GENOME-WIDE EPIGENETIC AND GENETIC INVESTIGATION OF MALE INFERTILITY

by

KENNY LOUIE

B.Sc., The University of British Columbia, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Reproductive and Developmental Sciences)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2018

Abstract

Genetic causes have been known to be involved in up to one-third of male infertility cases. Studies on the sperm and testis from infertile men have suggested that this population may also have higher rates of altered epigenetic modifications, particularly DNA methylation. Thus, we hypothesized that genetic and epigenetic causes may play a role together in male infertility.

We used Sanger sequencing to observe the DNA methylation at imprinted differentially methylated regions– *H19*, *IG-GTL2*, and *MEST* – in the sperm of infertile men with oligozoospermia to assess the occurrence of alterations. We analyzed semen samples from fifty-three men (9 fertile and 44 infertile – stratified by sperm concentration). Although we observed altered cases in 13% (3/23) of men with severe oligozoospermia and none among controls, this finding was not significant. We genotyped the same men at the *MTHFR* C677T single nucleotide polymorphism (SNP) to determine whether this methyl supply gene is associated with DNA methylation in the sperm. We observed a trend that all altered cases had the CT genotype at this SNP and in the severe oligozoospermic subgroup, suggesting a combinatorial effect. Motivated from these findings, we conducted two genome-wide investigations to identify novel genes with DNA methylation alterations and/or SNPs relating to male infertility. We evaluated the genome-wide DNA methylation in the testis of twenty-four men (8 fertile and 16 infertile). Using predictive models, we identified 359 CpGs with altered DNA methylation among the infertile men. Of these loci and using functional analyses, we identified *NDE1* which is a gene involved with cell cycle progression and centrosome formation. In the genome-wide SNP analysis of twenty men (13 fertile and 7 infertile), we identified 39 SNPs using machine learning models. Of these SNPs, candidates included *MTR* – a gene involved with the same folate pathway as

MTHFR – and *NIN* which is a gene involved with proper chromosome segregation during cell division.

In summary, we observed altered DNA methylation and SNPs in infertile men. We presented evidence that altered DNA methylation and changes in genetic sequence in genes involved with cell division and progression may impair spermatogenesis leading to male infertility.

Lay Summary

Infertility affects one in six couples worldwide, where either partner is equally likely to be a contributor to the condition. Although much research has focused on female infertility, the causes in up to half of male infertility cases are unknown. The preliminary investigations in the study of the regulation of DNA (termed epigenetics), as opposed to the DNA itself (termed genetics), has proved fruitful and may be a venture point for future studies.

We show that combined epigenetic and genetic changes may be a cause of male infertility. We surveyed the genome for new genes that may be involved with male infertility and determined a handful of new genes that when altered in epigenetic and/or genetic code, may contribute to the delayed progression or inhibition of sperm production. We believe that these genes may be key factors involved with the absence of sperm in infertile men.

Preface

This thesis is based on one of the many ongoing research programs in Dr. Sai Ma's laboratory. All work in this thesis received approval by the UBC's Research Ethics Board.

In Chapter 2, the research design was adapted from previous studies conducted by Agata Minor and Richard Ng – graduate students in Dr. Sai Ma's laboratory who also researched on male infertility. The addition of *MTHFR* genotyping was a design choice made in discussion between myself and Dr. Sai Ma. The recruitment of infertile cases was conducted by collaborating urologists Dr. Victor Chow and Dr. Kenneth Poon. The processing and cloning of samples was conducted in part by Richard Ng, Agata Minor, and myself. I led the data analysis with Richard Ng. A version of Chapter 2 has been published. Louie, K., Minor, A., Ng, R., Poon, K., Chow, V., & Ma, S. (2016) Evaluation of DNA methylation at imprinted DMRs in the spermatozoa of oligozoospermic men in association with *MTHFR* C677T genotype. Andrology, 4(5), 825-831. I wrote most of the manuscript and worked with the publishing company for revisions.

In Chapter 3, the experiments were designed by myself in discussion with Dr. Sai Ma. The recruitment of infertile cases was conducted by collaborating urologists Dr. Victor Chow and Dr. Kenneth Poon. I processed all the samples and sent them to the microarray facilities. I designed the analytical data mining pipeline.

In Chapter 4, the experiments were designed by Dr. Sai Ma. The recruitment of infertile cases was conducted by collaborating urologists Dr. Victor Chow. The processing of samples was conducted by Edgar Chan Wong – a previous graduate student in Dr. Sai Ma's laboratory. I designed the analytical data mining pipeline.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

450k	Illumina Infinium Human Methylation450 BeadChip
5-caC	5-carboxylcytosine
5-fC	5-formylcytosine
5-hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
5-mTHF	5-methylTHF
aCGH	Array-comparative genomic hybridization
AIS	Androgen insensitivity syndrome
AMH	Anti-Mullerian hormone
AR	Androgen receptor
ART	Assisted reproductive technologies
AS	Angelman syndrome
AZF	Azoospermia factor
BER	Base excision repair
BIRF	Balanced iterative random forest
BMP	Bone morphogenic protein
bp	Base pair
BTB	Blood-testis barrier
BWS	Beckwith Wiedemann syndrome
CAIS	Complete androgen insensitivity
CBAVD	congenital bilateral absence of the vas deferens
CER1	Cerberus 1

CFTR	Cystic fibrosis transmembrane regulator
CN	Copy number
CNV	Copy number variation
CpG	Cytosine-phosphate-guanine
CTCF	CCTC-binding factor
DBY	DEAD-box helicase 3
DDK1	Dickkopf 1
DDR1	Discoidin domain receptor 1
DHF	Dihydrofolate
DMP	Differentially methylated positions
DMR	Differentially methylated regions
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferases
DSB	Double strand breaks
DTT	Dithiothreitol
eBayes	Empirical Bayesian
EDTA	Ethylenediaminetetraacetic acid
ES	Ectoplasmic specialization
FDR	False discovery rate
FGF9	Fibroblast growth factor 9
FPP	Fertilization promoting peptide
FSH	Follicle-stimulating hormone
FGF9	Fibroblast growth factor 9

GEO	Gene Expression Omnibus
GO	Gene Ontology
GnRH	Gonadotropin-releasing hormone
GVRD	Greater Vancouver regional district
GWAS	Genome-wide association study
H3K9	Histone 3 Lysine 9
HDAC	Histone deactelyases
HPG	Hypothalamic-pituitary-gonadal
HMM	Hidden Markov Model
ICMART	International Committee Monitoring Assisted Reproductive Technologies
ICR	Imprinting control region
IQR	Interquartile range
KAP1	KRAB-associated protein 1
kb	Kilobase pair
KCNQ1OT1	KCNQ1 opposite transcript 1
KRAB	Krueppel-associated box
LH	Luteinizing hormone
LOOCV	Leave-one-out-cross-validation
MAIS	Mild forms of AIS
MBD4	Methylated DNA-binding protein
MeCP2	Methyl-CpG binding protein 2
miRNA	Micro RNA
ML	Machine learning

mRNA	Messenger RNA
MSY	Male-specific region on the Y chromosome
MTHFR	Methylenetetrahydrofolate reductase
ncRNA	Non-coding RNA
NGS	Next-generation sequencing
NIN	Ninein
NOA	Non-obstructive azoospermia
OA	Obstructive azoospermia
OAT	Oligoasthenoteratozoospermia
PAIS	Partial forms of AIS
PAR	Pseudoautosomal region
piRNA	Piwi-interacting RNA
PC	Principal component
PCA	Principal component analysis
PFI	Permutation feature importance
PGC	Primordial germ cells
PIWI	P-element-induced wimpy
PRDM1	PR domain zinc-finger protein 1
RBMY	RNA binding motif protein
REC8	Recombination protein 8
RF	Random forest
RHOX	Reproductive homeobox
RISC	RNA-induced silencing complex

RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SAM	S-adenosyl-L-Methionine
SC	Synaptonemal complex
SCOS	Sertoli cell only syndrome
SF1	Steroidogenic factor 1
SGO1	Shugoshin 1
siRNA	Small interfering RNA
SNP	Single nucleotide polymorphisms
SNP 6.0	Affymetrix Genome-Wide Human SNP array 6.0
SNRPN	Small nuclear ribonucleoprotein polypeptide N
SOX9	SRY-box 9
SOX17	SRY-box 17
SRY	Sex-determining region Y
StAR	Steroidogenic acute regulatory protein
SVD	Singular value decomposition
SVM	Support vector machines
SVM-RFE	Support vector machine recursive feature elimination
SWAN	Subset-quantile within array normalization
TAD	Topologically associated chromatin domains
TBC	Tubulobulbar complexes
TET	Ten-eleven translocation
TEX11	X-linked testis-expressed 11

THF	Tetrahydrofolate
TP1	Transition protein 1
TP2	Transition protein 2
tRNA	Transfer RNA
VR	Vasectomy reversal
WHO	World Health Organization
WNT3	Wingless-type MMTV integration site family member 3
ZFP57	Zinc finger protein 57 homolog

Acknowledgements

I would like to thank my supervisor Dr. Sai Ma for giving me the opportunity to research in her laboratory, providing me with the flexibility and patience to explore my curiosity, and being a wealth of experience in research and in life. I am also grateful to the members of my supervisory committee, Dr. Patrice Eydoux, Dr. Wan Lam, and Dr. Rajavel Elango, for their patience with my progress, and insights to the projects. I am indebted to the members of Dr. Sai Ma's laboratory who have directly and indirectly contributed to the projects presented in this thesis. I would like to thank Richard Ng, Luke Gooding, Samuel Schafer, Kate Watt, Annie Ren, Rowena Ho, Elizabeth Wu, Kevin Dong, Rebecca Vincent, Paloma Stanar for their help, guidance, and friendship.

I am indebted to Dr. Victor Chow and Dr. Kenneth Poon from the department of Urology, who performed the testicular biopsies for the research. Without them, the research would not be possible. I am grateful to the community at the BC Children's Hospital Research Institute (previously Child and Family Research Institute), who has nurtured an environment conducive to collaboration, mentorship, and excellence. I would like to especially thank Magda Price for her mentorship in bioinformatics. I am extremely grateful to the Canadian Institutes of Health Research (CIHR), who financially supported the studies presented in this thesis.

Finally, and most importantly, I would like to thank my family, especially my mother, and Andy for their patience and support through my graduate studies. I would also like to thank the loved ones by my side, especially Anna, for their continued encouragement and motivation. For anyone I may have left out, I truly apologize.

CHAPTER 1: INTRODUCTION

Infertility is defined as a disease of the reproductive system characterized by the inability to achieve a clinical pregnancy after twelve months or more of regular unprotected sexual intercourse (WHO-ICMART). The estimated global prevalence of infertility is 1.9%, or 48.5 million couples (Mascarenhas et al., 2012). Infertility due to the male partner is estimated to range from 20-70% of couples or >30 million men worldwide (Agarwal et al., 2015). Of these men, an estimated 37-58% of male infertility cases are idiopathic, i.e. the cause of the infertility could not be diagnosed (de la Calle et al., 2001; Moghissi and Wallach, 1983; Irvine, 1998; Hamada et al., 2012).

The genetics of male infertility is an active field of research due to the fact that approximately one-third of known male infertility cases have some genetic component to their condition. However, investigations into the epigenetics of infertile men have provided new insight into potential causes of male factor infertility. Epigenetic modifications of the genome are chemical alterations that work with – and not alter – the DNA sequence to modulate the expression of the associated genes. From these studies, we now better understand the role of epigenetic modifications on the developing germ cell. Epigenetic studies on the sperm of infertile men have suggested that they harbor different signatures as compared to fertile men. The focus of this thesis is the discovery of novel genes and/or pathways that may be epigenetically associated with male infertility, and to investigate the relationship between epigenetics and genetics in this context.

We will begin with an introduction into epigenetics, followed by the biology of male sex differentiation. The focus of the introduction will be to describe male sex development in the context of epigenetics and genetics. Following which, the latter half will include the increasing evidence of epigenetics in association with male infertility. Due to the advancement of microarray technologies for the life sciences and their abilities to investigate in a genome-wide manner, the last section of the introduction will be a brief overview of the microarray technologies used in this thesis and the challenges in deriving biological insight from microarrays.

1.1 Epigenetics

Epigenetics is the study of cellular mechanisms which change DNA, but without altering the DNA sequence itself, that results in stable, reversible, and heritable phenotypes in organisms. Epigenetic modifications provide an explanation for the differentiation of cells despite having the same genetic sequence, or genome (Jones et al., 1980; Hemberger et al., 2009). Thus, epigenetics has been largely implicated in development (Meissner et al., 2008), and shown to be the connection between the genome and the environment (Kelly et al., 2003). The following subsections describe the most commonly studied epigenetic modifications.

1.1.1 DNA methylation

DNA methylation is the most studied epigenetic modification. It involves the chemical modification of cytosine base pairs with a methyl group at the 5’ carbon of the pyrimidine ring, resulting in a 5-methylcytosine (5mC). DNA methylation reactions occur on cytosine base pairs (bp) following specific sequences, including CG, CHG, or CHH (where H refers to an A, T, or C bp). However, in mammals including humans, DNA methylation is almost exclusively found on CG (also termed CpG) sequences. Given the extensive research of DNA methylation of CpG sites, especially in the context of male infertility (section 1.3), it will be the focus of this thesis and herein referred to simply as **methylation** in the remainder of the thesis.

Methylation reactions are catalyzed by a family of enzymes known as DNA methyltransferases (DNMTs). This family of enzymes is essential for the establishment and maintenance of methylation. These enzymes catalyze the addition of methyl groups to bps derived from methyl donors such as S-adenosyl-L-Methionine (SAM). DNMT1 is shown to have a high affinity for hemi-methylated DNA and involved with the faithful copying of methylation on newly synthesized strands after DNA replication (Leonhardt et al., 1992). DNMT1 is therefore primarily involved with maintenance of methylation marks (Li et al., 1992). In fact, the mutation of DNMT1 (having only trace levels of activity) resulted in a one-third level of methylation in homozygous mutant cells. DNMT3A and DNMT3B are suggested to confer methylation to new sites, i.e. *de novo* methylation (Okano et al., 1999). DNMT3A is suggested to methylate paternally imprinted genes (Kaneda et al., 2004), while DNMT3B at centromeric repeats. DNMT3L shares a similar molecular structure to DNMT3A and DNMT3B, but has no active enzymatic activity on its own. DNMT3L acts as a co-factor to DNMT3A via increasing enzymatic activity (Jia et al., 2007).

Most CpGs are located within long stretches commonly found upstream of genes such as within promoters or within the first exons of genes (Bird et al., 1986). These long stretches are termed CpG islands. Although 70-80% of CpGs are methylated in the human genome (Jabbari and Bernardi, 2004), most CpG islands are unmethylated except those at imprinted genes (section 1.2.6) and at inactivated X chromosome regions. Methylation at CpG islands is associated with transcriptional silencing (Merlo et al., 1995). This is most evident in human cancer research, where there exist hundreds of reports of methylation induced silencing at tumour suppressor genes (Sakamoto et al., 2007). The bulky nature of methyl groups is thought to exclude transcription machinery at gene promoters and therefore prevent transcription and thus expression (Messerschmidt et al., 2014). Methylation is also implicated in the recruitment of protein complexes, e.g. methyl-CpG binding protein 2 (MeCP2), capable of regulating gene expression (Jones et al., 1998; Fuks, 2003; Ito, 2013). MeCP2 has been shown to block access of basal transcriptional machinery (Kaludov et al., 2000).

1.1.2 Non-coding RNAs

A non-coding RNA (ncRNA) is a ribonucleic acid (RNA) molecule that remains and functions as an RNA molecule. The most abundant types of ncRNAs include transfer RNAs (tRNA), ribosomal RNAs (rRNA), long ncRNAs, and small ncRNAs. The latter category is further split into micro RNAs (miRNA), small interfering RNAs (siRNA), and piwi-interacting RNAs (piRNA).

The functions of miRNAs include gene regulation and silencing of messenger RNAs (mRNA). “miRNAs” are first transcribed in the nucleus from DNA into pri-miRNAs which are then processed by the type III RNase Drosha forming pre-miRNAs (Zeng et al., 2005; Carthew et al., 2009). The export of pre-miRNAs into the cytoplasm triggers the type III RNase Dicer to further process the RNA, prior to one of the strands incorporating into the RNA-induced silencing complex (RISC). The combined RISC complex binds to mRNAs that are complementary to the miRNA's sequence, which either induces degradation or translational repression of the mRNA (depending on the level of complementarity, i.e. the more bases that are compliment the higher the rate of binding and thus more of the mRNAs would be degraded). "siRNAs" undergo similar processing as miRNAs (Carthew et al., 2009). However, piRNAs do not undergo Dicer processing and instead interact with the P-element-induced wimpy (PIWI) proteins to regulate genes (Bamezai et al., 2012). The localization of ncRNAs in the cytoplasm provides a heritable mechanism to retain gene expression regulation in mitotic and meiotic daughter cells (Yan, 2014).

