanisms in the Liver and Skeletal Muscle of Newborn and Elder Mice1.* Biology of Reproduction, 2013. **88**(3).

84.   Chen, M., et al., *Impaired Glucose Metabolism in Response to High Fat Diet in Female Mice Conceived by In Vitro Fertilization (IVF) or Ovarian Stimulation Alone.* PLOS ONE, 2014. **9**(11): p. e113155.

85.   Feuer, S.K., L. Camarano, and P.F. Rinaudo, *ART and health: clinical outcomes and insights on molecular mechanisms from rodent studies.* MHR: Basic science of reproductive medicine, 2012. **19**(4): p. 189-204.

86.   Donjacour, A., et al., *Sexually Dimorphic Effect of In Vitro Fertilization (IVF) on Adult Mouse Fat and Liver Metabolomes.* Endocrinology, 2014. **155**(11): p. 4554-4567.

87.   Donjacour, A., et al., *Use of a Mouse In Vitro Fertilization Model to Understand the Developmental Origins of Health and Disease Hypothesis.* Endocrinology, 2014. **155**(5): p. 1956-1969.

88.   Krapp, C., et al., *In Vitro Culture Increases the Frequency of Stochastic Epigenetic Errors at Imprinted Genes in Placental Tissues from Mouse Concepti Produced Through Assisted Reproductive Technologies.* Biology of Reproduction, 2014. **90**(2).

89.   Davies, M.J., et al., *Reproductive Technologies and the Risk of Birth Defects.* New England Journal of Medicine, 2012. **366**(19): p. 1803-1813.

90.   Kochanski, A., et al., *The impact of assisted reproductive technologies on the genome and epigenome of the newborn.* Journal of neonatal-perinatal medicine, 2013. **6**(2): p. 101-108.

91.   Chiba, H., et al., *DNA methylation errors in imprinting disorders and assisted reproductive technology.* Pediatrics International, 2013. **55**(5): p. 542-549.

92.   Delemarre-van de Waal, H.A., et al., *Cardiometabolic Differences in Children Born After in Vitro Fertilization: Follow-Up Study.* The Journal of Clinical Endocrinology & Metabolism, 2008. **93**(5): p. 1682-1688.

93.   Prein, J., et al., *Growth during infancy and early childhood in relation to blood pressure and body fat measures at age 8–18 years of IVF children and spontaneously conceived controls born to subfertile parents.* Human Reproduction, 2009. **24**(11): p. 2788-2795.

94.   Scherrer, U., et al., *Systemic and Pulmonary Vascular Dysfunction in Children Conceived by Assisted Reproductive Technologies.* Circulation, 2012. **125**(15): p. 1890-1896.

95.   Turan, N., et al., *DNA methylation and gene expression differences in children conceived in vitro or in vivo.* Human Molecular Genetics, 2009. **18**(20): p. 3769-3778.

96. Loke, Y.J., et al., *Association of in vitro fertilization with global and IGF2/H19 methylation variation in newborn twins.* Journal of Developmental Origins of Health and Disease, 2015. **6**(2): p. 115-124.

97. Lou, H., et al., *Assisted reproductive technologies impair the expression and methylation of insulin-induced gene 1 and sterol regulatory element-binding factor 1 in the fetus and placenta.* Fertility and Sterility, 2014. **101**(4): p. 974-980.e2.

98. Melamed, N., et al., *Comparison of genome-wide and gene-specific DNA methylation between ART and naturally conceived pregnancies.* Epigenetics, 2015. **10**(6): p. 474-483.

99. Song, S., et al., *DNA methylation differences between in vitro- and in vivo-conceived children are associated with ART procedures rather than infertility.* Clinical Epigenetics, 2015. **7**(1): p. 41.

100. Whitelaw, N., et al., *Epigenetic status in the offspring of spontaneous and assisted conception.* Human Reproduction, 2014. **29**(7): p. 1452-1458.

101. Feng, C., et al., *General imprinting status is stable in assisted reproduction–conceived offspring.* Fertility and Sterility, 2011. **96**(6): p. 1417-1423.e9.

102. Manning, M., et al., *Study of DNA-methylation patterns at chromosome 15q11-q13 in children born after ICSI reveals no imprinting defects.* Molecular Human Reproduction, 2000. **6**(11): p. 1049-1053.

103. Oliver, V.F., et al., *Defects in imprinting and genome-wide DNA methylation are not common in the in vitro fertilization population.* Fertility and Sterility, 2012. **97**(1): p. 147-153.e7.

104. Tierling, S., et al., *Assisted reproductive technologies do not enhance the variability of DNA methylation imprints in human.* Journal of Medical Genetics, 2010. **47**(6): p. 371.

105. Sen, A., et al., *Early life lead exposure causes gender-specific changes in the DNA methylation profile of DNA extracted from dried blood spots.* Epigenomics, 2015. **7**(3): p. 379-393.

106. Broberg, K., et al., *Arsenic exposure in early pregnancy alters genome-wide DNA methylation in cord blood, particularly in boys.* Journal of Developmental Origins of Health and Disease, 2014. **5**(4): p. 288-298.

107. Goodrich, J.M., et al., *Quality control and statistical modeling for environmental epigenetics: A study on in utero lead exposure and DNA methylation at birth.* Epigenetics, 2015. **10**(1): p. 19-30.

108. Huen, K., et al., *Effects of age, sex, and persistent organic pollutants on DNA methylation in children.* Environmental and Molecular Mutagenesis, 2014. **55**(3): p. 209-222.

109. Feber, A., et al., *ChAMP: 450k Chip Analysis Methylation Pipeline.* Bioinformatics, 2013. **30**(3): p. 428-430.

110. Incorporated, I. *HumanMethylation450 v1.2 Manifest File*. [cited 2016 April 27]; Available from: http://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html.

111. Price, E.M., et al., *Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array.* Epigenetics & Chromatin, 2013. **6**(1): p. 4.

112. Baccarelli, A.A., et al., *A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure.* Bioinformatics, 2013. **29**(22): p. 2884-2891.

113. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.

114. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**: p. 455.

115. Yang, X., et al., *Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer.* Cancer Cell, 2014. **26**(4): p. 577-590.

116. Smallwood, A. and B. Ren, *Genome organization and long-range regulation of gene expression by enhancers.* Current Opinion in Cell Biology, 2013. **25**(3): p. 387-394.

117. Franco, M.M., A.R. Prickett, and R.J. Oakey, *The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction1.* Biology of Reproduction, 2014. **91**(5).

118. Kolovos, P., et al., *Enhancers and silencers: an integrated and simple model for their function.* Epigenetics & Chromatin, 2012. **5**(1): p. 1.

119. Michael, D., et al., *Microarray analysis of sexually dimorphic gene expression in human minor salivary glands.* Oral Diseases, 2011. **17**(7): p. 653-661.

120. Ishida, M. and G.E. Moore, *The role of imprinted genes in humans.* Molecular Aspects of Medicine, 2013. **34**(4): p. 826-840.

121. Court, F., et al., *Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment.* Genome Research, 2014. **24**(4): p. 554-569.

122. White, C.R., et al., *High Frequency of Imprinted Methylation Errors in Human Preimplantation Embryos.* Scientific Reports, 2015. **5**: p. 17311.

123. Nordin, M., et al., *Epigenetic regulation of the Igf2/H19 gene cluster.* Cell Proliferation, 2014. **47**(3): p. 189-199.

124. Rakyan, V.K., et al., *Metastable epialleles in mammals.* Trends in Genetics, 2002. **18**(7): p. 348-351.

125. Waterland, R.A. and R.L. Jirtle, *Transposable Elements: Targets for Early Nutritional Effects on Epigenetic Gene Regulation.* Molecular and Cellular Biology, 2003. **23**(15): p. 5293.