1.1.3 Histone modifications

The higher order of DNA organization is the packaging of DNA around histones into nucleosomes – together forming chromatin (Cutter et al., 2015). Histones are proteins that reside within the nucleus of eukaryotic cells. The formation of an octamer of histone proteins connected by the H1 histone linker wraps 146 bp into nucleosome units providing approximately a 7-fold linear compaction. This allows the DNA to fit within the nucleus and provides protection from trans-acting factors. Thus, a lattice system is created that can regulate the exposure of DNA and thus transcription. A relaxed chromatin structure in principle is more likely to have gene expression activity, while a tightly packed structure excludes transcription machinery thereby silencing the packed genes. The negatively charged backbone of DNA winds around histone octamers due to electrostatic interactions. The modification to a specific histone subunit's N-terminus "tails" modulates the electronegativity of the protein, thereby altering how tightly DNA can bind (Kouzarides et al., 2007). This provides a controlled molecular modular system in controlling the availability of DNA to transcription machinery, thus regulating the expression of genes. For example, as mentioned earlier the binding of meCP2 (regulated by methylation on its binding site) to DNA recruits histone deactelyases (HDACs), which initiates the compaction of DNA and therefore transcriptional silencing (Jones et al., 1998). Acetylation of lysine 9 of the H3 protein (H3K9) of the histone octamer neutralizes the histone charge resulting in the loosening of DNA-histone interaction and therefore activating the DNA sequence (Struhl, 1998). The most common histone tail modifications include acetylation, phosphorylation, and methylation. Histone modifications can also directly interact with chromatin binding proteins. Histone H3 interacts with DNMT3L which in turn recruits the *de novo* methyltransferases to initiate methylation at the associated sequence; however, the methylation of H3K4 inhibits this interaction (Ooi et al., 2007).

1.2 Male development

The primary outcomes of male differentiation are the development of the testis, the ability for proper spermatogenesis, and male anatomical characteristics. The following section includes a brief overview of our current understanding of male sex development with a focus on the genetics and epigenetics of the biological processes.

1.2.1 Zygotic and embryonic development

Once the sperm enters the oocyte, the oocyte completes its second round of meiosis forming a haploid daughter cell containing the majority of the cytoplasm, the maternal pronucleus, and the paternal pronucleus from the sperm. The zygote undergoes a round of DNA replication, after which the pronuclei fuse into a single genome and immediately undergo yet another round of mitotic division, forming two blastomeres – and thus beginning the development into a multicellular organism.

The preimplantation embryo travels down the oviduct towards the uterus while continually to undergo mitotic divisions without increasing in mass, a process termed cleavage. After four divisions, the cleavage stage embryo consists of sixteen blastomeres, known as a morula. After further cell division, compaction, and blastulation – the forming of a blastocoel or central space within the cell – the conceptus takes the form of a blastocyst on the fifth day post fertilization. At this stage, the blastocyst approaches the site of implantation in the uterus. It hatches from the zona pellucida and has the potential to implant into the endometrial lining, where it will continue to develop into the fetus.

1.2.2 Zygotic epigenetic reprogramming

At the same time as the zygote is undergoing cleavage formation into a blastocyst, the pronuclei undergo a genome-wide erasure of methylation also known as epigenetic reprogramming (Messerschmidt et al., 2014). An exception to this reprogramming is imprinted genes (see section 1.2.5), where they escape from demethylation mechanisms. Demethylation is thought to be completed by the sixteen-cell morula stage (Kafri et al., 1992). Given the importance of methylation in the differentiation of cells, this genome-wide demethylation is thought to erase the parents' epigenetic signatures, transforming the embryo into a totipotent cell, capable of differentiating into a whole organism. The mechanisms of demethylation are not fully understood, however, there are proposed active and passive methods (Hill et al., 2014). Passive demethylation is thought to be due to replication dilution of methylation marks, in which methylation marks are not faithfully copied onto new daughter strands during DNA replication. Therefore, successive mitotic cycles in theory would continuously halve the total amount of methylation. Active demethylation is proposed to be mediated by enzymes. One such mechanism is via the base excision repair (BER) pathways. BER enzymes repair base pairs which have been altered. 5mC sites can be deaminated via 5mC deaminases (e.g. activation-induced deaminase or apolipoprotein B RNA-editing catalytic component) to produce thymine, resulting in a TG mismatch on the opposite strand. Glycosylases (e.g. methylated DNA-binding protein or MBD4) recognize this TG mismatch and cleaves the N-glycosidic bond between the ribose ring of the adjacent base pair and thymine, resulting in an abasic site that is then replaced with a new cytosine carrying no methylation. A newly proposed mechanism that has gained traction is the active demethylation via the ten-eleven translocation (TET) family of enzymes. TET enzymes are Fe(II)/α-ketoglutarate-dependent dioxygenases and include TET1, TET2, and TET3. These enzymes successively oxidize 5mC into 5-hydroxymethylcytosine (5hmC), then 5-formylcytosine (5fC), and finally to 5-carboxylcytosine (5caC), respectively. It is thought that 5fC and 5caC are recognized as a TG mismatch and thus excised to an abasic site and subsequently removed via the BER pathway (Hill et al., 2014). Furthermore, the active oxidation of 5mC into 5hmC via TET1 also interferes with DNMT1 in faithfully methylating daughter strands during DNA replication, thus also conferring a passive mechanism of demethylation.

1.2.3 Male differentiation

At six to eight weeks post fertilization, the developing urogenital ridge remains bipotent, i.e. capable of committing and differentiating into either the testis or ovaries (Shima and Morohashi, 2017). The master regulator of sex in humans is the sex-determining region Y (*SRY*) gene on the Y chromosome. SRY is a protein which when bound with the steroidogenic factor 1 (SF1) protein, forms a transcription factor which upregulates SRY-box 9 (*SOX9*) (Hanley et al., 2000; Kashimada et al., 2010). The result is a signaling cascade upregulating fibroblast growth factor 9 (FGF9) which in turn upregulates SOX9. Once SOX9 levels break a threshold within a specific window of time, the bipotential cells within the primordial gonads are now induced into the Leydig cells and the Sertoli cells. The development of the primary sex cords later become the seminiferous tubules and the testis. Therefore, the presence of the Y chromosome in the embryo inherited from the parents is the determining factor to inhibit the female anatomical structure development, and instead develop the testis and male germline; without the *SRY* gene the genital ridge differentiates into the ovaries.

The Sertoli cells and Leydig cells at this point will begin to secrete anti-Mullerian hormone (AMH) and testosterone, respectively. AMH causes the regression of the Mullerian ducts, which if left developing in the absence of AMH would form the female reproductive tract, including the fallopian tubes, uterus, uterine crevice, and the superior aspect of the vagina. Testosterone, on the other hand, stabilizes the Woffian ducts, which will continue to develop into the vas deferens, epididymis, and seminal vesicle (Hughes, 2001).

1.2.4 Primordial germ cells

The production of sperm begins in the germline, which itself differentiates early in embryonic development. The germline originates from pluripotent pre-implantation epiblast cells of the blastocyst (Tang et al., 2016). Following implantation, the blastocyst undergoes gastrulation forming the trilaminar embryonic disc. The exact mechanism of induction has not been fully uncovered but the following has been speculated (Tang et al., 2016). On embryonic day 17, the primordial germ cells (PGC) are specified in the forming mesoderm layer in the trilaminar human embryo. The bone morphogenic protein (BMP) signaling pathway including BMP4 and BMP8 are expressed as a ring-like domain that surrounds the epiblast. The posterior epiblast and hypoblast are speculated to be the site of wingless-type MMTV integration site family member 3 (WNT3) and BMP2 signaling together induces a few cells from the epiblast to express SRY-box 17 (SOX17). The expression of cerberus 1 (CER1) and dickkopf 1 (DDK1) occurs in the anterior hypoblast to restrict induction to the posterior epiblast, as CER1 and DDK1 are antagonists to WNT3 and BMP2. PR domain zinc-finger protein 1 (PRDM1) is upregulated in response to SOX17 expression, which the expression together is critical for cells to be specified into the germline lineage (Vincent et al., 2005). Around four weeks post fertilization, PGCs are localized in the yolk sac wall near the allantois (Tang et al., 2016). PGCs migrate through the hindgut to the forming genital ridges by embryonic day thirty-seven. By the sixth week, PGCs colonize the incipient genital ridge. It is the expression of SRY from pre-Sertoli cells that begins the differentiation of the male gonads (Albrecht et al., 2001). Gonadal PGCs continue to proliferate until the tenth week of development when they commit to gonocytes. Male gonocytes become mitotically quiescent until puberty, while female gonocytes asynchronously enter and arrest at meiotic prophase I.

1.2.5 Genomic Imprinting

One of the first identified phenomenon of epigenetics and methylation was genomic imprinting and imprinted genes. While most (somatic) genes are expressed from both the maternal and paternal allele, imprinted genes are regulated to have only expression from a single parental copy, i.e. a single dosage of expression. In humans there are approximately 100 known imprinted genes (Jirtle, 2017), with many more predicted and undergoing active research (Leudi et al., 2007). The mono-allelic expression dosage of these imprinted genes inherited by the embryo from the sperm and oocyte is thought to play a role in embryo survival, placental growth, and adaptivity in the fetus (Barton et al., 1991; Sandovici et al., 2012). Paternally expressed imprinted genes often promote fetal growth while maternally expressed genes restrict or control growth. The mutation of DNMT1 in the germline, resulting in altered methylation at genes, produced embryos that were stunted, delayed in development, and did not survive past midgestation (Li et al., 1992). The fine balance of the expression of imprinted genes is critical for proper development (Fowden et al., 2006).

Given their role in early development, the differential methylation patterns of imprints are thought to escape epigenetic reprogramming that occurs immediately after fertilization (Messerschmidt et al., 2014; Olek and Walter, 1997; Tremblay et al., 1997). The mechanisms are thought to involve the active maintenance of imprints via DNMT1 and zinc finger protein 57 homolog (ZFP57) (Mackay et al., 2006; 2008). The production of gametes with the correct imprints is therefore necessary for offspring health. The mono-allelic parent-specific gene expression is regulated by epigenetic mechanisms including methylation at regions known as differentially methylated regions (DMR), or imprinting control region (ICR) if it regulates more than one gene (Reik et al., 2007). Imprinted gene are often found in clusters where the methylation of an ICR affects the expression of multiple adjacent genes (Figure 1). Imprinted DMRs that are methylated in the spermatozoon (termed paternally methylated), are not methylated in the oocyte, and vice versa (the opposite is termed maternally methylated). Not all imprinted genes carry methylation marks in the sperm. Those that do include *H19* and *GTL2* (Kerjean, 2000; Guens et al., 2007). A well-studied insulator model highlighting the imprinting via methylation is the *H19*/*IGF2* locus (Nordin et al., 2014; Barlow and Bartolomei, 2014) (Figure 1). In this locus, the intergenic ICR on the paternal allele is normally methylated, which prevents the binding of CCTC-binding factor (CTCF) to the ICR. Transcription enhancers cis to the *H19* promoter due to the absence of CTCF skip over the *H19* gene and instead activate the expression of *IGF2*. In contrast, the maternal allele is normally devoid of methylation at the intergenic ICR. This allows CTCF to bind and physically prevent the enhancers from reaching the *IGF2* promoter. Instead, the enhancers activate the expression of the *H19* gene. Thus, the methylation imprints from the sperm and oocyte result in the mono-allelic parent-specific expression of both *H19* and *IGF2* (Barlow and Bartolomei, 2014). Only a single dosage of each gene will be present in a healthy fetus during development.

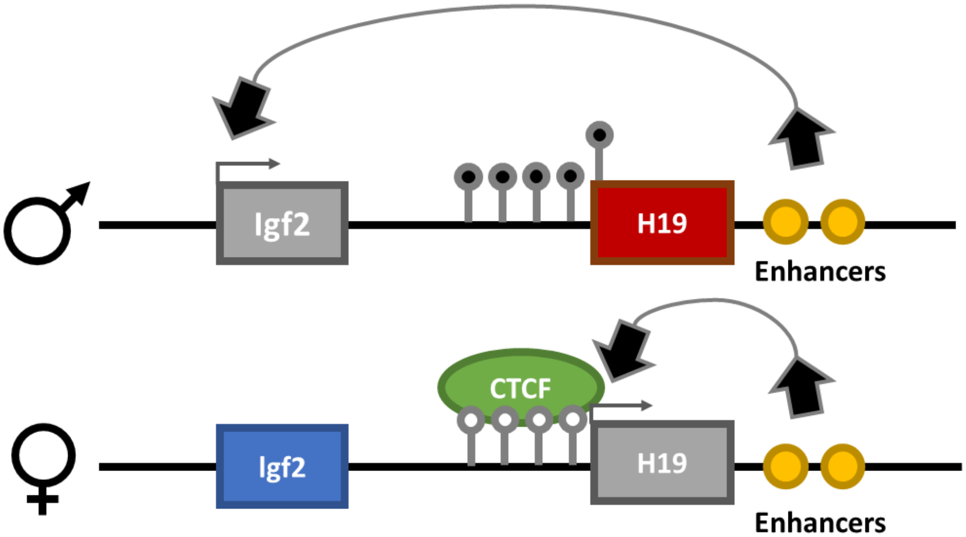


Figure 1. Genomic imprinting insulator model at the *H19*/*IGF2* locus
The mono-allelic parent-specific expression of *H19* and *IGF2*. DNA methylation illustrated as black beads at the intergenic imprinting control region between the two genes prevents binding of CTCF on the paternal allele. This allows the transcription enhancers cis to *H19* to skip over and instead upregulate the expression of *IGF2*. The maternal allele is devoid of DNA methylation at the ICR thereby allowing CTCF to bind preventing the enhancers from reaching IGF2 resulting in the expression of *H19*. Figure adapted from Nordin et al., 2014

1.2.6 PGC epigenetic reprogramming

PGCs contain high levels of methylation prior to entering the primordial gonads (Hajkova et al., 2002). However, before week four of human development and arrival at the gonadal ridge, PGCs show low global methylation levels. This is achieved by chromatin reorganization and comprehensive demethylation, including at most imprints (except *IGF2R* and *PEG10* ICRs), transposable elements, X inactivated regions, and promoters of methylation-sensitive germline genes (Gkountela et al., 2015; Guo et al., 2015; Tang et al., 2015). At this same time, 5hmC levels are detectable with gradual depletion by week nine. The *SOX17* and *PRDM1* gene regulatory network is thought to initiate and maintain this epigenetic program. Along with other factors, this gene regulatory network sustains robust repression of methylation pathways (DNMT3A and DNMT3B) and activation of TET-mediated hydroxymethylation. Recent evidence suggests that there may be regions resistant to even this second round of epigenetic reprogramming – genes which may be mediators of transgenerational epigenetic inheritance (Tang et al., 2015). These regions may be sensitive to environmentally induced changes that may persist and become heritable suggesting a mechanism of increasing evolutionary diversity (Tang et al., 2015). The mechanism of resistance is an active field of research; however, it has been found that there is enrichment of H3K9me3, Krueppel-associated box (KRAB)-associated protein 1 (KAP1), and ZFP57 motif – together may recruit residual methylation machinery. The demethylation process takes about four weeks in humans and doubling time of the PGCs is about six days. The mechanism is thought to be via both dilution and active as described earlier: the timing of this cycle is thought to be in line with 5mC and 5hmC dilution theory of demethylation. TET-mediated dynamics are also inferred from the earlier loss of methylation of PGCs as compared to mouse PGCS (Hackett et al., 2013; Yamaguchi et al., 2013).

The male PGCs are remethylated via *de novo* methylation at around sixteen weeks with imprints specific to the sperm and not the oocyte, i.e. paternal imprints (Kelly et al., 2003; Guo et al., 2015; Tang et al., 2015). This remethylation process is thought to be continuous throughout fetal development and the majority of signatures are complete prior to birth in the male (Messerschmidt et al., 2014). The remethylation of some genes including *H19* are not fully completed until prior to prophase I in the adult (Kerjean et al., 2000). By the spermatocyte stage, *H19* is found to be completely methylated. At the *H19* ICR in the sperm, the paternal and maternal alleles are asynchronously *de novo* methylated with the paternal undergoing methylation first (Davis et al., 2000). *GTL2* is known to be methylated while *MEST* remains unmethylated after reprogramming in the male germ cells (Kerjean et al., 2000). Methylation marks are thought to be stable and maintained throughout life in these cells (Davis et al., 1999; Kerjean et al., 2000; Marques et al., 2011).

1.2.7 Spermatogenesis

After puberty in the male, the maturation of germ cells into sperm capable of fertilization potential is a process called spermatogenesis. The duration is on average 74 to 120 days, and begins a new cycle every sixteen days. A total of 200 to 300 million sperm are produced daily in fertile men. The following section is a brief overview of the process of spermatogenesis. Known roles of methylation in these stages will be included.

1.2.7.1 The testis

The testis is the male reproductive gland. The function of the testis is to produce sperm and androgens, specifically the steroid hormone testosterone. The site where spermatogenesis begins is the at the basal membrane of the seminiferous tubules. As diploid spermatogonia undergo meiosis to produce four haploid mature sperm, they travel towards the lumen of the testis via a duct system in the seminiferous tubules. The spermatid is released into the lumen where they will travel via the efferent ductile to the epididymis and continue to mature.

1.2.7.2 Spermatogonia

The process of a diploid spermatogonia maturation into a haploid spermatocyte is termed spermatocytogenesis. Type A spermatogonia are differentiated from gonocytes and are arrested in mitosis until puberty. Nuclear staining of Type A spermatogonia show two distinct types, the type A_{Dark} (A_d) and type A_{pale} (A_p). Type A_d are thought to be reserve spermatogonia stem cells as they only undergo mitotic activity to replenish their population or commit to Type A_p spermatogonia. Type A_p spermatogonia actively undergo mitosis to maintain the pool of A_p stem cells, and once every sixteen days, a portion of type A_p cells commit to spermatogonia type B and enter meiosis producing primary spermatocytes (Figure 2).

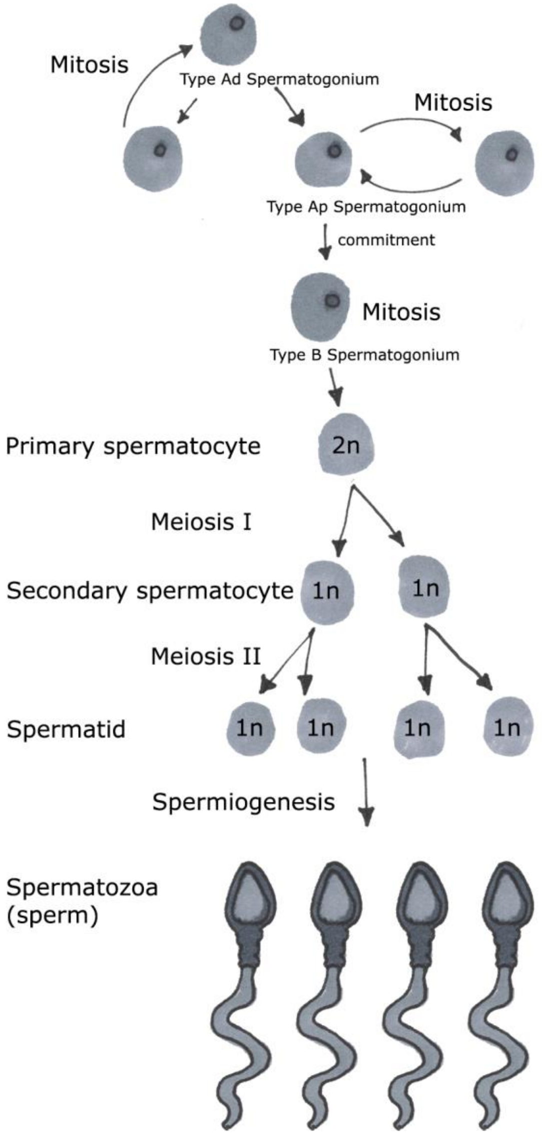


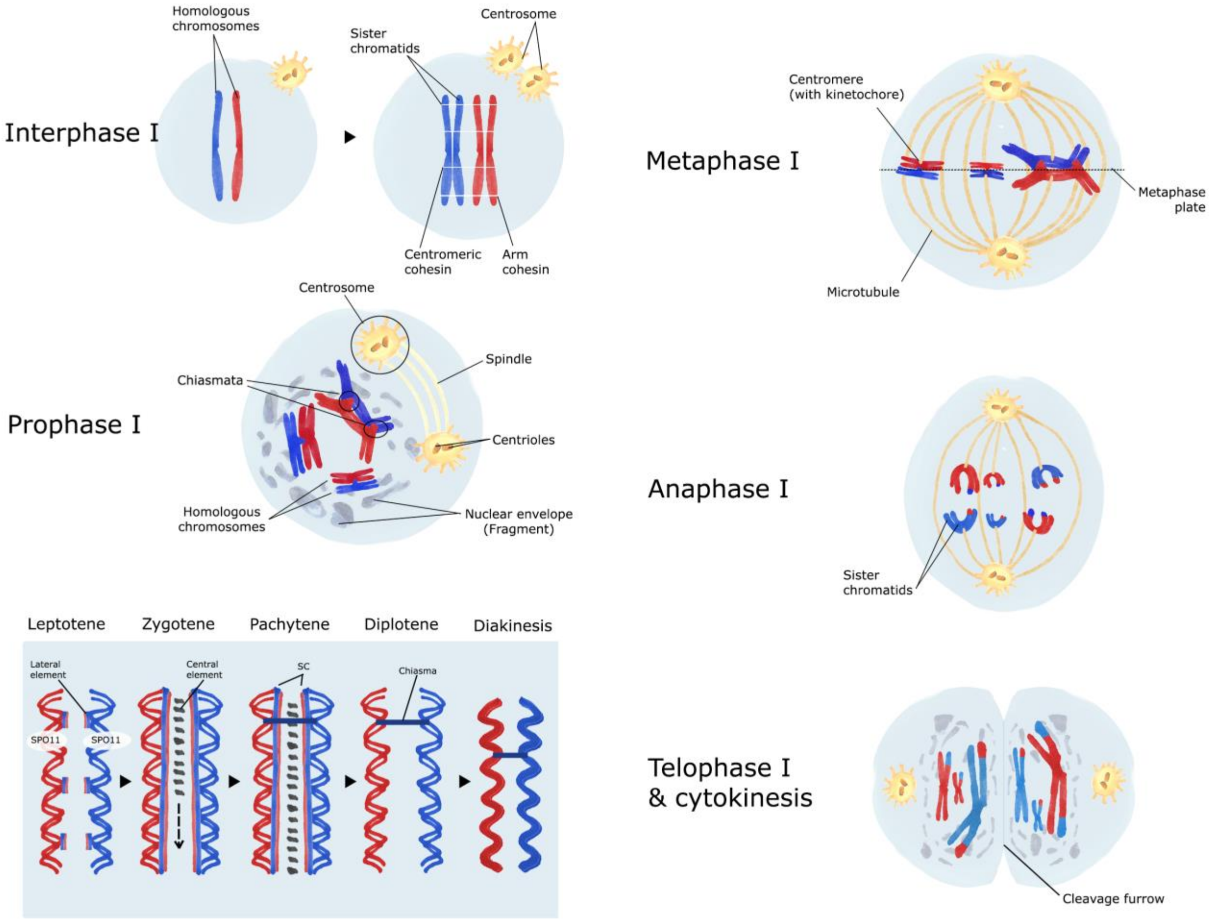
Figure 2. Spermatogenesis
Diploid spermatogonia undergo mitosis to either commit to meiosis or to replenish their population. Spermatogonia type B cells are committed to give rise to spermatocytes which enter into meiosis to further give rise to four haploid sperm. The site of spermatogenesis is in the seminiferous tubules in the testes. Sperm are released into the lumen of the tubule where they travel via contractions to the epididymis for spermiogenesis. Once matured, they are prepared for ejaculation.

1.2.7.3 Meiosis

Meiosis is a form of cell division that results in the regulated halving of chromosomes forming four haploid cells genetically distinct from the parent cell. The first step of the meiotic process in males is the migration of primary spermatocytes through the blood-testis barrier (BTB). The BTB is a tight, gap, and desmosome junction formed between adjacent Sertoli cells, which forms two distinct and separate spaces – the basal compartment which houses only spermatogonia and the adluminal compartment housing maturing spermatocytes. The function of the BTB is to control the environmental conditions of the adluminal space, including preventing toxins from entering the space. The process of migration involves the disassembly of the BTB as the primary spermatocyte crosses, followed by reassembly post-migration. This restructuring process involves highly regulated pathways that must be coordinated with cell cycle progression and cell migration. Once crossed, the primary spermatocyte progresses into prophase I of meiosis.

The primary spermatocyte begins in interphase of the cell cycle (Figure 3). The first portion of interphase of meiosis, the cell synthesizes the proteins and enzymes needed for growth later on – termed growth 1 or G_1 phase. The 2nd portion of interphase involves the replication of genetic material from a single molecule of DNA into two identical sister chromatids attached at a centromere – termed synthesis or S phase. Interphase is next followed by meiosis I.

Meiosis I



Meiosis II

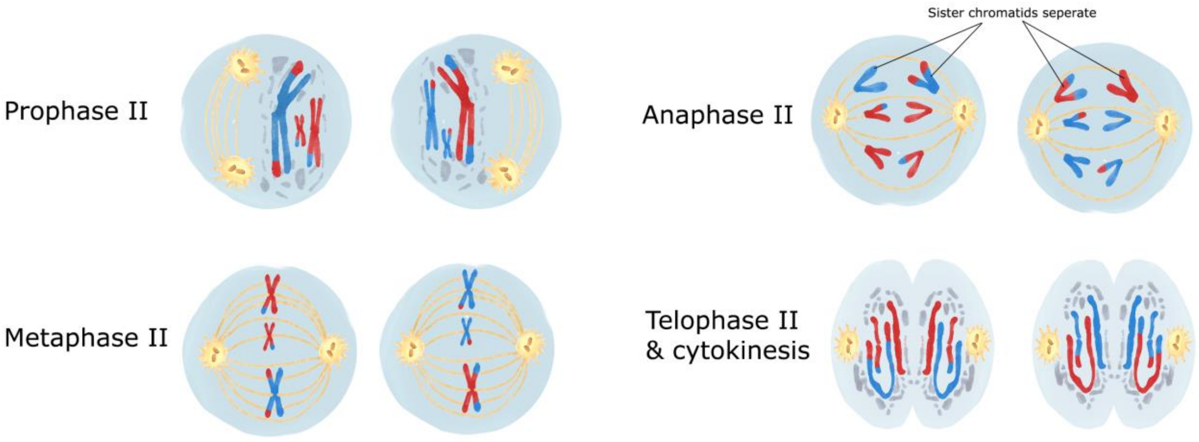


Figure 3. Meiosis
Meiosis is comprised of two specialized cell divisions that result in the reduction of ploidy number. Meiosis begins at interphase where DNA replication occurs (top left). Once committed, the cell enters prophase I (vertically down from interphase panel) where in itself is composed of five steps. At the end of prophase, the homologous chromosomes are paired together in a process called synapse and align at the cell's equator during metaphase I. Due to the presence of cohesins, the segregation of homologous chromosomes instead of sister chromatids occur during anaphase I. Telophase I and cytokinesis produces a cleavage burrow which separates the cytoplasm and nuclei into two daughter cells. Meiosis II is a similar process but does not reduce the ploidy number. Instead sister chromatids are separated during anaphase II.

1.2.7.3.1 Meiosis I

Meiosis I is often referred to as reductional division. At the end of interphase, the preleptotene primary spermatocyte is diploid (2n) with replicated genetic content in the form of sister chromatids on each chromosome. The genome at this point consist of two sets of chromosomes, i.e. homologous chromosomes, one from each parent and each set joined together as a tetrad (2n, 4c). The end result of meiosis I segregates homologous chromosomes producing two haploid daughter cells each with half number of chromosomes (Figure 3).

The main outcome of prophase I – the first phase of meiosis I – is the pairing and recombination of homologous chromosome pairs. Prophase I is broken down into numerous stages: leptotene, zygotene, pachytene, diplotene, and diakinesis. In **leptotene**, each chromosome which consists of two sister chromatids condense and form visible strands within the nucleus (Figure 3). Lateral elements of the synaptonemal complex assemble. In **zygotene**, the chromosomes pair together into homologous pairs facilitated by the transverse filaments of the synaptonemal complex (SC). Pairing is highly specific and exact as individual pairs are equal in length and in position of the centromere. Thus, pairing occurs only between homologous chromosomes, i.e. the same chromosome originating from the different parents. Pairing occurs in a zipper-like fashion and may start at any position along the length of the chromosome. The formed homologous pair is known as a tetrad or bivalent. The process of pairing is also known as synapsis. **Pachytene** stage is when homologous recombination occurs. Double strand breaks (DSB) in the DNA are induced by SPO11 thereby allowing DNA to cross over. The paired tetrads each containing two similar chromosomes – each with two sister chromatids – may form points of crossing over between non-sister chromatids, i.e. chromatids between homologous pairs. The points of contact are known as chiasmata and from these points, segments of DNA may cross over. The sex chromosomes also may pair and cross over the small region of homology known as the pseudoautosomal region. The process of creating chiasmata and crossing over is critical for the stability of pairing. DSBs that do not result in a crossover event are repaired. It has been shown that reduced recombination at this stage results in the improper segregation of chromosomes in later stages (Ferguson et al., 2007). The recombination during prophase I due to crossing over results in new combinations of DNA on chromosomes. This is the first occurrence of events where the daughter cells diverge from the genetic lineage of the parent cells. This is thought to produce offspring with genetic variability to increase survivability and adaptivity to changing environmental conditions. In the diplotene stage, the synaptonemal complexes break down and the homologous pairs form a slight separation, which can be seen as two threads under a microscope. DNA uncoils slightly and allows for transcription. However, chiasmata points remain holding the homologous pairs together – and will remain until anaphase I. In the last phase of prophase I – **diakinesis** – chromosomes further condense, the nucleoli disappear, the nuclear membrane disintegrates into vesicles, and the meiotic spindles initiate formation from two centromeres, which were migrating to the poles of the cell during the stages of prophase I. The centromeres were formed during S phase of interphase. Microtubule spindles invade into the nuclear region once the membrane disintegrates and connect to the tetrads at kinetochores – four on each tetrad. The pair of kinetochores on each sister chromatid fuses therefore each sister chromatid and thus whole homologous chromosome separates from each other towards opposite poles during metaphase I.

In the next phase, metaphase I, the tetrads move to the equator of the cell known as the metaphase plate (Figure 3). Diversity of genetic makeup after meiosis also occurs during metaphase I. The segregation of homologous chromosomes occurs according to the law of independent assortment: the independent orientation of the chromosome pairs along the metaphase plate during this phase allows for the random and independent distribution of chromosomes during later phases.

Next is anaphase I, where the cohesin arms – that were produced during S phase and bind the sister chromatids laterally along the chromosome arms – are degraded. However, the meiotic cohesins that bind near the centromeres are not degraded due to the specific subunits not present on the arm cohesins. These subunits are modified within the recombination protein 8 (REC8), where the arm cohesin REC8 proteins can be phosphorylated resulting in the removal by protease separase (Watanabe et al., 1999; Katis et al., 2010). The centromeric cohesins are protected from phosphorylation by shugoshin 1 (SGO1) recruitment of phosphatases. Therefore, the presence of centromeric cohesins but not arm cohesins enable the motor-like movement by the kinetochore (shortening of microtubules) to create tension on the centromeres instead of the arms - the end result is the separation of homologous chromosomes in anaphase I but not sister chromatids. The tension is a byproduct of at least one crossover event per tetrad. The cell will not progress past anaphase I unless there is tension on every tetrad. Therefore, the lack of crossing over may result in the arrest of meiosis I.

The first meiotic phase ends when the chromosomes arrive at the poles. At this point, known as telophase I, each daughter cell has half the number of chromosomes (haploid) but each chromosome has a pair of sister chromatids. The microtubule spindle network disappears and a new nuclear membrane surrounds each haploid set. The chromosomes uncoil to once again become chromatin. A pinch from the cell membrane is formed between the haploid regions. This is known as cytokinesis and the end result is two daughter cells each containing a haploid set. At this point, each daughter cell may enter an interphase known as interkinesis or interphase II.

1.2.7.3.2 Meiosis II

Meiosis II is similar to meiosis I with the exception of producing daughter cells with the same ploidy number, i.e. not a reductional division process. Instead, the daughter cells will contain only a single chromatid of each chromosome – often termed equational division or segregation and is analogous to mitosis. The phases of meiosis II are called the same as meiosis I, except with the designation of II (Figure 3).

Within prophase II, the SC and thus synapsing does not occur. Consequently, recombination and crossing over do not occur. As with prophase I, the chromosomes condense once again. Metaphase II is similar to metaphase I, while anaphase II proceeds with the removal of the centromeric cohesins. Therefore, the kinetochore and microtubule spindle network separate sister chromatids. This is made possible due to either the inhibition or degradation of

SGO1 (Liu et al., 2013). The end result is the production of two haploid and single chromatid spermatids, and overall four spermatids from the original primary spermatocyte.

1.2.7.4 Spermiogenesis

Haploid round spermatids undergo morphological changes before becoming mature sperm in a process called spermiogenesis. There are traditionally four phases: the Golgi phase, the cap phase, formation of tail phase, and the maturation phase.

The Golgi phase provides asymmetry to the radially symmetric round spermatids at this stage. At one end destined to be the head, the Golgi apparatus creates enzymes that will later become the acrosome. At the other end, a thickened mid-piece forms for the mitochondria to gather. One of the now obsolete centrioles begin to reform into the axoneme. At this stage, the spermatid DNA compacts to accommodate the reduction in cell size in later sizes. This is achieved via the replacement of histones and nucleosomes with specific basic nuclear proteins. The shaping of the sperm head and reduction of the body begins with the migration of the nucleus and acrosome to one side of the cell body. The acroplaxome and the ectoplasmic specialization (ES) on the Sertoli cells assist in reshaping of the nucleus from spherical to oval (O’Donnell, 2014). The ES also reorients the nucleus towards the basal membrane and anchor the spermatid to Sertoli cells. The elongation of the round spermatids is facilitated by the manchette – a microtubule bundle. The manchette traverses towards the rear of the spermatid and reshapes the nucleus along the way. The manchette is also thought to reposition the cytoplasm towards the back of the cell for removal in later steps.