126. Waterland, R.A., et al., *Maternal methyl supplements increase offspring DNA methylation at Axin fused.* genesis, 2006. **44**(9): p. 401-406.

127. Silver, M.J., et al., *Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment.* Genome Biology, 2015. **16**(1): p. 118.

128. Sanchez-Mut, J.V., et al., *Promoter hypermethylation of the phosphatase DUSP22 mediates PKA-dependent TAU phosphorylation and CREB activation in Alzheimer's disease.* Hippocampus, 2014. **24**(4): p. 363-368.

129. Sekine, Y., et al., *DUSP22/LMW-DSP2 regulates estrogen receptor-α-mediated signaling through dephosphorylation of Ser-118.* Oncogene, 2007. **26**: p. 6038.

130. Sekine, Y., et al., *Regulation of STAT3-mediated signaling by LMW-DSP2.* Oncogene, 2006. **25**: p. 5801.

131. Eddy, E.M., M. Goto, and D.A. O'Brien, *Speriolin is a novel human and mouse sperm centrosome protein.* Human Reproduction, 2010. **25**(8): p. 1884-1894.

132. Martin-Trujillo, A., et al., *Stability of Genomic Imprinting and Gestational-Age Dynamic Methylation in Complicated Pregnancies Conceived Following Assisted Reproductive Technologies.* Biology of Reproduction, 2013. **89**(3).

133. Yousefi, P., et al., *Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies.* Environmental and Molecular Mutagenesis, 2015. **56**(9): p. 751-758.

134.   Talens, R.P., et al., *Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology.* The FASEB Journal, 2010. **24**(9): p. 3135-3144.

135.   Jaffe, A.E. and R.A. Irizarry, *Accounting for cellular heterogeneity is critical in epigenome-wide association studies.* Genome Biology, 2014. **15**(2): p. R31.

136.   Lonsdale, J., et al., *The Genotype-Tissue Expression (GTEx) project.* Nature Genetics, 2013. **45**: p. 580.

137.   Uhlen, M., et al., *A pathology atlas of the human cancer transcriptome.* Science, 2017. **357**(6352).

138.   GTEx. *Release V7 (dbGaP Accession phs000424.v7.p2)*. 2018 March 29, 2018]; Available from: https://www.gtexportal.org/.

139.   EMBL-EBI. *Expression Atlas*. Available from: https://www.ebi.ac.uk/gxa/home/.

140.   ENCODE. *ENCODE: Encyclopedia of DNA Elements*. Available from: https://www.encodeproject.org/.

141.   Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature Biotechnology, 2010. **28**: p. 511.

142.   O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Research, 2016. **44**(D1): p. D733-D745.

143.   Harrow, J., et al., *GENCODE: The reference human genome annotation for The ENCODE Project.* Genome Research, 2012. **22**(9): p. 1760-1774.

144.   GENCODE. *https://www.gencodegenes.org/*. Available from: https://www.gencodegenes.org/.

145.   Esteller, M., *Non-coding RNAs in human disease.* Nature Reviews Genetics, 2011. **12**: p. 861.

146.   Harries, Lorna W., *Long non-coding RNAs and human disease.* Biochemical Society

        Transactions, 2012. **40**(4): p. 902.

147.   Salzman, J., et al., *Cell-Type Specific Features of Circular RNA Expression.* PLOS Genetics,

        2013. **9**(9): p. e1003777.

148.   Kim, T.-K., M. Hemberg, and J.M. Gray, *Enhancer RNAs: A Class of Long Noncoding RNAs*

        Synthesized at Enhancers. Cold Spring Harbor Perspectives in Biology, 2015. **7**(1).

149.   Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**: p. 101.

150.   Wu, H., et al., *Tissue-Specific RNA Expression Marks Distant-Acting Developmental*

        Enhancers. PLOS Genetics, 2014. **10**(9): p. e1004610.

151.   Kang, H.J., et al., *Spatio-temporal transcriptome of the human brain.* Nature, 2011. **478**: p.

        483.

152.   Svoboda, P., V. Franke, and R.M. Schultz, *Chapter Nine - Sculpting the Transcriptome During*

        the Oocyte-to-Embryo Transition in Mouse, in *Current Topics in Developmental Biology*, H.D.

        Lipshitz, Editor. 2015, Academic Press. p. 305-349.

153.   Fagerberg, L., et al., *Analysis of the Human Tissue-specific Expression by Genome-wide*

        Integration of Transcriptomics and Antibody-based Proteomics. Molecular & Cellular

        Proteomics, 2014. **13**(2): p. 397-406.

154.   Sendler, E., et al., *Stability, delivery and functions of human sperm RNAs at fertilization.*

        Nucleic Acids Research, 2013. **41**(7): p. 4104-4117.

155.   Johnson, G.D., et al., *Cleavage of rRNA ensures translational cessation in sperm at*

        fertilization. MHR: Basic science of reproductive medicine, 2011. **17**(12): p. 721-726.

156.   Johnson, G.D., et al., *Chromatin and extracellular vesicle associated sperm RNAs.* Nucleic

        Acids Research, 2015. **43**(14): p. 6847-6859.

157.	Jodar, M., A. Soler-Ventura, and R. Oliva, *Semen proteomics and male infertility.* Journal of Proteomics, 2017. **162**: p. 125-134.

158.	Cossetti, C., et al., *Soma-to-Germline Transmission of RNA in Mice Xenografted with Human Tumour Cells: Possible Transport by Exosomes.* PLOS ONE, 2014. **9**(7): p. e101629.

159.	Devanapally, S., S. Ravikumar, and A.M. Jose, *Double-stranded RNA made in C. elegans neurons can enter the germline and cause transgenerational gene silencing.* Proceedings of the National Academy of Sciences, 2015. **112**(7): p. 2133.

160.	Gòdia, M., G. Swanson, and S.A. Krawetz, *A History of Why Fathers' RNA Matters.* Biology of Reproduction, 2018: p. ioy007-ioy007.

161.	Jodar, M., E. Sendler, and S.A. Krawetz, *The protein and transcript profiles of human semen.* Cell and Tissue Research, 2016. **363**(1): p. 85-96.

162.	Krawetz, S.A., et al., *A survey of small RNAs in human sperm.* Human Reproduction (Oxford, England), 2011. **26**(12): p. 3401-3412.

163.	Jodar, M., et al., *Absence of sperm RNA elements correlates with idiopathic male infertility.* Science Translational Medicine, 2015. **7**(295): p. 295re6.

164.	Burl, R.B., et al., *Sperm RNA elements as markers of health.* Syst Biol Reprod Med, 2018. **64**(1): p. 25-38.

165.	Platts, A.E., et al., *Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs.* Human Molecular Genetics, 2007. **16**(7): p. 763-773.

166.	Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature Biotechnology, 2015. **33**: p. 290.

167.	Xue, Z., et al., *Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing.* Nature, 2013. **500**: p. 593.

168. Dang, Y., et al., *Tracing the expression of circular RNAs in human pre-implantation embryos.* Genome Biology, 2016. **17**(1): p. 130.

169. Bates, D., et al., *Fitting Linear Mixed-Effects Models Using lme4.* 2015, 2015. **67**(1): p. 48.

170. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

171. Flegel, C., et al., *Characterization of the Olfactory Receptors Expressed in Human Spermatozoa.* Frontiers in Molecular Biosciences, 2016. **2**(73).

172. Shaffer, J.P., *Multiple Hypothesis Testing.* Annual Review of Psychology, 1995. **46**(1): p. 561-584.

173. Kumar, L. and M. E. Futschik, *Mfuzz: A software package for soft clustering of microarray data.* Bioinformation, 2007. **2**(1): p. 5-7.

174. FUTSCHIK, M.E. and B. CARLISLE, *NOISE-ROBUST SOFT CLUSTERING OF GENE EXPRESSION TIME-COURSE DATA.* Journal of Bioinformatics and Computational Biology, 2005. **03**(04): p. 965-988.

175. Jan, S.Z., et al., *Unraveling transcriptome dynamics in human spermatogenesis.* Development, 2017. **144**(20): p. 3659.

176. Kalmar, A., et al., *Gene expression analysis of normal and colorectal cancer tissue samples from fresh frozen and matched formalin-fixed, paraffin-embedded (FFPE) specimens after manual and automated RNA isolation.* Methods, 2013. **59**(1): p. S16-S19.

177. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**: p. 75.

178. The, E.P.C., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**: p. 57.

179.   Sabo, P.J., et al., *Discovery of functional noncoding elements by digital analysis of chromatin structure.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(48): p. 16837-16842.

180.   Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Research, 2012. **22**(9): p. 1798-1812.

181.   Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**: p. 91.

182.   Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**: p. 376.

183.   Rosenkranz, D. and H. Zischler, *proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis.* BMC Bioinformatics, 2012. **13**(1): p. 5.

184.   Rosenkranz, D., *piRNA cluster database: a web resource for piRNA producing loci.* Nucleic Acids Research, 2016. **44**(D1): p. D223-D230.

185.   Smit, A., R. Hubley, and P. Green. RepeatMasker Open-3.0 1996-2010; Available from: http://www.repeatmasker.org.

186.   Smit, A.F., *Identification of a new, abundant superfamily of mammalian LTR-transposons.* Nucleic Acids Research, 1993. **21**(8): p. 1863-1872.

187.   DFAM. *MSTC (DF0001044)*. March 20,2018; Available from: http://dfam.org/entry/DF0001044.

188.   Franke, V., et al., *Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes.* Genome Research, 2017. **27**(8): p. 1384-1394.

189.   Raz, T., et al., *Protocol Dependence of Sequencing-Based Gene Expression Measurements.* PLOS ONE, 2011. **6**(5): p. e19287.

190. Schrom, E.-M., et al., *Chapter One - Regulation of Retroviral Polyadenylation*, in *Advances in Virus Research*, K. Maramorosch and F.A. Murphy, Editors. 2013, Academic Press. p. 1-24.

191. Borodulina, O.R., et al., *Polyadenylation of RNA transcribed from mammalian SINEs by RNA polymerase III: Complex requirements for nucleotide sequences.* Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 2016. **1859**(2): p. 355-365.

192. Krane, D.E. and R.C. Hardison, *Short interspersed repeats in rabbit DNA can provide functional polyadenylation signals.* Molecular Biology and Evolution, 1990. **7**(1): p. 1-8.

193. Heui-Soo, K., *Genomic Impact, Chromosomal Distribution and Transcriptional Regulation of HERV Elements.* Mol. Cells, 2012. **33**(6): p. 539-544.

194. Curinha, A., et al., *Implications of polyadenylation in health and disease.* Nucleus, 2014. **5**(6): p. 508-519.

195. Jachowicz, J.W., et al., *LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo.* Nature Genetics, 2017. **49**: p. 1502.

196. Lysiak, J.J., *The role of tumor necrosis factor-alpha and interleukin-1 in the mammalian testis and their involvement in testicular torsion and autoimmune orchitis.* Reprod Biol Endocrinol, 2004. **2**: p. 9.

197. Garcia, T.X. and M.C. Hofmann, *Regulation of germ line stem cell homeostasis.* Anim Reprod, 2015. **12**(1): p. 35-45.

198. Koch, S., et al., *Post-transcriptional Wnt Signaling Governs Epididymal Sperm Maturation.* Cell, 2015. **163**(5): p. 1225-1236.

199. Kerr, G.E., et al., *Regulated Wnt/Beta-Catenin Signaling Sustains Adult Spermatogenesis in Mice1.* Biology of Reproduction, 2014. **90**(1): p. 3, 1-12-3, 1-12.

200. De Robertis, Edward M. and D. Ploper, *Sperm Motility Requires Wnt/GSK3 Stabilization of Proteins.* Developmental Cell, 2015. **35**(4): p. 401-402.

201. Vidal, F., et al., *Gene trap analysis of germ cell signaling to Sertoli cells: NGF-TrkA mediated induction of Fra1 and Fos by post-meiotic germ cells.* Journal of Cell Science, 2001. **114**(2): p. 435-443.

202. Jin, W., et al., *Effect of NGF on the Motility and Acrosome Reaction of Golden Hamster Spermatozoa <i>In Vitro</i>.* Journal of Reproduction and Development, 2010. **56**(4): p. 437-443.

203. Michailov, Y., D. Ickowicz, and H. Breitbart, *Zn2+-stimulation of sperm capacitation and of the acrosome reaction is mediated by EGFR activation.* Developmental Biology, 2014. **396**(2): p. 246-255.

204. Shahar, S., et al., *Activation of Sperm EGFR by Light Irradiation is Mediated by Reactive Oxygen Species.* Photochemistry and Photobiology, 2014. **90**(5): p. 1077-1083.

205. Ducummon, C.C. and T. Berger, *Localization of the Rho GTPases and some Rho effector proteins in the sperm of several mammalian species.* Zygote, 2006. **14**(3): p. 249-257.

206. Irino, Y., et al., *Phospholipase Cδ4 Associates with Glutamate Receptor Interacting Protein 1 in Testis.* The Journal of Biochemistry, 2005. **138**(4): p. 451-456.

207. Ortiz-Ramírez, C., et al., *GLUTAMATE RECEPTOR-LIKE channels are essential for chemotaxis and reproduction in mosses.* Nature, 2017. **549**: p. 91.

208. Margolin, G., et al., *Integrated transcriptome analysis of mouse spermatogenesis.* BMC Genomics, 2014. **15**(1): p. 39.

209. Mulugeta Achame, E., et al., *Evaluating the Relationship between Spermatogenic Silencing of the X Chromosome and Evolution of the Y Chromosome in Chimpanzee and Human.* PLOS ONE, 2010. **5**(12): p. e15598.

210. Sin, H.-S., et al., *Human postmeiotic sex chromatin and its impact on sex chromosome evolution.* Genome Research, 2012. **22**(5): p. 827-836.

211. Ostermeier, G.C., et al., *Reproductive biology: delivering spermatozoan RNA to the oocyte.* Nature, 2004. **429**(6988): p. 154.

212. Krawetz, S.A., *Paternal contribution: new insights and future challenges.* Nat Rev Genet, 2005. **6**(8): p. 633-42.

213. Sachani, S., *Nucleoporin-mediated regulation of the Kcnq1ot1 imprinted domain*, in *Biochemistry*. 2016, University of Western Ontario: Electronic Thesis and Dissertation Repository. p. 3962.

214. Tsukamoto, S., A. Kuma, and N. Mizushima, *The role of autophagy during the oocyte-to-embryo transition.* Autophagy, 2008. **4**(8): p. 1076-1078.

215. Ntostis, P., et al., *Potential sperm contributions to the murine zygote predicted by in silico analysis.* Reproduction, 2017. **154**(6): p. 777-788.

216. Lee, M.T., A.R. Bonneau, and A.J. Giraldez, *Zygotic Genome Activation During the Maternal-to-Zygotic Transition.* Annual Review of Cell and Developmental Biology, 2014. **30**(1): p. 581-613.

217. Jodar, M., et al., *The presence, role and clinical use of spermatozoal RNAs.* Human Reproduction Update, 2013. **19**(6): p. 604-624.