The cap phase is characterized by the formation of an acrosome cap, which is a membrane bound organelle found attached at the front of the sperm head and between the nuclear and plasma membranes. The function of the acrosome is critical for fertilization as it releases hydrolytic enzymes that digest the zona pellucida thereby allowing the sperm to bind to the oocyte. The Golgi apparatus surrounds the condensed nucleus, where vesicles containing hydrolytic enzymes bud off and combine to the front head region of the spermatid. Cytoskeleton elements including the perinuclear theca and acroplaxome, anchor the acrosome in place.

The formation of the flagellum is the next phase. The function of the flagellum is to provide forward progressive motility to the mature sperm. Shortly after meiosis, the core of the flagellum called the axoneme is assembled from the other centriole. The structure is composed of a central pair of microtubules surrounded by nine additional pairs of microtubules to form a 9+2 arrangement. There are also dynein motor proteins on the outer pairs of microtubules providing movement. Dense fibers, fibrous sheath, and mitochondrial sheath are secondary structures that assemble around the axoneme during elongation. The Golgi provides proteins transported on the acroplaxome and manchette cytoskeleton network. The manchette also assists in the elongation of the flagellum. The ES junction and Sertoli cell cytoskeleton transport the elongated spermatids to the luminal edge. The elongated spermatid orients itself with its tail pointing towards the centre of the lumen and away from the epithelium in preparation for release. The maturation phase removes excess cytoplasm, known as the residual body of the regaud, phagocytosed by the surrounding Sertoli cells in the testes.

In the final stage termed spermiation, sperm obtain its remaining functions. The sperm are released into the lumen of the seminiferous tubule. This is made possible by modified endocytic structures called tubulobulbar complexes (TBC), which appear between the Sertoli cell and spermatid thereby removing the right ES adhesion between the cells. TBCs are also thought to assist in the removal of the remaining cytoplasm in the sperm. The spermatid head is extended into the lumen by a cytoplasmic Sertoli cell stalk, whereas the cytoplasm remains anchored. As the spermatids release into the lumen, the cytoplasm is stripped and left behind as the residual body. This removal accounts for up to 70% of the sperm's mass.

Through peristatic contractions and in testicular fluid released by the Sertoli cells, the immotile sperm make it to the epididymis where they will fully mature and gain motility. The motility gained is primarily for the travel within the female reproductive tract. The movement of sperm to the external environment and through the male reproductive tract is achieved through muscle contractions. A glycoprotein coat over the acrosome prevents the sperm from initiating fertilization prior to making contact with the egg. The removal of this coat is activated by the joint effort of the enzyme fertilization promoting peptide (FPP), which is produced in the epididymis, and heparin which is produced in the female reproductive tract. The process of activating the sperm for fertilization is known as capacitation.

Histones undergo a progressive replacement to transition proteins, TP1 and TP2, and then to protamines. The highly positive charge of the arginine amino acids in the protamine structure enable DNA to wind more tightly than with histones. The use of toroidal structures also provides a geometric advantage over histones in providing higher compaction. The transition process is facilitated via histone hyperacetylation in the elongating spermatid, which opens up the chromatin structure. The ratio of protamine 1 and protamine 2 is about 1 in fertile men (Oliva, 2006). An altered protamine ratio has been associated with infertility and sperm DNA damage. However, not all histones are replaced; about 5-15% of the histones in the paternal genome are retained in the sperm (Jenkins et al., 2012). The role of these histones is still not yet determined but speculated to be involved with embryonic development in the inherited offspring.

1.2.7.5 Hormonal regulation

The regulation of spermatogenesis and secondary sex characteristics is controlled by the hypothalamic-pituitary-gonadal (HPG) axis (Figure 4). The hypothalamus is a gland located in the brain and secretes in periodic pulses the hormone Gonadotropin-releasing hormone (GnRH) by GnRH-expressing neurons. The regulation of GnRH expression and release is an active area of research. Recent evidence suggests that kisspeptin, ghrelin, leptin, insulin, and the neurosteroid axis all have regulatory effects on GnRH production (Roseweir et al., 2008), and therefore on spermatogenesis as well. GnRH travels down to the anterior pituitary via the hypophyseal portal system binding to receptors on the secretory cells of the adenohypophysis. The response to GnRH is the periodic pulsatile release of luteinizing hormone (LH) and follicle-stimulating hormone (FSH), which are released systemically, i.e. into the blood stream. FSH pulses are not as distinct and secreted more regularly due to other feedback mechanisms. In males, LH binds to and activates LH receptors on Leydig cells activating a signaling cascade where the steroidogenic acute regulatory protein (StAR) imports cholesterol into the mitochondria (Luo et al., 2001). Cytochrome P450 catalyzes a series of enzymatic reactions to convert the cholesterol into testosterone. A portion of the testosterone undergoes aromatase conversion into estrogen. Both testosterone and estrogen enter systemic circulation where they provide negative feedback regulation to the secretion of GnRH in the hypothalamus as well as LH secretion from the anterior pituitary. FSH binds to receptors on Sertoli cells which activates the transcription of gene expression important for spermatogenesis (Griswold, 1998). Inhibin B is also produced by Sertoli cells and negatively regulates the secretion of FSH from the anterior pituitary. GnRH is upregulated by activin which is produced in peripheral tissues (Meethal and Atwood, 2005). Follistatin is yet another peripheral hormone which downregulates GnRH through the inhibition of activin (Figure 4).

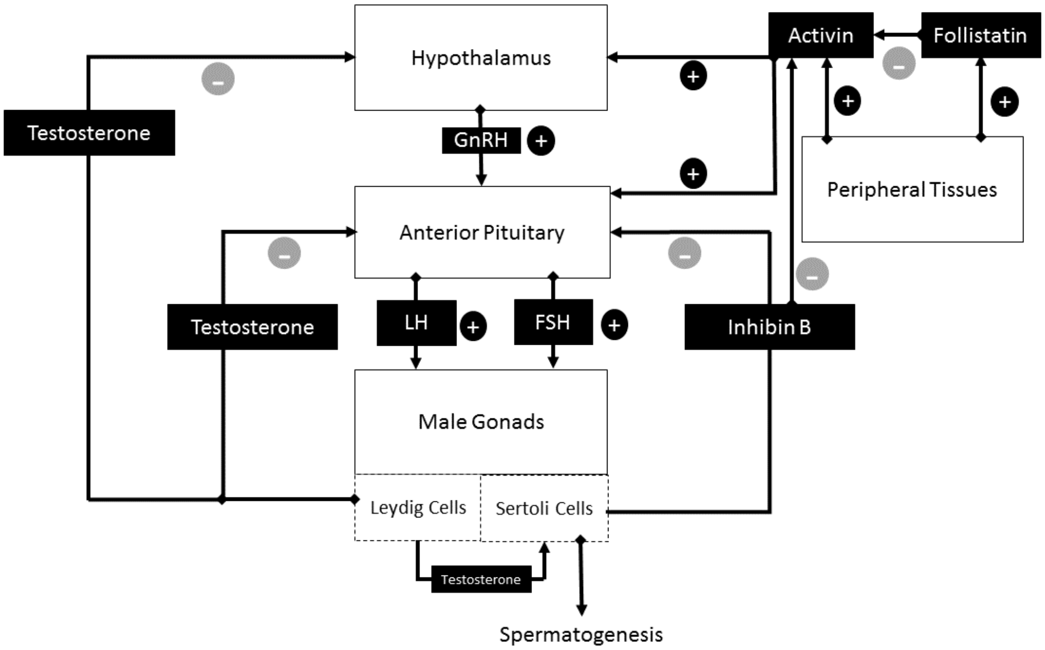


Figure 4. The feedback regulation of the hypothalamus-pituitary-gonadal axis
GnRH released by the hypothalamus in the brain travels to the anterior pituitary to release LH and FSH. These hormones in turn travel systemically to the gonads, i.e. testis in males, and induce the release of testosterone via Leydig cells. Testosterone promotes spermatogenesis through the Sertoli cells. Negative feedback loop of testosterone regulates GnRH, LH, and FSH release. Sertoli cells also negative regulates via Inhibin B. Activin from peripheral tissues play a role in upregulating the HPG axis.

The HPG axis is activated at puberty through the secretions of testosterone from the testes. Once active in males, the HPG axis is functional throughout life. This activation causes masculine anatomical changes to the male and begins the production of sperm. Testosterone has been shown to have effects on neural synapse development and migration. The result includes increased spatial reasoning, aggression, and sex drive; in females sex steroid hormones have sex differential effects including more neural connections between language areas resulting in better performance in communication as compared to males. The testes begin to produce less testosterone with age resulting in post-pubertal hypogonadism. Male characteristics are also decreased including progressive muscle loss, increase in visceral fat mass, loss of libido, impotence, and abnormal spermatogenesis.

1.3 Male infertility

Infertility is defined as the inability to conceive after one year of regular unprotected sexual intercourse. Infertility affects around 16% of couples, where up to half the cases are attributed to male infertility. As mentioned earlier, an estimated 37- 58% of male infertility cases are idiopathic (de la Calle et al., 2001; Moghissi and Wallach, 1983; Irvine, 1998; Hamada et al., 2012). The evaluation of an infertile male case begins with the semen analysis for sperm parameters.

1.3.1 Semen analysis

A routine semen analysis is used in the clinic to identify several semen parameters useful in the diagnosis of a male infertility case. The WHO's semen analysis guide refers to the 5th percentile lower reference limits (Table 1): Oligozoospermia is defined as a sperm count below

15 million sperm/mL; asthenozoospermia is defined as progressive motility below 32% of all counts; teratozoospermia is defined as normal sperm morphology forms below 4% of the counts. An infertile male can be a combination of phenotypes, where the observation of all three categories is known as oligoasthenoteratozoospermia (OAT). Complete absence of sperm from the ejaculate is known as azoospermia, which is commonly subdivided into two common phenotypes: non-obstructive azoospermia (NOA) and obstructive azoospermia (OA) – based on the pathology of azoospermia. NOA patients may have spermatogenesis but at very low levels (hypospermatogenesis), arrest at a certain step in meiosis (maturation arrest), or a complete absence of germ cells (Sertoli cell only syndrome or SCOS). OA patients usually have an obstruction in the vas deferens preventing sperm from leaving the urological tract. Azoospermic men can often have sperm in their seminiferous tubules. Histological analysis of the testicular seminiferous tubules can identify sperm used for ART and help differentiate between OA and NOA infertile men.

Table 1. WHO 5th edition 2010 lower reference limits for common semen parameters

Semen parameter	5th percentile lower reference limit (95% confidence intervals)
Semen volume (mL)	1.5 (1.4-1.7)
Total sperm number (10^6 per ejaculate)	39 (33-46)
Sperm concentration (10^6 per mL)	15 (12-16)
Total motility (PR + NP, %)	40 (38-42)
Progressive motility (PR, %)	32 (31-34)
Vitality (live sperm, %)	58 (55-63)
Sperm morphology (normal forms, %)	4 (3.0-4.0)

1.3.2 Causes of male infertility

Male infertility has been associated with non-genetic and genetic causes. Non-genetic causes include hormonal imbalances, cryptorchidism (undescended testes), acquired obstruction in the vas deferens, varicoceles, chronic illness, stress, immunology factors, drugs, chemotherapy, and radiation. There is also increasing evidence for exposure to environmental conditions including toxic waste, chemicals, or harsh work conditions to be linked to male infertility (Whorton et al., 1977; Oliva et al., 2001). The following section will go into details regarding the genetic factors of male infertility.

1.3.3 Genetic causes

Genetic causes are suggested to be involved in 15-30% of male infertility cases (Krausz et al., 2015). In approximately 20% of azoospermic men a genetic cause can be found (Röpke and Tüttelmann, 2017). The most severe genetic causes are whole chromosomal abnormalities such as Klinefelter syndrome, structural chromosomal abnormalities such as translocations, Y chromosome microdeletions, and Cystic Fibrosis gene mutations. Beyond these well characterized genetic causes, there are numerous single nucleotide polymorphisms (SNP) associated with male infertility and spermatogenic failure – many of which are in X-linked or autosomal genes. The following section will be a general overview of the known genetic causes of male infertility.

1.3.3.1 Chromosome Abnormalities

The prevalence of chromosome abnormalities in infertile men is up to 5% (O'Brien et al., 2010), including 13.7% of azoospermic men and 4.6% of oligozoospermic men as compared to 0.38% in the general population (Van Assche et al., 1996). Autosomal abnormalities such as translocations and inversions are more often found in oligozoospermic men. Sex chromosome abnormalities are more common among azoospermic men.

The first most common type of chromosomal abnormality in male infertility is chromosomal translocation. Translocations are the rearrangement of DNA between non-homologous chromosomes and can be either balanced (even exchange with no genetic material extra or missing) or unbalanced (uneven exchange resulting in extra or missing genes). Balanced translocations are usually silent or harmless in the individual; however, spermatogenesis and therefore fertility may be affected. The translocation of genetic material between non-homologous chromosomes may reduce the ability of these chromosomes from pairing during meiosis. Given that recombination and crossing over is critical to proper segregation of homologs, there is therefore a higher risk for improper segregation (nondisjunction) and thus an imbalanced number of chromosomes in daughter cells. Robertsonian translocations are the result of the fusion of two acrocentric chromosomes – that is, a chromosome in which the centromere is located quite close to one end – giving rise to one extremely small chromosome that may be lost and one large metacentric chromosome, i.e. X shaped chromosomes with the centromere in the middle and the two arms of almost equal size. The resulting karyotype in the individual is only 45 instead of 46. There is usually no direct effect on phenotype as the genes lost are minimal; however, as with other translocations there is a risk for producing unbalanced gametes due to the complete lack of a pairing partner for the chromosome during meiosis. Robertsonian translocations have a prevalence of 0.8% in infertile men.

The second major type of chromosomal abnormality that affects fertility is aneuploidy. Aneuploidy is the deviation from the correct number of chromosomes according to cell type. The sex chromosome aneuploidies are of special interest in fertility due to the fact that the sex chromosome genes are mostly hemizygous (only one copy) with the exception of the pseudoautosomal regions (PAR1 and PAR2). One common sex chromosome aneuploid in men that results in infertility is Klinefelter syndrome, which is characterized by the gain of an extra X chromosome. The most common karyotype is 47,XXY which is found in 80-90% of Klinefelter syndrome men. The prevalence of Klinefelter syndrome is 3% of all infertile men (Bojesen et al., 2003) and 14% of NOA men (Krausz et al., 2014). An estimated 1 in 1000 newborn boys are affected by 47,XXY. The extra X chromosome interferes with male sexual development with varied results including the following: microchidism (small testes), cryptorchidism (undescended testes), hypospadias (the opening of the urethra is on the underside of the penis), micropenis, learning disabilities, and delayed speech and language development. The effect of microchidism on fertility is due to the inadequate levels of testosterone produced by the Leydig cells. This in turn affects the development of male sexual characteristics during puberty. The supernumerary sex chromosomes are also considered to impair meiosis resulting in various phenotypes relating to spermatogenesis.

1.3.3.2 Single nucleotide polymorphisms

Also referred to as single nucleotide variations, SNPs are mutations to the genetic code resulting in the alteration in gene expression and/or protein structure that occurs in at least 1% of the population. Surprisingly there are numerous SNPs described in literature that are not on the Y chromosome relating to male infertility. In higher order mammals including humans, due to the instability of the Y chromosome as a result of ampliconic regions exhibiting non-homologous recombination, the present-day Y chromosome have lost most of the ancestral Y genes, many of which may have been involved with spermatogenesis and fertility (Röpke and Tüttelmann, 2017). Many of these genes have instead moved to other chromosomes including the X chromosome. The following section will describe some of the known genes and their mutations that show high expression in the testes and may be involved with male infertility.

1.3.3.2.1 Cystic fibrosis transmembrane regulator

Cystic fibrosis is a recessive autosomal genetic disorder affecting most commonly the cystic fibrosis transmembrane regulator (CFTR) which regulates chlorate transportation (Chen et al., 2012). There are more than 1500 mutations known in this transmembrane protein. The incidence is approximately one in 2000-3000 individuals in Europe. CTCF mutations have been associated with congenital bilateral absence of the vas deferens (CBAVD) typically resulting in OA. CBAVD occurs in up to 25% of OA patients (Vogt, 2004) and up to 97-98% of male cystic fibrosis patients. Hallmarks of CBAVD includes bilateral agenesis of the vas deferens and atrophy or absence of the seminal vesicles and a large portion of the epididymis. The most common mutation associated with CBAVD is F508del with a prevalence of 17% among these patients, which results in the misfolding of the protein.

1.3.3.2.2 Methylenetetrahydrofolate reductase

The substrates which DNMTs use for methylation reactions are in part derived from the folate cycle (Figure 5). Folic acid enters the cell and is reduced into dihydrofolate (DHF) and next to tetrahydrofolate (THF). A methyl group is transferred to THF from serine forming 5,10-methyleneTHF (5,10-mTHF). Methylenetetrahydrofolate reductase (MTHFR) is the rate limiting enzyme in the cycle and catalyzes 5,10-mTHF to 5-methylTHF (5-mTHF). The production of SAM or AdoMat (Figure 5) through methionine produces a methyl donor that can be used directly for methylation reactions of histones, CpGs, etc. A deficiency in MTHFR activity reduces SAM thereby reducing the available methyl donor pool for methylation reactions. In humans, the complete loss of MTHFR does not usually result in a live birth while a homozygous nonsense mutation results in severe neonatal abnormalities (Goyette et al., 1995). SNPs within the *MTHFR* gene has been shown to be linked to male infertility in multiple ethnic populations (Gupta et al., 2011; Naqvi et al., 2014; Gong et al., 2015). The most common mutation associated is the C677T SNP which switches the 677th bp from thymine to cytosine, which produces a thermolabile variant with significantly decreased enzyme activity (Frosst et al., 1995).