218. Werber, M., et al., *The tissue-specific transcriptomic landscape of the mid-gestational mouse embryo.* Development, 2014. **141**(11): p. 2325-2330.

219. Kojima, Y., O.H. Tam, and P.P.L. Tam, *Timing of developmental events in the early mouse embryo.* Seminars in Cell & Developmental Biology, 2014. **34**: p. 65-75.

220. Cao, S., et al., *Specific gene-regulation networks during the pre-implantation development of the pig embryo as revealed by deep sequencing.* BMC Genomics, 2014. **15**(1): p. 4.

221. Fan, X., et al., *Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos.* Genome Biology, 2015. **16**(1): p. 148.

222.   Yan, L., et al., *Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells.* Nature Structural &Amp; Molecular Biology, 2013. **20**: p. 1131.

223.   Team, R.C., *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

224.   Sozen, B., A. Can, and N. Demir, *Cell fate regulation during preimplantation development: A view of adhesion-linked molecular interactions.* Developmental Biology, 2014. **395**(1): p. 73-83.

225.   Miller, D., *Confrontation, Consolidation, and Recognition: The Oocyte's Perspective on the Incoming Sperm.* Cold Spring Harbor Perspectives in Medicine, 2015. **5**(8).

226.   Youngson, N.A., et al., *Obesity-induced sperm DNA methylation changes at satellite repeats are reprogrammed in rat offspring.* Asian Journal of Andrology, 2016. **18**(6): p. 930-936.

227.   Shea, Jeremy M., et al., *Genetic and Epigenetic Variation, but Not Diet, Shape the Sperm Methylome.* Developmental Cell, 2015. **35**(6): p. 750-758.

228.   Catasti, P., et al., *DNA repeats in the human genome.* Genetica, 1999. **106**(1): p. 15-36.

229.   Yaron, Y., et al., *Centromere sequences localize to the nuclear halo of human spermatozoa.* Int J Androl, 1998. **21**(1): p. 13-8.

230.   Linnemann, A., *Analysis of Nuclear Scaffold/Matrix Attachment: The Role of Genome Organization in Transcription:*

Chapter 5: Microarray and High Throughput Genomic Sequencing to Analyze Sperm MARs, in *Ph.D. Thesis : Analysis of Nuclear Scaffold/Matrix Attachment: The Role of Genome Organization in Transcription*. 2009, Wayne State University: Center for Molecular Medicine and Genetics. p. 184.

231. Spadafora, C., *A LINE-1–encoded reverse transcriptase–dependent regulatory mechanism is active in embryogenesis and tumorigenesis.* Annals of the New York Academy of Sciences, 2015. **1341**(1): p. 164-171.

232. Giordano, R., et al., *Reverse Transcriptase Activity in Mature Spermatozoa of Mouse.* The Journal of Cell Biology, 2000. **148**(6): p. 1107.

233. Theunissen, Thorold W., et al., *Molecular Criteria for Defining the Naive Human Pluripotent State.* Cell Stem Cell, 2016. **19**(4): p. 502-515.

234. Spadafora, C., *The "evolutionary field" hypothesis. Non-Mendelian transgenerational inheritance mediates diversification and evolution.* Progress in Biophysics and Molecular Biology, 2018. **134**: p. 27-37.

235. Davidson, E.H. and R.J. Britten, *Regulation of Gene Expression: Possible Role of Repetitive Sequences.* Science, 1979. **204**(4397): p. 1052-1059.

236. Britten, R.J. and E.H. Davidson, *Gene Regulation for Higher Cells: A Theory.* Science, 1969. **165**(3891): p. 349.

237. Johnson, G.D., et al., *Nuclease Footprints in Sperm Project Past and Future Chromatin Regulatory Events.* Scientific Reports, 2016. **6**: p. 25864.

238. Ioannou, D., et al., *A new model of sperm nuclear architecture following assessment of the organization of centromeres and telomeres in three-dimensions.* Scientific Reports, 2017. **7**: p. 41585.

239. Velazquez Camacho, O., et al., *Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA:DNA hybrid formation.* eLife, 2017. **6**: p. e25293.

240.	Scherthan, H., et al., *Contrasting behavior of heterochromatic and euchromatic chromosome portions and pericentric genome separation in pre-bouquet spermatocytes of hybrid mice.* Chromosoma, 2014. **123**(6): p. 609-624.

241.	*Dibutyl phthalate; CASRN 84-74-2*, in *Integrated Risk Information System (IRIS) Chemical Assessment Summary*, U.S.E.P. Agency, Editor. 1987.

242.	GmbH, W.C.D., *Prescribing information for ASACOL (mesalamine) delayed-release tablets*, W.C.D. GmbH, Editor. 2015:

	https://www.accessdata.fda.gov/drugsatfda_docs/label/2015/019651s025lbl.pdf.

243.	Allergan, I., *ASACOL HD- mesalamine tablet, delayed release* 2013:

	https://dailymed.nlm.nih.gov/dailymed/.

244.	Dias, B.G. and K.J. Ressler, *Parental olfactory experience influences behavior and neural structure in subsequent generations.* Nature Neuroscience, 2013. **17**: p. 89.

245.	Estill, M.S., R. Hauser, and S.A. Krawetz, *RNA element discovery from germ cell to blastocyst.* Nucleic Acids Research, 2018: p. gky1223-gky1223.

246.	Goodrich, R., E. Anton, and S.A. Krawetz, *Isolating mRNA and small noncoding RNAs from human sperm*, in *Methods in Molecular Biology*. 2013. p. 385-96.

247.	Goodrich, R.J., G.C. Ostermeier, and S.A. Krawetz, *Multitasking with molecular dynamics Typhoon: quantifying nucleic acids and autoradiographs.* Biotechnol Lett, 2003. **25**(13): p. 1061-5.

248.	Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.

249.	Rueda, A., et al., *sRNAtoolbox: an integrated collection of small RNA research tools.* Nucleic Acids Research, 2015. **43**(W1): p. W467-W473.

250.   Ha, M. and V.N. Kim, *Regulation of microRNA biogenesis.* Nature Reviews Molecular Cell Biology, 2014. **15**: p. 509.

251.   Iwasaki, Y.W., M.C. Siomi, and H. Siomi, *PIWI-Interacting RNA: Its Biogenesis and Functions.* Annual Review of Biochemistry, 2015. **84**(1): p. 405-433.

252.   Mao, S., et al., *Evaluation of the effectiveness of semen storage and sperm purification methods for spermatozoa transcript profiling.* Systems Biology in Reproductive Medicine, 2013. **59**(5): p. 287-295.

253.   Mao, S., et al., *A comparison of sperm RNA-seq methods.* Systems Biology in Reproductive Medicine, 2014. **60**(5): p. 308-315.

254.   Jodar, M., et al., *Response to Comment on "Absence of sperm RNA elements correlates with idiopathic male infertility".* Science Translational Medicine, 2016. **8**(353): p. 353tr1.

255.   Gupta, A., et al., *Gene expression profiles in peripheral blood mononuclear cells correlate with salience network activity in chronic visceral pain: A pilot study.* Neurogastroenterology & Motility, 2017. **29**(6): p. e13027.

256.   Iborra, M., et al., *Identification of serum and tissue micro-RNA expression profiles in different stages of inflammatory bowel disease.* Clinical & Experimental Immunology, 2013. **173**(2): p. 250-258.

257.   Hübenthal, M., et al., *Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease.* PLOS ONE, 2015. **10**(10): p. e0140155.

258.   Gurram, B., et al., *Plasma-induced signatures reveal an extracellular milieu possessing an immunoregulatory bias in treatment-naive paediatric inflammatory bowel disease.* Clinical & Experimental Immunology, 2016. **184**(1): p. 36-49.

259.   Crawley, Scott W., et al., *Intestinal Brush Border Assembly Driven by Protocadherin-Based Intermicrovillar Adhesion.* Cell, 2014. **157**(2): p. 433-446.

260. Okazaki, N., et al., *Protocadherin LKC, a new candidate for a tumor suppressor of colon and liver cancers, its association with contact inhibition of cell proliferation.* Carcinogenesis, 2002. **23**(7): p. 1139-1148.

261. Ribeiro, M.A., et al., *Integrative transcriptome and microRNome analysis identifies dysregulated pathways in human Sertoli cells exposed to TCDD.* Toxicology, 2018. **409**: p. 112-118.

262. Holgersen, K., et al., *High-Resolution Gene Expression Profiling Using RNA Sequencing in Patients With Inflammatory Bowel Disease and in Mouse Models of Colitis.* Journal of Crohn's and Colitis, 2015. **9**(6): p. 492-506.

263. Mirza, A.H., et al., *Transcriptomic landscape of lncRNAs in inflammatory bowel disease.* Genome Medicine, 2015. **7**(1): p. 39.

264. Gillberg, L., et al., *Nitric oxide pathway-related gene alterations in inflammatory bowel disease.* Scandinavian Journal of Gastroenterology, 2012. **47**(11): p. 1283-1298.

265. Cooke, J., *Mucosal genome-wide methylation changes in inflammatory bowel disease.* Inflammatory bowel diseases, 2012. **18**(11): p. 2128-2137.

266. Camilleri, M., et al., *RNA sequencing shows transcriptomic changes in rectosigmoid mucosa in patients with irritable bowel syndrome-diarrhea: a pilot case-control study.* American journal of physiology. Gastrointestinal and liver physiology, 2014. **306**(12): p. G1089-G1098.

267. Hong, S.N., et al., *RNA-seq Reveals Transcriptomic Differences in Inflamed and Noninflamed Intestinal Mucosa of Crohn's Disease Patients Compared with Normal Mucosa of Healthy Controls.* Inflammatory Bowel Diseases, 2017. **23**(7): p. 1098-1108.

268. Taman, H., et al., *Transcriptomic Landscape of Treatment—Naïve Ulcerative Colitis.* Journal of Crohn's and Colitis, 2018. **12**(3): p. 327-336.

269. Peters, L.A., et al., *A functional genomics predictive network model identifies regulators of inflammatory bowel disease.* Nature Genetics, 2017. **49**: p. 1437.

270. Mo, A., et al., *Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease.* Genome Medicine, 2018. **10**(1): p. 48.

271. Nassan, F.L., et al., *Dibutyl-phthalate exposure from mesalamine medications and serum thyroid hormones in men.* International Journal of Hygiene and Environmental Health, 2019. **222**(1): p. 101-110.

272. Uhlén, M., et al., *Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.

273. Atlas, T.H.P. *PRSS21*. 2019 February 1, 2019]; Available from: https://www.proteinatlas.org/ENSG00000007038-PRSS21/tissue.

274. Kleinschmidt, E.G. and D.D. Schlaepfer, *Focal adhesion kinase signaling in unexpected places.* Current Opinion in Cell Biology, 2017. **45**: p. 24-30.

275. Dattilo, V., et al., *SGK1 affects RAN/RANBP1/RANGAP1 via SP1 to play a critical role in pre-miRNA nuclear export: a new route of epigenomic regulation.* Scientific Reports, 2017. **7**: p. 45361.

276. Wang, D.-H., D.-D. Ma, and W.-X. Yang, *Kinesins in spermatogenesis†.* Biology of Reproduction, 2017. **96**(2): p. 267-276.

277. Zou, Y., A.O. Sperry, and C.F. Millette, *KRP3A and KRP3B: Candidate Motors in Spermatid Maturation in the Seminiferous Epithelium1.* Biology of Reproduction, 2002. **66**(3): p. 843-855.

278. Ayer-LeLievre, C., et al., *Nerve growth factor mRNA and protein in the testis and epididymis of mouse and rat.* Proceedings of the National Academy of Sciences, 1988. **85**(8): p. 2628.

279. Seidl, K. and A.F. Holstein, *Evidence for the presence of nerve growth factor (NGF) and NGF receptors in human testis.* Cell and Tissue Research, 1990. **261**(3): p. 549-54.

280. Jin, W., et al., *Cellular localization of NGF and its receptors trkA and p75LNGFR in male reproductive organs of the Japanese monkey, Macaca fuscata fuscata.* Endocrine, 2006. **29**(1): p. 155-160.

281. Li, C., et al., *Detection of nerve growth factor (NGF) and its specific receptor (TrkA) in ejaculated bovine sperm, and the effects of NGF on sperm function.* Theriogenology, 2010. **74**(9): p. 1615-1622.

282. Perrard, M.-H., et al., *Cytostatic Factor Proteins Are Present in Male Meiotic Cells and β-Nerve Growth Factor Increases Mos Levels in Rat Late Spermatocytes.* PLOS ONE, 2009. **4**(10): p. e7237.

283. Lin, K., et al., *Nerve growth factor promotes human sperm motility in vitro by increasing the movement distance and the number of A grade spermatozoa.* Andrologia, 2015. **47**(9): p. 1041-1046.

284. Shun, Z., et al., *Endogenous EGF maintains Sertoli germ cell anchoring junction integrity and is required for early recovery from acute testicular ischemia/reperfusion injury.* REPRODUCTION, 2013. **145**(2): p. 177-189.

285. Cheng, J., et al., *Regulation of Sertoli-Germ Cell Adhesion and Sperm Release by FSH and Nonclassical Testosterone Signaling.* Molecular Endocrinology, 2011. **25**(2): p. 238-252.

286. Jiang, J.-T., et al., *Prenatal exposure to di-n-butyl phthalate (DBP) differentially alters androgen cascade in undeformed versus hypospadiac male rat offspring.* Reproductive Toxicology, 2016. **61**: p. 75-81.

287. Ryu, J.Y., et al., *Identification of differentially expressed genes in the testis of Sprague-Dawley rats treated with di(n-butyl) phthalate.* Toxicology, 2007. **234**(1): p. 103-112.

288.   O'Connor, J.C., S.R. Frame, and G.S. Ladics, *Evaluation of a 15-Day Screening Assay Using Intact Male Rats for Identifying Antiandrogens.* Toxicological Sciences, 2002. **69**(1): p. 92-108.

289.   Tremblay, J.J., *Molecular regulation of steroidogenesis in endocrine Leydig cells.* Steroids, 2015. **103**: p. 3-10.

290.   Lasko, J., et al., *Calcium/calmodulin and cAMP/protein kinase-A pathways regulate sperm motility in the stallion.* Animal Reproduction Science, 2012. **132**(3): p. 169-177.

291.   Schlingmann, K., et al., *Calmodulin and CaMKII in the Sperm Principal Piece: Evidence for a Motility-Related Calcium/Calmodulin Pathway.* Journal of Andrology, 2007. **28**(5): p. 706-716.

292.   Ahmad, K., et al., *Regulation of human sperm motility and hyperactivation components by calcium, calmodulin, and protein phosphatases.* Archives of andrology, 1995. **35**(3): p. 187.

293.   Zhang, X., et al., *Inhibition of PPARα attenuates vimentin phosphorylation on Ser-83 and collapse of vimentin filaments during exposure of rat Sertoli cells in vitro to DBP.* Reproductive Toxicology, 2014. **50**: p. 11-18.