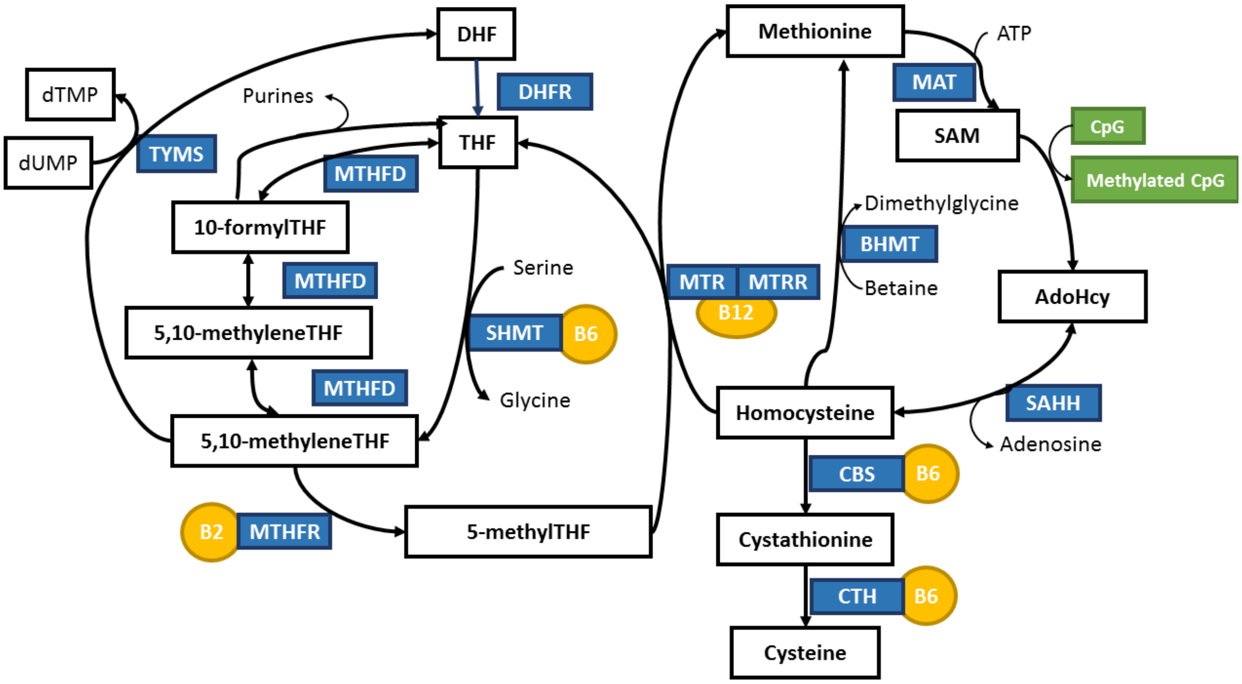


Figure 5. The folate cycle
Methylation reactions depend on the pool of methyl donors such as S-adenosyl methionine or SAM. The production of SAM derives from the reduction of folate. Folate enters the cycle via the reduced form THF. The transfer of a methyl group from serine to THF produces 5,10-methyleneTHF. The rate limiting enzyme of this cycle is MTHFR which catalyzes the 5,10-methyleneTHF to 5-methylTHF – ultimately producing SAM. Figure adapted from Hiraoka et al., 2017

1.3.3.3 Copy number variations

Since the development of genome-wide technologies such as SNP arrays or array-comparative genomic hybridization (aCGH), the ability to detect submicroscopic deletions and duplications has been possible – termed copy number variations (CNV). CNVs have been suspected as normal or non-lethal in our genome, where the majority of CNVs are not related to disease (Redon et al., 2006; Röpke and Tüttelmann, 2017). Although there are numerous candidate X-linked CNVs in male infertility, none have a definitive cause-effect relationship and been verified in an independent study (Krausz et al., 2012; Chianese et al., 2014; Giacco et al., 2014). In contrast, the microdeletions of the male-specific region on the Y chromosome (MSY) are a clear example of CNVs linked to male infertility – so much that it is routine practice to detect for these deletions in the clinical setting.

1.3.3.3.1 Y chromosome microdeletions

Y chromosome microdeletion (YCM) is an example of CNVs and is a genetic disorder characterized by the deletion of genes on the Y chromosome. YCM are thought to be the 2nd leading genetic cause of male infertility: up to 20% of oligozoospermic and 16% of azoospermic men have some form of these microdeletions (Oliva et al., 1998). The high incidence of Y chromosome microdeletions is attributed to the intrinsic instability of the chromosome (Bachtrog, 2013). The Y chromosome divergently evolved from the X chromosome (Charlesworth et al., 2005). While doing so, the chromosome gained male specific genes for the determination of the testis. However, in doing so the Y chromosome lost homology with the X chromosome, which is essential for recombination during prophase I of meiosis. While other chromosomes have copy error correction mechanisms via pairing during recombination, the Y chromosome does not have a homolog pair during meiosis. The absence of a partner in this male MSY region is suggested to be the driving force for the high rate of inversions, duplications, and deletions in this region. Without a pairing partner during meiosis, stability is partially generated from intrachromosomal pairing from repeats and palindromes in the form of hairpin structures. However, this also creates opportunities for recombination crossing over events with the potential for strand slippage or unequal crossing over events – leading to further inversions, deletions, duplications, and deletions.

The MSY region houses spermatogenic genes found within the long arm called the azoospermia factor (AZF) region. The AZF region is the most commonly deleted region within the MSY and is subdivided into the AZFa, AZFb, and AZFc. The deletions of genes within this region is a cause of infertility ranging from oligozoospermia to azoospermia. Within the AZFa region, the deletion of the Y-linked DEAD-box helicase 3 (DBY) gene results in SCOS. DBY deletions occur in up to 1% of cases (O’Brien et al., 2010). The AZFb region houses the Y-linked RNA binding motif protein (RBMY) gene. Deletions within this region result in azoospermia including spermatogenic arrest at the primary spermatocyte stage. The AZFc region is the most commonly deleted region, accounting for up to 70% of YCM cases (Oliva et al., 1998). The phenotypical result of AZFc microdeletions are varied, ranging from normozoospermia (normal sperm parameters) to azoospermia.

1.3.3.3.2 X-linked genes

In contrast to the unstable Y chromosome, the X chromosome has been predicted to vary little in mammals (Mueller et al., 2013). Comparisons of the X chromosome between mouse and humans have confirmed this stability. However, it was found that 10% of human and 16% of mouse X-chromosomal genes did not have orthologs in other species. In fact, these unique genes were found to reside in the ampliconic regions which made them difficult to study due to the high repetition. RNA sequencing showed that almost all these genes were expressed only in males and predominantly in the testis. Therefore, it has been suggested to be evolutionarily stable and evolving toward a male specialization. The X chromosome has been found to have 1098 genes, 99 of which encode proteins expressed in the testis (Ross et al., 2005). The androgen receptor (*AR*) gene is a proposed X-linked gene that causes androgen insensitivity syndrome (AIS) – two polymorphisms, CAG and GGN, located on exon 1 code for polyglutamine and polyglycine respectively. The longer lengths of these triplet codons are associated with decreased transcriptional activity of the gene *in vitro* (Pan et al., 2016). On the spectrum of AIS patients, *AR* mutations can lead to complete androgen insensitivity (CAIS) resulting in female phenotypes in karyotypic males (Röpke and Tüttelmann, 2017). Partial forms of AIS (PAIS) include ambiguous genitalia and mild forms (MAIS) where patients may exhibit hypospadias, gynecosmastia, or spermatogenic impairment. However, mutations in *AR* seem to be a rare cause of isolated male infertility (Hiort et al., 2000). Furthermore, the studies although show significance, the effect sizes are small and may not be clinically relevant (Tüttelmann et al., 2007), and that the GGN repeats are found in the normal population not associated with male infertility (Rajender et al., 2006).

The X-linked testis-expressed 11 (*TEX11*) gene codes for a protein critical for male germ cell recombination during meiosis (Adelman et al., 2008). Tex11 knockout male mice are azoospermic with maturation arrest at the pachytene stage due to the inability to repair DNA

DSBs. This gene has been recently associated with spermatogenic arrest in men with idiopathic infertility (Yatsenko et al., 2015), where the loss of 99 kb involving three exons of *TEX11* were identified in two azoospermic patients. An additional five CNVs were identified in 2.4% of the azoospermia patients and 15% (*n* = 33) of the azoospermic patients with maturation arrest, but not in the normozoospermic controls. This was further supported by immunohistochemical analysis of the testes showing cytoplasmic *TEX11* expression in late spermatocytes, round and elongated spermatids in controls, but this expression was absent in azoospermic patients. A recent finding also linked the autosomal *TEX15* as a cause in maturation arrest (Okutman et al., 2015).

The human reproductive homeobox (*RHOX*) genes are a cluster of genes on the X chromosome including *RHOXF1*, *RHOXF2,* and *RHOXF2B*. These genes are expressed in the oocytes and male germ cells. The mutations of these genes have been linked to impairment to regulate downstream genes including transcription factors and chaperons of the HSP70 family. Mutations in these genes have been identified in severe oligozoospermic patients (Borgmann et al., 2016).

The *ANOS1* gene (also known as *KAL1*) is located on the short arm of the X chromosome and encodes the extracellular matrix protein anosmin 1 which plays a role in the migration of GnRH-producing neurons (Cariboni et al., 2004). Gene deletions and mutations were identified in patients with hypogonadotropic hypogonadism (Costa-Barbosa et al., 2014).

The *USP26* gene belongs to a family of deubiquitinating enzymes that are responsible for processing inactive ubiquitin precursors, removing ubiquitin from cellular adducts, and rescuing macromolecules from degradation (Röpke and Tüttelmann, 2017). This gene resembles the *AR* gene and modulates its ubiquitination and therefore regulates its activity. SNPs and CNVs of *USP26* have been reported in both fertile and infertile men; the conflicting results make this gene unclear as a cause of male infertility.

1.3.4 Epigenetic causes

The investigation of epigenetic causes in infertile men were first prompted by the significantly higher risk of imprinting disorders among ART children (Maher et al., 2003). These disorders included Beckwith Wiedemann (BWS), Angelman (AS), Prader-Willi, and Silver Russell syndromes (Maher et al., 2003; Sutcliffe et al., 2005; Bowdin et al., 2007). BWS is characterized by overgrowth while AS patients exhibit mental retardation, speech impairment, and behavioral problems. The cause of these two syndromes are linked to a loss of function on the maternal allele at the KCNQ1 opposite transcript 1 (*KCNQ1OT1*) in BWS and small nuclear ribonucleoprotein polypeptide N (*SNRPN*) in AS patients. Both these genomic regions are ICRs. While the syndromes can occur in the regular population via deletions, mutations, uniparental disomy, the major causes of BWS and AS in the ART population tend to be the loss of methylation at their respective ICRs on the maternal allele (Maher et al., 2003; Grafodatskaya et al., 2013). Although the cause of higher imprinting disorders and methylation alterations is still an active area of research (Sakian et al., 2015; Vincent et al., 2016), the parental infertility – especially the inheritance of methylation from the sperm from infertile fathers – have been suspected as a contributing cause (Kobayashi et al., 2009), and is now an emerging research field in male infertility and ART. Given that ART newborns conceived from infertile fathers via the microscopic extraction of sperm via testicular biopsies and direct injection of the sperm into the oocyte, there is a high possibility of inherited methylation defects from the sperm of infertile men (see next section). Although the connection is unclear, it is plausible that the proper methylation in the sperm in infertile men may be a requirement for fertility given these dire consequences.

1.3.4.1 DNA methylation

The initial studies of methylation in the sperm was foremost to correlate altered methylation at imprinted genes in the sperm with outcomes of pregnancies via ART. Altered methylation at the *H19* and *IGF2* DMR were associated with a decrease in fertilization rate (Boissonnas et al., 2010). Identical altered methylation patterns were identified in paired samples of father's sperm and ART conceptus tissue (Kobayashi et al., 2009). Altered methylation at histone retained regions in idiopathic infertile men were also associated with poor blastocyst grading (Denomme et al., 2017). From these studies, it was clear that methylation was an associated factor in male infertility. Altered methylation at a number of genes have been identified over the past decade linking to male infertility (Kobayashi et al., 2007; Poplinski et al., 2010; El Hajj et al., 2011; Marques et al., 2004, 2008, 2010; Boissonnas et al., 2010; Laurentino et al., 2014; Louie et al., 2016). Studies investigating the methylation in testicular biopsies identified altered methylation at the promoter of *MTHFR* and discoidin domain receptor 1 (DDR1) of azoospermic men (Khazamipour at al., 2009; Ramasamy et al., 2014). The importance of methylation in spermatogenesis and male infertility was demonstrated by the use of hypomethylating agents in mice (Kelly et al., 2003). These agents include 5-aza-2'-deoxycytidine which incorporate into DNA and permanently bind DNMTs. The administration of this agent in adult mice resulted in decreased testes and epidydimal weights, and decreased sperm counts. Germ cells were also observed to be lost via germ cell apoptosis. As expected, pregnancies from the sperm of these mice resulted in decreased overall pregnancy rates, increased embryo loss, increased abnormal embryos, and preimplantation loss in females (Kelly et al., 2003). DNMT mutations were also evidence for the role of methylation in spermatogenesis. DNMT3A mutant mice were infertile, with reduced testes size and only few round spermatids (Yaman and Grandjean, 2006). Furthermore, the sperm were identified to have methylation loss at the *H19* and *IG-GTL2* DMRs. DNMT3L mutant mice are affected with meiotic abnormalities associated with meiotic arrest (Webster et al., 2005).

1.3.4.2 MicroRNA

The study of miRNAs in male infertility is an emerging field of research. The germ cell has been found to harbour numerous variants of ncRNAs, and are thought to play a role in spermatogenesis and fertility (Papaioannou and Nef, 2010). The sperm despite being transcriptionally inactive, are thought to carry miRNAs. Recent studies have identified miRNAs to be a diagnostic marker for male infertility (Wang et al., 2011; Abu-Halima et al., 2014). The inheritance of these miRNAs is also thought to regulate developmental regulators during early embryonic development (Hammoud et al., 2009).

1.4 Microarrays

DNA microarrays screen thousands of genes simultaneously via measuring signals from genetic samples using thousands of oligonucleotide probes bound to a silicon chip (Fodor, 1997). This technology has provided biological and medical researchers the ability to make inferences and predictions about a clinical question or phenotype in relation to gene expression, methylation, or mutations. The goal of a microarray experiment for medical researchers is typically to identify discriminant genes capable of classifying patients as either healthy or diseased and learn about the underlying biology of such phenotypes. The following section provides a brief introduction to the microarrays used in this thesis, the Affymetrix Genome-Wide Human SNP array 6.0 (SNP 6.0) and the Illumina Infinium Human Methylation450 BeadChip (450k), for genotyping and methylation analysis respectively, followed by an overview of the difficulties in analyzing such data. This is followed by an introduction to the methods used for analysis.

1.4.1 Affymetrix Genome-Wide Human SNP Array 6.0

The current technologies available for the identification of segmental duplication or deletions in the human genome include aCGH, genotyping microarrays, and next-generation sequencing (NGS). The detection of allele-specific copy numbers (CN) and SNPs on microarrays provide an advantage over aCGH in the detection of copy-neutral regions. Although NGS detects significantly more loci than microarrays, the substantial amount of data generated is more complex to analyze with limited support given that it’s a newer technology. The SNP 6.0 array contains over 1.8 million genomic markers, including 946,000 probes for the detection of CNVs and 906,600 for SNPs.

1.4.2 The Illumina Infinium Human Methylation450 BeadChip

The 450k array is a silicon based microarray that measures methylation at 485,512 CpG positions across the genome. The majority of CpG sites are selected at known or predicted regions that could potentially regulate gene expression. The 450k microarray measures methylation in one of two ways: type I probes measure methylation using two different colours, each from different probe types; type II probes measure using a single probe capable of emitting two different colors (fluorochromes) corresponding to either the methylated or unmethylated. Physically, a single 450k slide holds twelve arrays (6 x 2 grid), where each array measures a single biological sample. Facilities arrange 8 slides on a single plate therefore measuring a total of 96 arrays (8 x 12 slides) or samples per slide.

1.4.3 Standard 450k analytic pipeline

While the analysis of SNPs and CNVs is more straight forward, a standard pipeline for 450k analysis has not been established. The following section is a recommended 450k pipeline produced by some research groups and is a brief summary partially adopted from Wright et al. (Table 2; Wright et al., 2016).

Table 2. Major analytical steps in the 450k analysis pipeline

Analysis	Rationale
Sample filtering	Sample signals are compared to control probe signals on the 450k to identify those samples with inadequate detection. Samples with poor signal detection may be inaccurate due to poor sample quality.
Probe filtering	Similar to sample filtering, probes that inadequately detect methylation from numerous samples might be unreliable. Probes with sequences that bind to a SNP may be confounding due differential binding between samples and therefore unreliable. Unless the SNP has

	been previously assessed in the samples, it is preferable to remove such probes.
Within-array normalization	Removes background noise and corrects for technical dye based (red/green), intensity, and probe based (type I/II) differences within the array technology.
Batch effects	Samples are assessed for differences in technical differences, e.g. run on different days or facilities, and may remove these technical variations either through sample filtering again or statistical correction techniques.
Cell composition	Unless otherwise sorted, samples typically contain multiple cell types and therefore multiple methylation profiles. There may be different proportion of cells between samples and therefore introduce noise into the methylation signals. Statistical methods and the curating of methylation profiles from pure cell samples should be applied to estimate and correct for cell composition.
Differential methylation positions and regions	Methylation is assessed at specific and at broader genomic regions for differences between samples. This step can be assessed with many statistical techniques.
Biological and clinical interpretation	This step assesses the relationship of the significant hits with the disease or condition of interest. Manual exploration of literature is one option, while the use of numerous databases on the prediction of gene pathways has also been applied.

Adapted from Walker et al., 2017

Illumina provides a default software called GenomeStudio for the analysis of their microarrays. Although the interface is easy and intuitive to use providing convenient methods for generating data visualizations, the functionality of the software is limited and lacks support for more advanced and newer methods (Wright et al., 2016). The recommended software by researchers and which will be used in this thesis (Chapters 3 and 4) will be the free and open source R programming language. R was designed and is currently updated as a scripting language specifically for statistical computing and data visualization at an advanced level. With an active developing community, R is constantly updated with new packages taking advantage of the most cutting-edge techniques and analyses. Within the research field of genomics, R has numerous packages built and freely available for analyses including for the 450k microarray. This package repository is known as the Bioconductor project (Huber et al., 2015). A significant advantage of R due to its scripting language and open source repository of packages is that analyses are highly reproducible. Scripts designed for a specific dataset and project are essentially step-by-step documentation of the analytical workflow. Thus, the use of R for 450k analyses is highly favourable and well suited for reproducibility of research. For the remainder of this thesis, all analyses mentioned will be in reference to using R.

1.4.3.1 Sample filtering

The 450k microarray contains several types of control probes as reference when detecting the quality of the methylation signals from the samples. These probes include the evaluation of bisulfite conversion efficiency and background fluorescence levels to check the quality of the wet lab experimental steps (Triche et al. 2013). The quality control algorithm is to determine samples with a control probe intensity value outside the clustering of the other samples. The comparison of these values may provide insight into poor quality samples. The R Bioconductor package minfi (Aryee et al., 2014) provides a filtering function which compares the log mean intensity values of the raw methylated values against those of the unmethylated values. Samples of poor quality tend to deviate towards lower median values in both directions. In addition to the control probes, samples may be considered for removal based on the overall beta value density distribution plots, where samples that deviate from the general distribution may be removed. However, these differences in the beta density may be biologically significant and therefore should be evaluated carefully.

1.4.3.2 Probe filtering

The removal of specific probes that are of poor quality due to the technical experimentation, inherently designed poorly, or that are not informative for the study goals may improve the robustness and clarity of the analysis. Typically, probes are removed if their intensity levels are at or near background intensity levels as detected from the >600 negative control probes, signifying that the probe may have been altered or not properly manufactured. The usual method is to filter out probes that do not pass a threshold intensity value (*P*>0.01) in a specified portion of the samples, e.g. 10% of samples. Alternatively, a filtering criterion may be applied to probes that failed to hybridize if a specified number of beads are not detected (Morris et al., 2014); the 450k has a median of 14 probes per target sequence. In addition to poorly performing probes, some of the 450k probes bind to SNP sites. The reason is that the SNP probes may have differential binding affinities to specific SNP variants. Therefore, the methylation signal observed from such a SNP probe may be confounded by genetic diversity between samples and thus may reflect not a methylation difference but a genetic difference. The annotation of such probes is documented in the current version of the 450k manifest on Bioconductor (IlluminaHumanMethylation450kmanifest 0.4.0; Hansen and Aryee, 2012) and can be filtered manually or via minfi’s built-in function. Research teams may additionally remove specific probes based on the specific experimental design requirements. For example, researchers studying methylation in females only may opt to remove all probes detecting methylation on the Y chromosome. Lastly, recent evidence has identified that some 450k probes bind to multiple genomic sequences, rendering these probes non-specific (Price et al., 2013; Chen et al., 2013; Pidsley et al., 2016). The removal of such probes may improve the confidence in the methylation data.

1.4.3.3 Within-array normalization

The intensity measurement of each probe can be subdivided into a signal component, i.e. true intensity level of methylation, and a noise component, i.e. intensity measurement due to technical artifacts. The 450k's array design in using different types of probes requires adjustments to reduce variations and thus improve the signal to noise ratio of each intensity value. Thus, the goal of normalization is to adjust for non-specific background adjustments, red/green dye bias of type II probes, and rescaling of type I and type II probe differences (Morris et al., 2014). The use of advanced model-based background correction methods uses the intensity level of type I probes outside of their specified colour band (*n* = 135,501 probes) and has been shown to be produce improved results as compared to using the negative control probes (*n* = 42) (Triche et al., 2013; Ritchie et al., 2007). These methods include the normal-exponential convolution using out-of-band probes ("noob" method; Triche et al., 2013). The rescaling of type I and type II probes is necessary due to the fact that type II probes show a smaller range of beta values and larger variance between repeated measures as compared to type I probes (Wilhelm-Benartzi et al., 2013). This is further compounded by the fact that the placement of type I and type II probes differs in functional regions thereby introducing bias in differential methylation. One approach is the quantile normalization that is provided in the minfi R package, where the main idea underlying this correction method is that only modest changes are expected between experimental classes. An alternative method is the Funnorm normalization procedure where it has no major assumptions and has been shown to be effective in designs where there is a global methylation shift in samples, such as the case in cancer vs. normal tissues (Fortin et al., 2014). The subset-quantile within array normalization (SWAN) (Maksimovic et al., 2012) method is a within array normalization method that matches the type I and type II probes across subsets of probes differentiated by CpG content. This is followed by applying a quantile normalization technique, i.e. making all type I and type II probe distributions the same. This method has been shown to perform better than other methods (Fortin et al., 2014; Aryee et al., 2014). Yet another method is the stratified quantile normalization which is similar to SWAN but stratifies by region (Touletimat and Tost, 2012).

1.4.3.4 Batch effect analysis and correction

The term batch refers to a grouping of samples that undergo experimentation together. The issue with this is that often there is batch-to-batch variation which may confound the methylation results in that the intensities may be due to batch variation (Harper et al., 2013) as opposed to real biological significances. This may include batches run on different days or that the samples are such large sizes that they cannot fit on one microarray chip. This is especially important in designs where due to uncontrollable circumstances the groups of interest also correlate with the batch design, i.e. all infertility cases on one batch while controls on the other. This is often unavoidable for samples under a rolling collection schedule and that the recruitment of cases are infrequent such as with severe male infertility cases. If batch effects are suspected, the design of an experiment can mitigate such confounding effects through carefully randomizing experimental groups across array processing steps. Alternatively, downstream statistical models can be applied to reduce batch effects including the COMBAT function provided by the SVA package in R (Leek et al., 2017); however, unknown batch structures may exist that are not known and not easily predicted.

1.4.3.5 Cell composition correction

Unless otherwise sorted, most sample analyzed are usually composed of more than one cell type. For example, peripheral blood samples are composed of several cell types. Given that methylation profiles vary between cell types, samples that do have different cell proportions perhaps from random chance may confound the methylation results. Unless samples are identical in cell proportions, the differences between samples may present differential methylation that may be interpreted as biological significance but are only a result of slight differences in a certain proportion of cells, e.g. abnormally high proportions of eosinophil count in a patient due to an allergy reaction at the time of blood draw. Statistical corrections can be performed to estimate heterogeneity or proportion of cell types found in samples (Houseman et al., 2012). These methods are robust and rely on the methylation profiles (450k or newer) generated from pure cell populations. By estimating the proportion of cell types in each sample, a correction can be applied statistically when estimating methylation at each probe. However, not all samples have methylation profiles for their composed cell types, and thus may rely on newer innovations including reference-free approaches such as EWASher (Zou et al., 2014) and RefFreeEWAS (Houseman et al., 2014).

1.4.3.6 Calculation of differentially methylated positions and regions

The calculation of differentially methylated positions (DMP) or DMRs has been a topic of active research. The consensus on a single robust approach has not been decided with numerous studies using the 450k platform performing different calculations for DMPs and DMRs. There are many reasons such as the considerations for probes that are spatially close together show correlation in methylation. This correlation may be due to a real biological phenomenon and/or measurement errors (Jaffe et al., 2012); nevertheless, our biological understanding of methylation is that it is a coordinated effort to regulate a region, and not just a single CpG, thus the methylation of probes on the 450k that are within the same region may not be independent and thus may violate some assumptions of statistical methods, e.g. parametric t-test. To remedy this technical problem, instead of investigating each probe and thus CpG site independently, DMRs set at a specified region distance are investigated (Jaffe et al., 2012). This approach has been shown to be more robust and higher reproducibility than individual DMPs and can be found in the minfi R package as the bumphunter function. However, this approach is limited to only 20% of the 450k's probes due to the sparse placement of probe design; a significant of probes are not positioned within 1 kb of the next neighboring probe site (Ong et al., 2014). Furthermore, the complexity of this approach makes applications to more complex or individualized experimental designs more difficult; there is less flexibility for these functions for customization such as in multivariate designs. However, the use of a moderated t-test has been shown to be an effective method in determining single DMPs as sample sizes increase (Li et al., 2015) and many 450k studies have elected to use this method instead of bumphunting (Feinberg et al., 2015; Dere et al., 2017); thus, the consensus for a selection method for differential methylation analysis is not yet clear. More methods are presented in section 1.4.4.

The conversion of the methylation beta values to M-values (logit beta values) has been shown to promote normality and reduce heteroscedacity, i.e. unequal variance between groups, at the extreme beta values (Du et al., 2010). Therefore, the use of M-values in linear modelling would provide more robust results as compared to beta values. Furthermore, the manual filtering of more probes may reduce the burden of multiple testing; however, the use of the false discovery rate should still be applied for the correction for multiple comparisons (Benjamini et al., 1995; Michels et al., 2013).

1.4.3.7 Biological and clinical interpretation

This analytical step is intended to reveal potential biological mechanisms from the 450k findings with the associated clinical information. There are several interpretation approaches including comparing results directly with other publicly available datasets, using functional enrichment algorithms, comparing the genes of the DMPs or DMRs with literature, and investigating further the regulatory context with different types of epigenomic datasets (Wright et al., 2016). The comparison and combination with other publicly available datasets allows for the validation of results if there are similar findings, or for the mapping of novel gene pathways of unannotated genes by making the assumption that similar phenotypes should have similar genotypes (Suthram et al., 2010; Broen et al., 2011). This method in fact predicts and proposes new biological pathways of genes. Publicly available datasets can be found on the Gene Expression Omnibus (GEO; Edgar et al., 2002). The functional enrichment analysis uses databases of known pathways and applies algorithms – such as the protein analysis through evolutionary relationships (PANTHER; Mi et al., 2013; Thomas et al., 2003) that is part of the Gene Ontology (GO) Consortium – to find the likelihood that a determined set of genes from 450k results is associated with a specific pathway. The application of such analyses is inferring the biological processes, mechanisms, and canonical pathways that may be differentially methylated between patient groups. However, DMRs may be situated nearby numerous genes and the anticipated effect by methylation differences are not clear. The data curated by ENCODE (The ENCODE Project Consortium) and the Roadmap Epigenomics projects have developed a regulatory enrichment analysis (Hoffman et al., 2013). These datasets allow for the combination of different regulatory datasets, e.g. promoter region or chromatin state, together with the DMR data to better infer the pathways involved.

1.4.4 Feature selection

With advancements of microarray and sequencing technologies, a massive amount of data is now routinely produced. New analytical methods must be developed in order to make sense of the data. This is because of the small-*n*-large-*p* problem, i.e. small number of cases studied and large number of features or genes studied for each case, and that traditional methods are not optimized for identifying complex biological processes (Everson et al., 2015). The need to select a subset of informative features are needed (Dunning et al., 2008; Lynch et al., 2009), especially for a microarray where there is sparse measurement points and the vast majority of features are most likely uninformative.

In the section below, we will present the methods in evaluating differentially methylated genes or CpGs beginning with classic regression statistics, followed by machine learning (ML) algorithms. This thesis introduces ML as opposed to focusing on traditional statistical methods due to the overwhelming amount of new studies highlighting improved prediction accuracies of these algorithms (Changwon et al., 2014; Guyon et al., 2002). Many of these ML algorithms have been applied to cancer research for over a decade now; however, the use in epigenetic research in specific medical fields such as infertility has been sparse. Their good prediction is a result of being multivariate in nature and being built with a high level of accuracy that is suitable for genetic diagnosis and drug discovery (Guyon et al., 2002). It has enabled the accurate classification and diagnosis of patients with specific tumour types. The results of such analyses have enabled patients to receive the correct medication and treatments thereby enabling higher patient outcomes. The goal the following subsection is to provide a general background to ML concepts used in the analysis of microarray data. Although sequencing data can also be applied to such techniques, it is not the focus of this thesis.

The goal of feature/gene selection is first and foremost to identify a subset of genes which can classify cases by a phenotype with high accuracy, thereby producing a subset of genes that may be relevant the underlying biology. The inclusion of irrelevant and redundant information may harm performance of some ML algorithms (Wang et al., 2005). A known problem in classification is to find ways to reduce the dimensionality of *n* features (or genes in genomic studies) to overcome the risk of overfitting, i.e. a model error when a function too closely fit to a limited dataset and is inaccurate when generalizing to other datasets (Guyon et al., 2002). Overfitting occurs when the number of *n* features or genes is much larger than the number of patterns, or cases. This is very common among biological studies where microarrays are often conducted on few samples, while the microarrays with increasing advancements every year can be >450,000. A decision function in such cases such as a linear model can separate the data, but will perform poorly when generalizing to other datasets and thus make inferences about the underlying biology of the genes involved with the disease. Thus, the need to reduce the dimensionality arises and feature selection (often called gene selection in biological studies) aids in removing genes that are irrelevant to the classification problem. This in turn may discover a set of genes that have biological meaning to the disease at hand and is generalizable (thus not overfitted) to other cases.

Two general strategies are applied, filters and wrappers (Langley, 1994). Filters start by providing a general score to the starting training set. By removing genes, a score is provided for the gene. If the score is above a threshold, the gene is added to the feature subset. Wrappers incorporate ML algorithms when selecting features: subsets of features will be iteratively selected and a ML approach will provide a score through cross validation of training sets. A third type of feature selection method is the embedded type, which is a combination of both a filter and wrapper. The filter will be used to derived a subset of genes in which a ML algorithm will conduct a score. This is repeated until the best score is found. A feature set is finalized when the addition or removal of a gene does not improve the scoring metric. It is also possible to have multiple subsets that have the same score and therefore a feature subset can continually be modified so long as the score remains the same. Filters are generally much faster than wrappers due to the exhaustive nature of wrappers, which can be computationally expensive. The tradeoff is that wrappers provide higher accuracy. Of note, it is also conceivable to select the best possible set of features satisfying a model criterion by exhaustive enumeration of all subset of features; however, this is impractical for large numbers of features (e.g. >450,000 CpG sites) due to the large number of combination of subsets. Therefore, the need for methods to discriminate subset of features is needed. The section below will provide a brief introduction to many of the common filter methods used in research, particularly in microarray analysis.

1.4.4.1 Filter methods

1.4.4.1.1 Statistics

The most commonly applied filter approach is the use of a statistic comparing between groups. Parametric testing such as a t-test/linear modelling is widely used in all fields. Applying an empirical Bayesian moderation to t-test have produced generally acceptable results in methylation research (Li et al., 2015). Non-parametric testing such as Wilcoxon-rank sum test and logistic regression are applied as well, especially given that DNA methylation have been shown to not follow the normal distribution.

1.4.4.1.2 Principal component analysis

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations into values of linearly uncorrelated (orthogonal) variables termed principal components (PC). PCs are then ranked by variance. Often, PCA is used as a feature space reduction method (Duda, 1973), where the majority of variance (and usually the interesting data) is on the first few PCs while the remaining PCs (and thus features) contribute to very little of the differences. Thus, as a filtering method, PCA can identify features that contribute most to the variance in the observations. Approaches based on PCA have been previously used for the mining of genes from expression microarray data (Roden et al. 2006).

1.4.4.2 Embedded methods

ML algorithms such as support vector machines (SVM) and random forest (RF) are supervised classifiers, i.e. using *a priori* knowledge about a group, a trained classifier can label new unseen samples. As stated above, ML algorithms can be used for feature selection as wrappers. In this method, a search function is applied whereby genes are removed iteratively and a ML classifier will evaluate the accuracy of the subset of genes. This is repeated until the ML classifier identifies a reduction in accuracy. The gene set prior to the reduction is deemed the best set. An embedded method applies a filtering method instead of a search function.