294.   Koehler, J.K., E.D. Nudelman, and S. Hakomori, *A collagen-binding protein on the surface of ejaculated rabbit spermatozoa.* The Journal of Cell Biology, 1980. **86**(2): p. 529.

295.   Li, S., et al., *Crucial role of estrogen for the mammalian female in regulating semen coagulation and liquefaction in vivo.* PLOS Genetics, 2017. **13**(4): p. e1006743.

296.   Busada, J.T. and C.B. Geyer, *The Role of Retinoic Acid (RA) in Spermatogonial Differentiation1.* Biology of Reproduction, 2016. **94**(1): p. 10, 1-10-10, 1-10.

297.   Morales, Y., et al., *Biochemistry and regulation of the protein arginine methyltransferases (PRMTs).* Archives of Biochemistry and Biophysics, 2016. **590**: p. 138-152.

298.   Alver, B.H., et al., *The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers.* Nature Communications, 2017. **8**: p. 14648.

299.   Meier, K. and A. Brehm, *Chromatin regulation: How complex does it get?* Epigenetics, 2014. **9**(11): p. 1485-1495.

300.   Hupalowska, A., et al., *CARM1 and Paraspeckles Regulate Pre-implantation Mouse Embryo Development.* Cell, 2018. **175**(7): p. 1902-1916.e13.

301.   Barrachina, F., et al., *Novel and conventional approaches for the analysis of quantitative proteomic data are complementary towards the identification of seminal plasma alterations in infertile patients.* Molecular &amp; Cellular Proteomics, 2018: p. mcp.RA118.001248.

302.   Clavería, C., et al., *Myc-driven endogenous cell competition in the early mammalian embryo.* Nature, 2013. **500**: p. 39.

303.   Kanatsu-Shinohara, M., et al., *Myc/Mycn-mediated glycolysis enhances mouse spermatogonial stem cell self-renewal.* Genes & Development, 2016. **30**(23): p. 2637-2648.

304.   Alvarez-Fernandez, M., et al., *Crystal Structure of Human Cystatin D, a Cysteine Peptidase Inhibitor with Restricted Inhibition Profile.* Journal of Biological Chemistry, 2005. **280**(18): p. 18221-18228.

305.   Ivan, M. and W.G. Kaelin, *The EGLN-HIF O2-Sensing System: Multiple Inputs and Feedbacks.* Molecular Cell, 2017. **66**(6): p. 772-779.

306.   Schuster, A., et al., *SpermBase: A Database for Sperm-Borne RNA Contents1.* Biology of Reproduction, 2016. **95**(5): p. 99, 1-12-99, 1-12.

307.   Wong, L.H., et al., *Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere.* Genome research, 2007. **17**(8): p. 1146-1160.

308.   Blower, Michael D., *Centromeric Transcription Regulates Aurora-B Localization and Activation.* Cell Reports, 2016. **15**(8): p. 1624-1633.

309.   Yaron, Y., et al., *Centromere sequences localize to the nuclear halo of human spermatozoa.*
         International Journal of Andrology, 1998. **21**(1): p. 13-8.

310.   Wang, H., et al., *Di-n-Butyl Phthalate (DBP) Exposure Induces Oxidative Damage in Testes of*
         Adult Rats AU - Zhou, Dangxia. Systems Biology in Reproductive Medicine, 2010. **56**(6): p.
         413-419.

311.   Kocer, A., et al., *Oxidative DNA damage in mouse sperm chromosomes: Size matters.* Free
         Radical Biology and Medicine, 2015. **89**: p. 993-1002.

312.   Peng, H., et al., *A novel class of tRNA-derived small RNAs extremely enriched in mature*
         mouse sperm. Cell Research, 2012. **22**: p. 1609.

313.   Krawetz, S.A., et al., *A survey of small RNAs in human sperm.* Human Reproduction, 2011.
         26(12): p. 3401-3412.

314.   Martinez, G. and C. Köhler, *Role of small RNAs in epigenetic reprogramming during plant*
         sexual reproduction. Current Opinion in Plant Biology, 2017. **36**: p. 22-28.

315.   Rosenkranz, D., *piRNA cluster database: a web resource for piRNA producing loci.* Nucleic
         Acids Research, 2015. **44**(D1): p. D223-D230.

316.   piRNAdb. *hsa-piR-27080*. piRNA Database version 1.7.5 2019 [cited 2019 February 1];
         Available from: https://www.pirnadb.org/information/pirna/hsa-piR-27080.

317.   Zamudio, N., et al., *DNA methylation restrains transposons from adopting a chromatin*
         signature permissive for meiotic recombination. Genes & Development, 2015. **29**(12): p.
         1256-1270.

318.   Mugat, B., et al., *MicroRNA-Dependent Transcriptional Silencing of Transposable Elements in*
         Drosophila Follicle Cells. PLOS Genetics, 2015. **11**(5): p. e1005194.

319.   Green, C.D., et al., *Impact of Dietary Interventions on Noncoding RNA Networks and mRNAs*
         Encoding Chromatin-Related Factors. Cell Reports, 2017. **18**(12): p. 2957-2968.

320. Hamdorf, M., et al., *miR-128 represses L1 retrotransposition by binding directly to L1 RNA.* Nature Structural &Amp; Molecular Biology, 2015. **22**: p. 824.

321. Roberts, J.T., S.E. Cardin, and G.M. Borchert, *Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences.* Mobile Genetic Elements, 2014. **4**(3): p. e29255.

322. Kojima, K.K., *Human transposable elements in Repbase: genomic footprints from fish to humans.* Mobile DNA, 2018. **9**(1): p. 2.

323. Chen, L., et al., *Co-exposure to environmental endocrine disruptors in the US population.* Environmental Science and Pollution Research, 2019.

324. Kaur, G., L.A. Thompson, and J.M. Dufour, *Sertoli cells – Immunological sentinels of spermatogenesis.* Seminars in Cell & Developmental Biology, 2014. **30**: p. 36-44.

325. Kappelman, M.D., et al., *The Prevalence and Geographic Distribution of Crohn's Disease and Ulcerative Colitis in the United States.* Clinical Gastroenterology and Hepatology, 2007. **5**(12): p. 1424-1429.

326. Naro, C., et al., *An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation.* Developmental Cell, 2017. **41**(1): p. 82-93.e4.

327. Mecklenburg, J.M. and B.P. Hermann, *Mechanisms Regulating Spermatogonial Differentiation*, in *Molecular Mechanisms of Cell Differentiation in Gonad Development*, R.P. Piprek, Editor. 2016, Springer International Publishing: Cham. p. 253-287.

328. Teletin, M., et al., *Chapter Seven - Roles of Retinoic Acid in Germ Cell Differentiation*, in *Current Topics in Developmental Biology*, D. Forrest and S. Tsai, Editors. 2017, Academic Press. p. 191-225.

329. Lalancette, C., et al., *Paternal contributions: New functional insights for spermatozoal RNA.* Journal of Cellular Biochemistry, 2008. **104**(5): p. 1570-1579.

330. Network, f.t.R.M., et al., *The presence, role and clinical use of spermatozoal RNAs.* Human Reproduction Update, 2013. **19**(6): p. 604-624.

331. Shaha, C., R. Tripathi, and D.P. Mishra, *Male germ cell apoptosis: regulation and biology.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2010. **365**(1546): p. 1501-1515.

332. Conine, C.C., et al., *Small RNAs Gained during Epididymal Transit of Sperm Are Essential for Embryonic Development in Mice.* Developmental Cell, 2018. **46**(4): p. 470-480.e3.

333. Sharma, U., et al., *Small RNAs Are Trafficked from the Epididymis to Developing Mammalian Sperm.* Developmental Cell, 2018. **46**(4): p. 481-494.e6.

334. Belleannée, C., *Extracellular microRNAs from the epididymis as potential mediators of cell-to-cell communication.* Asian Journal of Andrology, 2015. **17**(5): p. 730-736.

335. Anton, E. and S.A. Krawetz, *Spermatozoa as biomarkers for the assessment of human male infertility and genotoxicity* Systems Biology in Reproductive Medicine, 2012. **58**(1): p. 41-50.

336. Neuhaus, N., et al., *Single-cell gene expression analysis reveals diversity among human spermatogonia.* MHR: Basic science of reproductive medicine, 2017. **23**(2): p. 79-90.

337. Ramón, M., et al., *Understanding Sperm Heterogeneity: Biological and Practical Implications.* Reproduction in Domestic Animals, 2014. **49**(s4): p. 30-36.

338. Kauffman, A.S., K. Bojkowska, and E.F. Rissman, *Critical periods of susceptibility to short-term energy challenge during pregnancy: Impact on fertility and offspring development.* Physiology & Behavior, 2010. **99**(1): p. 100-108.

339. Gollwitzer, E.S. and B.J. Marsland, *Impact of Early-Life Exposures on Immune Maturation and Susceptibility to Disease.* Trends in Immunology, 2015. **36**(11): p. 684-696.

340. Sinclair, A.W., et al., *Diethylstilbestrol-induced mouse hypospadias: "window of susceptibility".* Differentiation, 2016. **91**(1): p. 1-18.

341. Haugen, A.C., et al., *Evolution of DOHaD: the impact of environmental health sciences.* Journal of Developmental Origins of Health and Disease, 2015. **6**(2): p. 55-64.

342. Edwards, R.G. and P.C. Steptoe, *Control of human ovulation, fertilization and implantation.* Proceedings of the Royal Society of Medicine, 1974. **67**(9): p. 932-936.

343. Yamada, M., et al., *Genetic Drift Can Compromise Mitochondrial Replacement by Nuclear Transfer in Human Oocytes.* Cell Stem Cell, 2016. **18**(6): p. 749-754.

344. Woods, D.C. and J.L. Tilly, *Autologous Germline Mitochondrial Energy Transfer (AUGMENT) in Human Assisted Reproduction.* Seminars in reproductive medicine, 2015. **33**(6): p. 410-421.

345. Kang, E., et al., *Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations.* Nature, 2016. **540**: p. 270.

346. Yong, E., *A Reckless and Needless Use of Gene Editing on Human Embryos*, in *The Atlantic*. 2018: https://www.theatlantic.com/science/archive/2018/11/first-gene-edited-babies-have-allegedly-been-born-in-china/576661/.

347. Argyle, C.E., J.C. Harper, and M.C. Davies, *Oocyte cryopreservation: where are we now?* Human Reproduction Update, 2016. **22**(4): p. 440-449.

348. Hendrickson, P.G., et al., *Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons.* Nature Genetics, 2017. **49**: p. 925.

349. Okae, H., et al., *Genome-Wide Analysis of DNA Methylation Dynamics during Early Human Development.* PLOS Genetics, 2014. **10**(12): p. e1004868.

350. Delihas, N., *A family of long intergenic non-coding RNA genes in human chromosomal region 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence.* PloS one, 2018. **13**(4): p. e0195702-e0195702.

351. Nanostring Technologies, I. *System Selection Guide*. 2019 [cited 2019 May 1]; Available from: https://www.nanostring.com/products/ncounter-systems-overview/ncounter-overview-system-selection-guide.

352. Atlas, T.H.P. *GP6*. 2019 [cited 2019 February 1]; Available from: https://www.proteinatlas.org/ENSG00000088053-GP6/tissue.

353. NCBI. *cytochrome P450, family 1, subfamily A, polypeptide 1 [Homo sapiens (human)]*. 2014 12/07/2014 [cited 2014 December 11]; Available from: http://www.ncbi.nlm.nih.gov/gene/1543.

354. GeneCards. *Microtubule-Actin Crosslinking Factor 1*. 2015 1/8/2015 [cited 2015 February 1]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=MACF1.

355. NCBI. *LIM homeobox 8 [Homo sapiens (human)]*. 2014 12/07/2014 [cited 2014 December 11]; Available from: http://www.ncbi.nlm.nih.gov/gene/431707.

356. NCBI. *uncharacterized LOC102503427 [Homo sapiens (human)]*. 2014 12/07/2014 [cited 2014 December 11]; Available from: http://www.ncbi.nlm.nih.gov/gene/?term=loc102503427.

357. GeneCards. *Cysteine-Rich Secretory Protein LCCL Domain Containing 2*. 2016 [cited 2016 January 3].

358. Chiquet, B.T., et al., *CRISPLD2: a novel NSCLP candidate gene.* Human Molecular Genetics, 2007. **16**(18): p. 2241-2248.

359. Atlas, T.H.P. *Chromosome 7 open reading frame 50*. [cited 2016 January 3]; Available from: http://www.proteinatlas.org/ENSG00000146540-C7orf50/tissue.

360. NCBI. *ZNF503 antisense RNA 2* 2016 11/15/2015 [cited 2016 January 3]; Available from: http://www.ncbi.nlm.nih.gov/gene/100131213.

361. GeneCards. *Growth Factor Receptor-Bound Protein 7*. 2016 [cited 2016 January 3]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=GRB7&keywords=GRB7.

362. GeneCards. *Testis Expressed 14*. 2016 [cited 2016 January 3]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=TEX14.

363. NCBI. *Nuclear factor I/X (CCAAT-binding transcription factor)*. 2016 12/6/2015 [cited 2016 January 3]; Available from: http://www.ncbi.nlm.nih.gov/gene/4784.

364. GeneCards. *Hypoxia Inducible Lipid Droplet-Associated*. 2016 [cited 2016 January 3]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=HILPDA.

365. NCBI. *Theg spermatid protein*. 2016 12/06/2015 [cited 2016 January 3]; Available from: http://www.ncbi.nlm.nih.gov/gene/51298.

366. NCBI. *Histone deacetylase 4*. 2016 12/12/2015 [cited 2016 January 3]; Available from: http://www.ncbi.nlm.nih.gov/gene/9759.

367. NCBI. *Crystallin, alpha A [Homo sapiens (human)]*. 2015 1/13/2015 [cited 2015 January 15]; Available from: http://www.ncbi.nlm.nih.gov/gene/1409.

368. NCBI. *protein tyrosine phosphatase, receptor type, N polypeptide 2 [Homo sapiens (human)]*. 2014 12/07/2014 [cited 2014 December 11]; Available from: http://www.ncbi.nlm.nih.gov/gene/5799.

369. NCBI. *Mucin 8 [Homo sapiens (human)]*. 2015 1/13/2015 [cited 2015 January 15]; Available from: http://www.ncbi.nlm.nih.gov/gene/100129528.

370. GeneCards. *Polypeptide N-Acetylgalactosaminyltransferase 9*. 2015 1/8/2015 [cited 2015 February 1]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=GALNT9.

371. NCBI. *FERM domain containing 4A [Homo sapiens (human)]*. 2015 10/13/2015 [cited 2015 January 17]; Available from: http://www.ncbi.nlm.nih.gov/gene/55691.

372.   GeneCards. *SKI/DACH Domain Containing 1*. 2015 1/8/2015 [cited 2015 February 2];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=SKIDA1.

373.   GeneCards. *Cyclin-Dependent Kinase Inhibitor 1C*. 2015 1/8/2015 [cited 2015 February 2];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=CDKN1C.

374.   GeneCards. *Solute Carrier Family 22, Member 18*. 2015 1/8/2015 [cited 2015 February 1];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=SLC22A18.