1.4.4.2.1 Support vector machine recursive feature elimination

The goal of SVMs is to find the optimal separating hyperplane which maximizes the margin between groups in training data. This is a supervised machine learning algorithm and therefore requires training data with appropriate labels. SVMs are capable of discovering informative patterns (Guyon et al., 1996). Standard SVM is a linear classification using a dot product. However, a model for more complicated relationships can be achieved by replacing each dot product with a non-linear kernel function such as a Gaussian radial basis or Polynomial kernel. Gamma is the free parameter of the Gaussian radial basis function. SVMs have been shown to perform well for high dimensional microarray datasets (Furey et al., 2000).

The SVM recursive feature elimination (SVM-RFE) is a feature selection method originally designed for microarrays (Guyon et al., 2002). In the simplest form, SVM-RFE is composed of the following steps: 1) train a standard SVM model; 2) compute rankings of features; and 3) remove the feature with the worst rank. The step which “filters” the probes is ranking all features in the SVM primal problem, i.e. the coefficients to the SVM hyperplane. SVM-RFE is inherently able to reduce redundancy in the genes selected, which has been shown to improve the accuracy of gene sets (Jäger et al., 2002).

1.4.4.2.2 Balanced iterative random forest

RF is a machine learning algorithm used for classification developed by Leo Breman. The RF algorithm involves the growing of classification trees using randomly selected samples, i.e. bootstrapping, from the data to form training and testing sets. Each tree is a decision tree whereby within each node, the training set is partitioned into different classes with the split determined by a subset of randomly selected predictors or in this case methylation at genes. The trees are then aggregated to form a decision as a “forest” in a process known as bootstrap aggregating or bagging. The robustness of the classification by RF is a result of the randomness during bootstrapping and the random testing within each tree with a subset of features.

The balanced iterative random forest (BIRF) is a tweaked version of the RF algorithm. In cases where there is an imbalance in samples between groups, a normal RF will have a higher probability of randomly picking more samples from a certain group thereby influencing the training of the tree. In a balanced RF, a set proportional number of cases from each group used to grow trees. BIRF uses a balanced RF to train a forest. In addition, genes are iteratively removed based on an importance score (Gini index) and accuracy of the trained model is evaluated. This process is repeated until the accuracy of the model is lower than the previous model. Therefore, the best model is the subset of genes prior to the drop.

BIRF was shown to outperform SVM-RFE and other classifiers in the case of imbalanced datasets, i.e. unequal sample sizes in each group (Anaissi et al., 2013). The authors of the algorithm achieved 7-12% better accuracy over a variant of RFE, multiple SVM-RFE, in childhood leukemia datasets. Other studies have found RF to have comparable if not better performance and shorter computation time versus other methods (Díaz-Uriarte et al., 2006; Yao et al., 2015).

1.5 Rationale and hypotheses

Spermatogenesis is a complex process involving the coordination and expression of an estimated >2000 genes (Aston, 2014). Even before puberty, the development of the male sex and the male gonads is achieved through crucial timing of genetic and epigenetic events. Studies investigating the sperm of infertile men have identified that altered methylation at imprinted genes is more prevalent among infertile men (Marques et al., 2001; Laurentino et al., 2015). Given the importance of methylation in the regulation of expression, we suspect that methylation may also be regulating numerous genes involved with spermatogenesis. Thus, our hypotheses are that: (1) altered methylation at imprinted and somatic genes may be more prevalent in the sperm and testis of infertile men; and (2) the risk for altered methylation may be modulated by SNPs. While numerous studies have investigated methylation in infertile men, the majority have focused on the sperm (Marques et al., 2008; Boissonnas et al., 2010; Laurentino et al., 2015). Furthermore, there appears to be discordance between results. This is highlighted in the genome-wide association studies (GWAS) investigating SNPs in infertile men, where resequencing studies have routinely identified false-positives from previous results. Finally, there is sparse information regarding the analytical use of integrated analyses in the combined investigation of epigenetics and genetics in male infertility. The work presented in this thesis will address the following specific objectives.

1.5.1 Specific objectives

Objective 1a: To determine if altered methylation at imprinted genes are more prevalent in the sperm of more severe cases of infertile men.

Objective 1b: To determine if altered methylation at imprinted genes are more prevalent in men with the *MTHFR* C677T SNP.

Objective 2: To determine if there is altered methylation at non-imprinted genes in testicular cells in infertile men with impaired spermatogenesis.

Objective 3a: To determine if there are SNPs more prevalent among infertile men with impaired spermatogenesis.

Objective 3b: To determine if there is a combination of genes with altered methylation and SNPs more prevalent among infertile men with impaired spermatogenesis.

CHAPTER 2: INVESTIGATION OF DNA METHYLATION AT IMPRINTED GENES IN THE SPERM OF INFERTILE MEN IN ASSOCIATION WITH THE *MTHFR* GENOTYPE

2.1 Introduction

Infertility affects more than 48 million couple worldwide and causes significant stress on the lives of these couples (O'Moore et al., 1983; Newton et al., 1999; Cousineau et al. 2007). Of these couples, half are estimated to be attributed to male factor infertility. The most common trait among these men is oligozoospermia caused by impaired spermatogenesis (Stuppia et al., 2015). Although genetics have played a large role in identifying the cause in 15-30% of these cases (Krausz et al., 2015), the etiology of a large portion of cases remains unresolved, where an estimated 37-58% of all male infertility cases are idiopathic (de la Calle et al., 2001; Moghissi and Wallach, 1983; Irvine, 1998; Hamada et al., 2012). Epigenetics is proposed as an avenue of research to resolve some of these cases. Epigenetics is the study of stable and heritable modifications to DNA without altering the sequence itself that modulate the expression of genes. Methylation is an epigenetic modification which regulates the expression of genes via the modification of cytosine bps (most commonly found next to a guanine, CpG) with a methyl group (Smallwood & Kelsey, 2012). CpG islands which are long stretches of CpG sites usually near genomic elements such as promoters and enhancers are thought to directly regulate the expression of adjacent genes via physical impediment of transcriptional machinery (Messerschmidt et al., 2014). Therefore, the methylation in the male germ cells may be important for expression signatures critical for proper spermatogenesis.

The male PGCs during development undergo extensive methylation remodeling. Given the role of methylation in cellular differentiation, the purpose of such an epigenetic reprogramming is to reset the methylation marks that were indicative of the precursor cells (PGCs are induced from epiblast cells). The newly erased clean slate PGCs undergo a remethylation program throughout development and most of the marks are established and maintained in the mitotically arrested germ cells prior to birth (Tang et al., 2016; Kerjean et al., 2000; Marques et al., 2011). However, some regions including the imprinted ICR at *H19*/*IGF2* are not fully methylated until just prior to meiosis in the adult (Kerjean et al., 2000). The methylation of imprinted genes/regions are vital for the mono-allelic parent-specific expression needed for proper growth and development in the offspring (Kappil et al., 2015). Thus, the methylation in the germ cells may be important for the germ cell progression through spermatogenesis. Indeed, errors in DNA methylation at imprinted genes in the sperm have been linked to male infertility and dysfunctional spermatogenesis (Kobayashi et al. 2007; Marques et al., 2008; Minor et al., 2011; Urdinguio et al., 2015; Laqqan et al., 2017; Santi et al., 2017). In mice studies, the use of hypomethylating agents in adult mice result in reduced testes and epidydimal weights, and reduced sperm counts (Kelly et al., 2003) – showing a direct cause-effect relationship between methylation and spermatogenesis. In human studies, other than association studies between male infertile cases and methylation in the sperm, the origins of these alterations, however, are still not well understood.

The role of the folate cycle may be one clue into the mechanism of methylation in the sperm. The mechanistic addition of a methyl group to CpGs relies in part to the methyl pool in the form of SAM via the folate cycle. The rate-limiting enzyme in this cycle is MTHFR. The deficiency of MTHFR reduces SAM production, which in turn has been shown to alter cellular methylation processes (Chen et al., 2001; Friso et al., 2002; Castro et al., 2004). Therefore, the alteration in folate derived methyl pool via reduction in SAM may be a mechanism for the altered methylation in the sperm of infertile men. The severe deficiency in folate intake is unlikely in western societies as folate is fortified in most grain products. Furthermore, the complete loss of MTHFR does not result in a live birth in humans, while the homozygous nonsense mutation which produces a truncated MTHFR enzyme results in severe neonatal deformities – both of which would be unlikely in an infertile male adult case. However, the *MTHFR* C677T SNP is a common mutation in humans which produces a thermolabile variant of MTHFR with reduced catalytic abilities (Frosst et al., 1995). Therefore, it is plausible that the inheritance of this SNP may limit the available methyl donor pool for the critical epigenetic reprogramming events in the PGC reprogramming during gametogenesis. Similarly, the methyl pool may be reduced in instances during the active maintenance of methylation marks postnatally. It may even be further extrapolated to *in utero* development whereby if the SNP is inherited from the mother, the lack of methyl donors for the mother on top of a possible low folate diet may contribute to altered methylation in the developing fetus. Epidemiological studies have identified that the *MTHFR* C677T SNP has been associated with a variety of populations of male infertility (Gupta et al., 2011; Naqvi et al., 2014; Gong et al., 2015). However, the investigation of this SNP in relation to methylation alterations in the sperm has not been extensively studied.

2.2 Rationale

Given that not all studies have shown that infertile men have altered DNA methylation in their sperm (Santi et al., 2017), methylation may not be a typical phenotype of all infertile men. Therefore, there may be certain conditions that regulate the altered methylation in the sperm that is common among infertile men. Further, given the importance of MTHFR in the cellular availability of methyl pool, it is plausible that the C677T SNP may increase the risk for methylation alterations in infertile men.

The aim of this research study is to determine whether the *MTHFR* C677T SNP is associated with methylation in the sperm, and if so, is there a subset of men that are most at risk for these alterations. The primary objective is to evaluate the *MTHFR* C677T genotype in addition to the methylation of three known and susceptible imprinted genes (*H19*, *IG-GTL2*, *MEST*) for methylation alterations in the sperm of oligozoospermic men. The infertile men will be stratified by sperm concentration to determine whether severity of oligozoospermia is associated with methylation and/or the SNP.

2.3 Methods

2.3.1 Patient recruitment

Male partners of couples attending the UBC Centre for Reproductive Health or local fertility clinics in the Greater Vancouver regional district (GVRD) were recruited for this study. Only men who were unable to achieve a natural pregnancy after >twelve months of unprotected intercourse and who had >two semen samples with reported abnormal sperm parameters (World Health Organization, 2010) were included in this study. Control men with proven fertility (<one

year old biological child as confirmed via birth certificate) were included in this study. Participation included the donation of a semen sample into a semen collection cup, filling out a patient questionnaire regarding medical background, retrieval of medical history through the patient's chart, and drawing of peripheral blood into an EDTA vacutainer collection tube (at BC Children's Hospital Blood Clinic). Given the known association of chromosomal abnormalities and Y-microdeletions with male infertility (Krausz et al., 2015), only men with 46XY karyotype and the absence of Y-chromosome microdeletions in the AZF region were included, as taken from patient charts. Oligozoospermic infertile men were stratified by sperm concentration: normal (>15 million sperm/mL), moderate (5-15 million sperm/mL), severe (1-5 million sperm/mL), and very severe (<1 million sperm/mL). Informed written consent was obtained by a research coordinator prior to sample donation. This study was approved by the Research Ethics Board of the University of British Columbia (H06-03547).

2.3.2 Semen sample processing

After three days of abstinence, semen samples were donated by consenting men and immediately were incubated for 30 minutes at 37°C for liquefaction, i.e. the natural breaking down of semen fluid for an overall less gelatinous texture via prostate enzymes. Semen samples were analyzed under a bright-field Zeiss Axioplan microscope (Carl Zeiss Microscopy, Germany) and counted for sperm concentration using the Makler Counting Chamber (Sefi-Medical Instruments, Ltd, Haifa, Israel) according the manufacturer's instructions. Semen samples were then aliquoted into multiple 1.5 mL microfuge tubes and washed with modified HEPES buffered Human Tubal Fluid (mHTF; Vitrolife, San Diego, CA, USA) via 2-3 gentle inversions, and finally pelleted via low speed centrifugation for 30 seconds. All semen samples were washed three times prior to sperm isolation. Washed sperm samples with moderate to high sperm concentration and the presence of motility were subjected to standard swim-up procedure (Jameel, 2008). For swim-up, washed semen samples in 1.5 mL microfuge tubes were filled with a small amount of mHTF were incubated at 37°C for two hours tilted at a 45° position. After the incubating period, evaluation for pure sperm on a glass slide using a phase contrast microscope (Nikon, Tokyo, Japan) was conducted from an approximately 5 uL sample from the top layer of the tube. If enough sperm was observed, the top layer of the mHTF medium was carefully transferred to a 0.5 mL microfuge tube. Alternatively, samples with low or no motility or low sperm concentration could not undergo standard swim-up and thus were subjected to manual micromanipulation for sperm isolation. After washing and centrifugation from the previous step, semen samples were resuspended in mHTF and 10 uL was plated onto the centre of a 60 x15 mm petri dish. The droplet was briefly evaluated for the concentration of sperm and other somatic cells via a phase contrast microscope (Nikon, Tokyo, Japan) equipped with Hoffman modulating optics, thermal stage set at 37°C, and micromanipulators. Highly concentrated samples were diluted with mHTF until the field of vision allowed for custom-made micropipettes to isolate single sperm. Up to 10 more 10 uL mHTF droplets were positioned around the sample droplet and finally overlaid with mineral oil (Sigma-Aldrich Canada Ltd, Oakville, ON). Sperm were picked and deposited into a designated clean mHTF droplet; other droplets were used for cleaning the micropipette. Upon picking 200 sperm, the mHTF droplet with the clean sperm was transferred using a 10 uL micropipette tip into a 0.5 mL centrifuge tube. The position of the clean droplet on the petri dish was checked to ensure most of the sperm was transferred to the microfuge tube.

2.3.3 DNA extraction and bisulfite conversion

Swim-up sperm samples was digested using a modified protocol adapted from Doerksen et al. 2000: 3 mL of sperm lysis buffer (20 mM Tris pH 8.0, 10 mM Dithiothreitol (DTT), 150 mM NaCL, and 10 mM ethylenediaminetetraacetic acid (EDTA) pH 8.0), 1 mL of 10% sodium dodecyl sulfate (SDS), and 50 µl of 5µg/mL proteinase K (Invitrogen Canada Inc, Burlington, ON). Digested sperm samples were incubated at 60°C in a water bath overnight. DNA was extracted from digested samples by standard salt extraction method, including a wash with 70% ethanol, and resuspension in TE buffer (10 mM Tris, 1 mM EDTA pH 8.0).

The limited number of sperm from the micromanipulated samples required a more sensitive protocol (Manning et al. 2001). Isolated sperm samples were lyzed in 20 µL of alkaline lysis buffer (200 mM KOH and 50 mM of DTT) to decondense the sperm nuclei. Lyzed samples were immediately frozen at -80°C for a minimum of 3 days. Decondensed samples were thawed at room temperature before incubation at 80°C for 15 minutes on a thermoblock. A neutralization buffer (0.9 M Tris-HCL, 0.3 M KCL, and 0.2M HCL) was added to samples afterwards.

The extracted DNA samples (from either protocol) were subjected to sodium bisulfite for the detection of cytosine with methyl groups, i.e. DNA methylation. Bisulfite deaminates unmethylated cytosine bases into uracil, while the methylated cytosine bases are protected from this modification, thus remaining as cytosine bases. Thus, detection via sequencing post bisulfite modification allows for the single-base pair resolution detection of DNA methylation, which will read as cytosine bases, while unmethylated cytosine will read as thymidine bases. The EZ DNA Methylation Gold Kit (Zymo Research, Orange, CA) was used for sodium bisulfite of sperm samples according to manufacturer’s protocol, with the exception of incubation of samples for a reduced time to two hours only to limit the degradation of the already limited amount of DNA. Bisulfite modified DNA samples were stored at -20°C for short term storage prior to use, while for long term storage samples were stored at -80°C.

2.3.4 DNA Amplification

The sequences evaluated at the *H19*, *IG-GTL2*, and *MEST* DMRs were taken as previously published (Kerjean et al., 2000; Guens et al., 2007; Minor et al., 2011) (Table 3). The primers were selected based on their published lack of bias towards either maternal or paternal allele (Kerjean et al., 2000; Guens et al., 2007). Within the *H19* DMR sequence analyzed (NCBI accession number AF087017.1, from 6128 bp to 6299 bp), an informative SNP at the seventh CpG site (basepair 6124; SNP #1073516) may be used for allele-specific analysis (Table 3). However, the SNP nature of this site excludes it from DNA methylation analysis.

Table 3. Genomic sequences of investigated imprinted DMRs

Imprinted DMR	Sequence
H19	ctcctt[cg]gtctcac[cg]cctggatggca[cg]gaattggttgtagttgtggaat[cg]gaagtggc[cg][cg][**c**g][1]g[cg]gcagtgcaggctcacacatcacagcc[cg]agcc[cg]ccccaactggggtt[cg]cc[cg]tggaaa[cg]tcc[cg]ggtcacccaagcca[cg][cg]t[cg]cagggttca[cg]gg
IG-GTL2	cc[cg][cg]gctcaccagttgcc[cg][cg]actcaccaggtgcctg[cg]gctcaccagttgcctgtggctcaccagctgcc[cg]tggctcaccagctgcc[cg]tggcttacagttgcc[cg]aggctcacagttgcccatggcttgctaattgccag[cg]atttgccaattg[cg]agtggtt[cg]ccagttgcc[cg][cg]gtc[cg]ctaaacc[cg]taatcct
MEST	g[cg]ggctctg[cg]g[cg]cc[cg]gtgctctgcaa[cg]ctg[cg]g[cg]gg[cg]gcatgggataa[cg][cg]gccatggtg[cg]c[cg]agat[cg]cctc[cg]caggtgagtgtg[cg]gtgggaa[cg]agggggtgtggctgg[cg]gccctgggactaggg[cg]cagg[cg]ag[cg]gaggactgtgtgcc[cg]tgtcc

[1] Bolded C is a C/T single nucleotide polymorphism (SNP #1073516), therefore, this CpG will not be analyzed for methylation.