375.   NCBI. *BAI1-associated protein 2 [Homo sapiens (human)]*. 2015 10/13/2015 [cited 2015

January 10]; Available from: http://www.ncbi.nlm.nih.gov/gene/10458.

376.   GeneCards. *Guanine Nucleotide Binding Protein (G Protein), Alpha Activating Activity

Polypeptide, Olfactory Type*. 2015 1/8/2015 [cited 2015 February 1]; Available from:

http://www.genecards.org/cgi-bin/carddisp.pl?gene=GNAL.

377.   GeneCards. *Charged Multivesicular Body Protein 1B*. 2015 1/8/2015 [cited 2015 February 1];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=CHMP1B.

378.   GeneCards. *Signal Peptide Peptidase Like 2B*. 2015 1/8/2015 [cited 2015 February 2];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=SPPL2B.

379.   GeneCards. *LSM7 Homolog, U6 Small Nuclear RNA Associated (S. Cerevisiae)*. 2015 1/8/2015

[cited 2015 February 1]; Available from: http://www.genecards.org/cgi-

bin/carddisp.pl?gene=LSM7.

380.   GeneCards. *Peroxidasin Homolog (Drosophila)*. 2015 10/7/2014 [cited 2015 January 15];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=PXDN.

381.   NCBI. *Peroxidasin [Homo sapiens (human)]*. 2015 1/13/2015 [cited 2015 January 15];

Available from: http://www.ncbi.nlm.nih.gov/gene/7837.

382.   GeneCards. *G Protein-Coupled Receptor 35*. 2015 10/7/2014 [cited 2015 January 15];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=GPR35.

383. GeneCards. *Aquaporin 12B*. 2015 10/7/2015 [cited 2015 January 15]; Available from:

http://www.genecards.org/cgi-bin/carddisp.pl?gene=AQP12B.

384. GeneCards. *Regulator Of G-Protein Signaling 12*. 2015 10/7/2014 [cited 2015 January 15];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=RGS12.

385. GeneCards. *Epiplakin 1*. 2015 1/8/2015 [cited 2015 February 1]; Available from:

http://www.genecards.org/cgi-bin/carddisp.pl?gene=EPPK1.

386. NCBI. *neuron navigator 1* 2016 12/06/2015 [cited 2016 January 3]; Available from:

http://www.ncbi.nlm.nih.gov/gene/89796.

387. GeneCards. *Tubulin Polymerization Promoting Protein*. 2016 [cited 2016 January 3];

Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=TPPP.

ABSTRACT

DEFINING THE EFFECT OF ENVIRONMENTAL PERTURBATION ON THE MALE GERMLINE

by

MOLLY S. ESTILL

August 2019

Advisor: Dr. Stephen A. Krawetz

Major: Molecular Medicine and Genetics

Degree: Doctor of Philosophy

Periconceptional environment, according to the Developmental Origins of Health and Disease (DOHaD) theory, influences offspring phenotype, primarily via epigenetic mechanisms. Although the paternal component in humans is poorly understood, both maternal and paternal peri-conceptional environment are now believed to contribute to this phenomenon. Manipulation of the early embryo for treating human infertility, is suspected of contributing to offspring abnormalities through epigenetic mechanisms. To directly address the effects of common assisted reproductive technology procedures on the offspring epigenome, the DNA methylation profiles of newborns conceived naturally, or through the use of intrauterine insemination (IUI), or *in vitro* fertilization (IVF) using Fresh or Cryopreserved (Frozen) embryo transfer, were compared. In addition to a reduction of epigenetic aberrations in the IVF conceptions using cryopreservation, metastable epialleles also exhibited altered methylation with fertility status. ART, embryo nutrition, and fertility status are thus suggested to have a lasting epigenetic effect of on the developing embryo. While the paternal contribution to the human embryo is uncertain, sperm deliver a collection of proteins and RNA to the zygote. To identify the entire cadre of intergenic spermatozoal RNAs, RNA Element (RE) discovery algorithm (REDa) was developed and applied to a spectrum of germline, embryonic, and somatic tissues. This highlighted extensive transcription throughout the human genome

and yielded previously unidentified human RNAs. Human spermatogenesis was found to exhibit extensive intergenic transcription and pervasive repetitive sequence expression. By analyzing the collection of novel and annotated spermatozoal RNAs in sperm samples from the Mesalamine and Reproductive Health Study (MARS), the effect of endocrine disruptor exposure on human sperm RNA profiles was determined. Sperm RNA profiles among men and their relationship to di-butyl phthalate (DBP) was longitudinally assessed across binary (high or background) DBP crossover exposures. Numerous changes in the composition of sperm RNA elements were detected during the acute and recovery phases, which suggest that exposure to, or removal from high DBP, produces effects that require longer than one spermatogenic cycle to resolve, if at all. Overall, chronic phthalate exposure influences the male germline, and acts on the dynamic RNA expression during human spermiogenesis.

AUTOBIOGRAPHICAL STATEMENT

EDUCATION

Aug. 2013 – May 2019   Center for Molecular Medicine and Genetics
Wayne State University School of Medicine, Detroit, MI, USA
Ph.D., Molecular Biology and Genetics

Aug. 2010 – Feb. 2013   Department of Biology
Villanova University, Villanova, PA, USA
M.S., Biology

Aug. 2005 – May 2009   Department of Biology
Albion College, Albion, MI
B. A., French and Biology

SELECTED ACHIEVEMENTS

2018   First place Graduate Poster Presentation, Wayne State University Presidential Sesquicentennial Symposium

2016   Thomas C. Rumble University Graduate Fellowship, Wayne State University School of Medicine

2011   Graduate Assistantship, Villanova University

2005-2009   Trustee Scholarship, Albion College

SELECTED PUBLICATIONS

Estill, M. S., Hauser, R., and Krawetz, S. A. (2018) RNA element discovery from germ cell to blastocyst. Nucleic Acids Research. Volume 47, Issue 5, 18 March 2019, Pages 2263–2275. doi: 10.1093/nar/gky1223

Nassan, F. L., Korevaar, T., Coull, B., Skakkebæk, N., Krawetz, S. A., Estill, M., … Hauser, R. (2018) Dibutyl-Phthalate Exposure from Mesalamine Medications and Serum Thyroid Hormones in Men. Int J Hyg Environ Health. 2019 Jan;222(1):101-110. doi: 10.1016/j.ijheh.2018.08.008.

Wu, H., Estill, M. S., Shershebnev, A., Suvorov, A., Krawetz, S. A., Whitcomb, B. W., . . . Pilsner, J. R. (2017). Preconception urinary phthalate concentrations and sperm DNA methylation profiles among men undergoing IVF treatment: a cross-sectional study. Human Reproduction, 32(11), 2159-2169. doi:10.1093/humrep/dex283

Estill, M. S., & Krawetz, S. A. (2016). The Epigenetic Consequences of Paternal Exposure to Environmental Contaminants and Reproductive Toxicants. Current Environmental Health Reports, 3(3), 202-213. doi:10.1007/s40572-016-0101-4

Estill, M. S., Bolnick, J. M., Waterland, R. A., Bolnick, A. D., Diamond, M. P., & Krawetz, S. A. (2016). Assisted reproductive technology alters deoxyribonucleic acid methylation profiles in bloodspots of newborn infants. Fertility and Sterility, 106(3), 629-639.e610. doi:https://doi.org/10.1016/j.fertnstert.2016.05.006

Estill, M., Kerwin-Iosue, C. L., & Wykoff, D. D. (2015). Dissection of the PHO pathway in Schizosaccharomyces pombe using epistasis and the alternate repressor adenine. Current Genetics, 61(2), 175-183. doi:10.1007/s00294-014-0466-6