Table 4. *H19*, *IG-GTL2*, and *MEST* primers and PCR cycling conditions

Imprinted DMR	Primers	PCR Product Size (bp)	CpGs amplified (n)

H19	Forward: aggtgttttagttttatggatgatgg Reverse: tcctataaatatcctattcccaaataacc Nested Forward: tgtatagtatatgggtatttttggaggttt	172	18[1]
IG-GTL2	Forward: gtggatttgtgagaaatgattygt Reverse: ccattataaccaattacaataccac Nested Forward: gttagttgtttgtggtttattagttg	205	15
MEST	Forward: tygttgttggttagttttgtayggtt Reverse: aaaaataacacccctcctcaaat Nested Reverse: cccaaaaacaaccccaactc	173	21

[1] Due to the single nucleotide polymorphism, this CpG was excluded from downstream analyses

The DNA at this step is in even smaller amounts from the approximately 200 sperm picked due to DNA extraction steps and degradation from bisulfite conversion. A semi-nested PCR amplification procedure was used, which involves two rounds of amplification, where the PCR primers in the first round amplifies a broader region around the target DMRs. The second round involves the use of one of the primers from the first round, in addition to the introduction of a nested primer, which is specific to the DMR of interest. PCR protocol and cycling conditions are described in Tables 4-7. A total of five PCR reactions were setup for each patient sample for each DMR, allowing for the amplification of different sperm as opposed to a single reaction which may be skewed due to amplification bias. Therefore, a typical procedure would include 18 PCR reactions.

Table 5. PCR protocol for first round of semi-nested amplification

Component	Stock Concentration	Final Concentration	Volume to add for single reaction (μL)	Volume to add for negative reaction (μL)
PCR Buffer	10X	1X	2.5	2.5
$MgCl_2$	50mM	1.5mM	0.75	0.75
dNTPs	5mM	0.2mM	1	1
Forward Primer	5uM	0.5uM	2.5	2.5
Reverse Primer	5uM	0.5uM	2.5	2.5
DNA template	--	--	0.5	0
Taq	5U	0.5U	0.1	0.1
Distilled H_2O	--	--	15.15	15.65
Total			25	25

Table 6. Cycling conditions for both rounds of semi-nest PCR amplification

Temperature (Celsius)	Time (minutes)
For 1 cycle:	
94	05:00
For 35 cycles:	
94	00:45
59	00:45
72	00:60
For 1 cycle:	
72	10:00

Table 7. PCR protocol for 2nd round of semi-nested amplification

Component	Stock Concentration	Final Concentration	Volume to add for single reaction (µL)	Volume to add for negative reaction (µL)
PCR Buffer	10X	1X	2.5	2.5
$MgCl_2$	50mM	1.5mM	0.75	0.75
dNTPs	5mM	0.2mM	1	1
Forward Primer	5uM	0.5uM	2.5	2.5
Reverse Primer	5uM	0.5uM	2.5	2.5
DNA template	--	--	2	0
Taq	--	--	0.1	0.1
Distilled H_2O	--	--	13.65	15.65
Total			25	25

2.3.5 Molecular Cloning

The PCR products were verified for successful amplification and size (Table 3) on a 1% agarose gel at 100V for 1.5 hours. Briefly, this protocol used TAE buffer (Tris, acetic acid, EDTA), 3 µL of 100 bp ladder, 3 µL of PCR product, and 4X gel loading dye. Visualization of the bands were achieved using SYBR Safe DNA gel stain (Fisher Scientific, Ottawa, ON, Canada) under a gel trans-illuminator with ultraviolet (UV) light. DNA bands were extracted and purified using the GenElute Gel Extraction Kit (Sigma-Aldrich Canada Ltd, Oakville, ON) following the manufacturer's protocol. Briefly, the remaining PCR product for each reaction was loaded onto a 0.5% agarose TAE gel with 7 µL of gel loading dye and 3 µL of 100bp ladder. DNA separation was achieved by running the gel at 100V for 1.5 hours. Extraction of successful bands from the gel was achieved using a razor under a gel trans-illuminator under UV light.

The gel purified DNA fragments were cloned using the pGEM-T Easy Vector System (Promega, Madison, WI, USA) according to manufacturer's protocol with adjustments. Ligation reactions were setup using 2.5 µL of 2X Rapid Ligation Buffer, 0.5 µL of pGEM-T Easy Vector, 0.5 µL of T4 DNA Ligase, and 50 ng of purified PCR DNA, and distilled H_2O for a total reaction volume of 5 µL. Ligation reactions were incubated at 4°C overnight. JM109 high efficiency competent cells (Promega, Madison, WI, USA) were thawed for 5 minutes, added to the ligation reactions, and allowed to sit on ice for 20 minutes. Transformation of the PCR fragments were achieved by heat shocking the cells at 42°C for 2 minutes. Immediately, 380 µL of Luria-Bertani (LB) broth (Invitrogen Canada Inc, Burlington, ON) were added to the cells and allowed to incubate at 37°C for 1.5 hours with occasional shaking. The samples were then plated on agar (Sigma-Aldrich Canada Ltd, Oakville, ON) plates containing ampicillin, IPTG, and X-Gal for the screening of cell colonies carrying the vector (alive via ampicillin resistance), and with a PCR insert in the vector (white as opposed to blue). Plates were incubated upside-down at 37°C overnight.

2.3.6 Colony PCR

At this step, each plate represents a single PCR reaction and therefore contains at least a single sperm from the studied patient. A total of 2-3 white colonies were selected from each plate resulting in a total of 10 colonies for each gene. A total of 30 colonies for each patient was evaluated for the PCR product via colony PCR following the same semi-nested protocol (Tables 4-7). In parallel, the same colonies were inoculated in 500 µL of LB broth in 1.5 mL centrifuge tubes. Colony PCR products were separated on a 1% agarose TAE gel run at 100V for 1.5 hours and checked for correct DNA bands under UV light. The inoculated colonies corresponding to colony PCR products showing the correct band were further inoculated in 50 mL falcon tubes (Fisher Scientific, Ottawa, ON, Canada) supplemented with 4.5 mL LB broth supplemented with 100 µg/mL ampicillin and incubated overnight at 37°C for a maximum of 16 hours. The entire PCR amplification procedure was repeated until an average of 5 unique colonies were successful for each gene for each patient.

2.3.7 DNA extraction from colonies

Plasmid DNA was extracted and purified from the inoculated colonies using the Qiagen plasmid buffer set (Qiagen, Mississauga, ON, Canada) according to the manufacturer's protocol with adjustments. The reactions were scaled down to about a third of the protocol's volume, i.e. 250 µL of P1, P2, and P3 buffer was instead added. The DNA was precipitated in 800 µL of chilled isopropanol (Fisher Scientific, Ottawa, ON, Canada) and the pellet was washed in 500 µL of 70% ethanol. After drying, the plasmid DNA was resuspended in 30 µL of sterile nuclease free H_2O. The concentration of DNA was measured by spectrophotometry (Eppendorf Canada, Mississauga, ON).

2.3.8 Sequencing

Samples were ready for sequencing when at least five PCR reactions were successfully cloned for each gene for each sample. Complete batches were sent out to McGill University and Génome Québec Innovation Centre (Montreal, QC, Canada). A total of 3-5 µg of DNA in a total of 10 µL volume was submitted for traditional Sanger sequencing using SP6 sequencing primer (5'-tatttaggtgacactatag-3') using the Applied Biosystems 3730xl DNA analyzer. The files containing raw sequence data (FASTA file format) were downloaded from facility’s web application, Nanuq.

2.3.9 Sequencing data processing

FASTA files were aligned manually in Microsoft Excel using search functions, according to the non-modified genomic sequence (Table 3). CpGs were manually checked for DNA methylation status and tabulated for each clone. DNA methylation status of the CpGs were converted to bead diagrams for ease of interpretation using QUantification tool for Methylation Analysis (QUMA) (Kumaki et al., 2008). Only unique clones were displayed, where black beads represented DNA methylation and white empty beads were DNA non-methylated.

2.3.10 *MTHFR* C677T genotyping

Peripheral blood samples were extracted for DNA following the Puregene Gentra blood kit (Qiagen, Mississauga, ON, Canada). DNA concentration of blood samples was determined by spectrophotometry (Eppendorf Canada, Mississauga, ON). Restriction Fragment Length Polymorphism (RFLP) was used to determine the C677T SNP. PCR amplification of *MTHFR* was carried out in 25 µL reactions containing 0.625 U HotStarTaq DNA (Qiagen, Mississauga, ON, Canada), 1 x PCR Buffer (Qiagen, Mississauga, ON, Canada), 0.2 mM dNTPs, 0.5 uM forward (5’-tgaaggagaaggtgtctgcggga-3’) and reverse primers (5’-aggacggtgcggtgagagtg-3’) (Naqvi et al., 2014), and 250-300 ng of DNA. The following cycling conditions were used for amplification: 1 cycle of 95°C for 15 minutes; 44 cycles of 95°C for 45 seconds, 65°C for 1 minute, and 72°C for 1.5 minutes; followed by 1 extension cycle of 72°C for 5 minutes. Restriction digest reactions were carried out in 20 µL reactions containing 17.5 µL of PCR product and 5 U of Hinf1 (New England Biolabs, Ipswich, MA, USA). Digestion reactions were incubated at 37°C overnight and separated on a 3% agarose gel (Invitrogen Canada Inc, Burlington, ON) in 1 x TAE buffer with SYBR Safe and 5 µL of 100bp ladder (Invitrogen Canada Inc, Burlington, ON) run at 140V for 1 hour. PCR bands were visualized using a gel trans-illuminator under UV light and genotypes were determined by size: a single 198 bp band indicated the CC genotype; a 198 bp band, 175 bp band, and 23 bp band indicated the CT genotype; and a 175 bp band and a 23 bp band indicated the TT genotype.

2.3.11 Statistical analysis

The DNA methylation for each clone was analyzed in a binary fashion. A clone was categorized as 'altered' when ≥50% of the CpG sites were incorrectly DNA methylated according to the normal imprinting of the gene, i.e. *H19* and *IG-GTL2* DMRS are paternally methylated therefore should have 100% DNA methylation in the sperm, whereas the *MEST* DMR is maternally methylated and should have 0% DNA methylation in the sperm. A patient's DNA methylation for each DMR was given a binary status as well, where 'altered' for a DMR was set at ≥50% of their clones being altered for that DMR.

The frequency of altered men was compared between groupings (infertility and *MTHFR* genotype) using Fisher's exact two-tailed test, where significance was accepted at P<0.05 after correction for multiple comparisons using the Bonferroni post hoc test. Analysis of variance (ANOVA) test was used to determine age differences between groupings.

2.4 Results

2.4.1 Patient demographics and clone metrics

A total of fifty-three men were investigated in this study: 3 men who were normal in terms of sperm count, 8 who were moderate, 23 who were severe, 10 who were very severe, and 9 fertile control men. A total of 903 clones were analyzed in the final analysis: 320 at the *H19* DMR, 266 *IG-GTL2*, and 317 *MEST*. On average, 6 unique clones were studied per case at the *H19* DMR, 6 at the *IG-GTL2* DMR, and 6 at the *MEST* DMR. In terms of infertility status, on average 8 unique clones were evaluated per case within the fertile controls, 6 within the normal subgroup, 4 within the moderate subgroup, 6 within the severe subgroup, and 5 within the very severe subgroup. In some cases, multiple PCR amplification procedures (>5 reactions) were conducted due to failed reactions to produce a PCR product or to produce a cloned target. The likely explanation is due to limited amounts of starting DNA due to degradation or loss during DNA extraction from sperm. Overall, a comparable number of unique clones were investigated across all DMRs and infertility subgroups, with a slight decrease among the more severe oligozoospermic men. Furthermore, the mean ages between subgroups were comparable (*P*<0.05) (Table 8).

Table 8. Age statistics of subgroups

Subgroup	Mean Age (years)	Standard Deviation (±years)	*P*-value [d]
Fertile [a]	34.1	2.4	-
Normal [b]	N/A	N/A	ns
Moderate	32.9	5.7	
Severe	35.7	5.7	
Very Severe	32.9	1.7	
CC [c]	33.6	4.8	ns
CT	35.6	5.0	
TT	33.5	2.9	

[a] Fertile group consists of proven fertile men with normal semen parameters

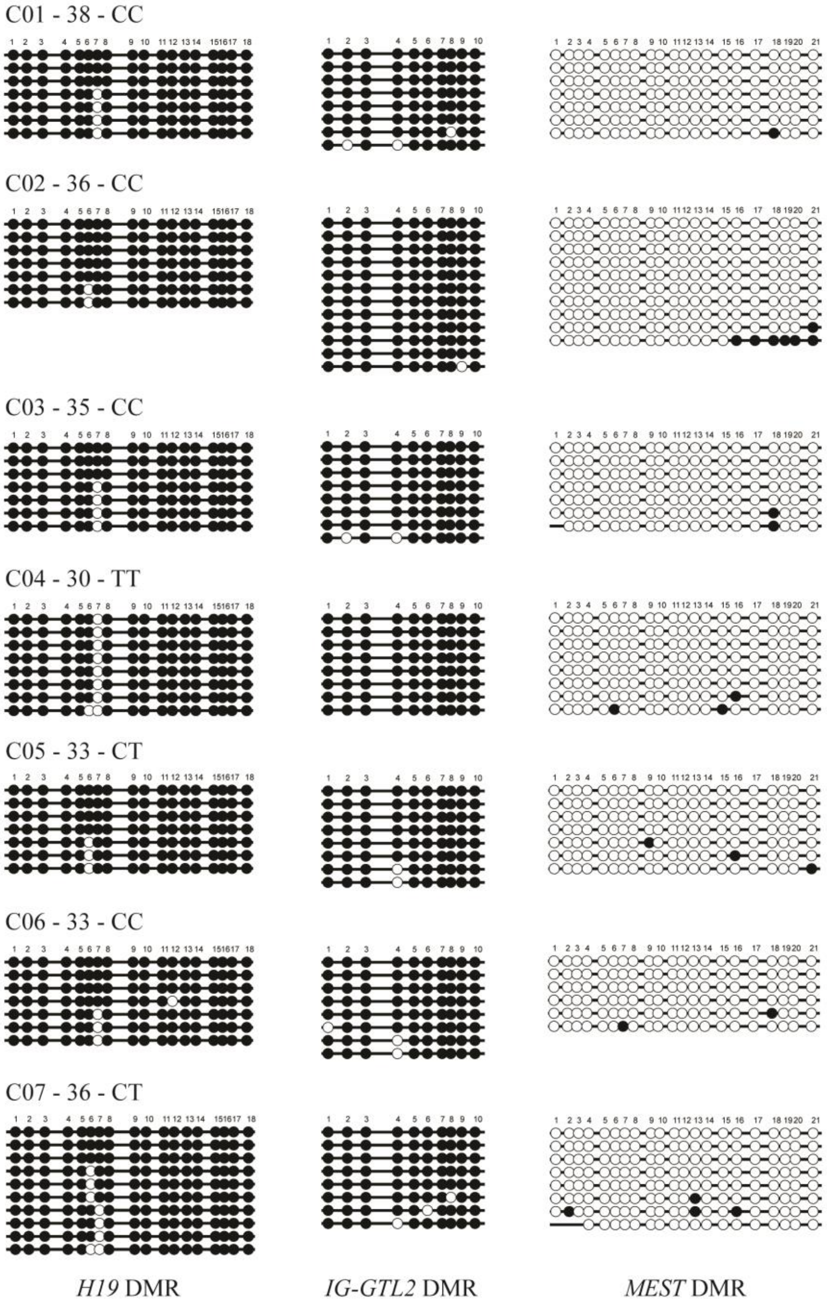
[b] Infertile men with a sperm concentration of >15 million sperm/mL (normal), 5-15 million sperm/mL (moderate), 1-5 million sperm/mL (severe), and <1 million sperm/mL (very severe)

[c] *MTHFR* C677T genotype status
[d] ANOVA, where significance is P<0.05

2.4.2 Methylation analysis at imprinted genes in association with oligozoospermia severity

A total of eleven men (21% of the cohort) were found to have at least 1 altered clone at any DMR (P09, P13, P14, P15, P19, P20, P21, P22, P23, P33, P39) (Figure 6). A total of 9 men were found to have at least 1 altered *H19* DMR clone, where 8 belonged to the severe subgroup (36% of subgroup) and 1 to the very severe subgroup (10% of subgroup). A total of 4 severe (17% of subgroup) and 1 very severe (10% of subgroup) were found to have at least 1 altered *MEST* clone. There were no men with *IG-GTL2* DMR altered clones.

Fertile Control Men



C01 - 38 - CC

C02 - 36 - CC

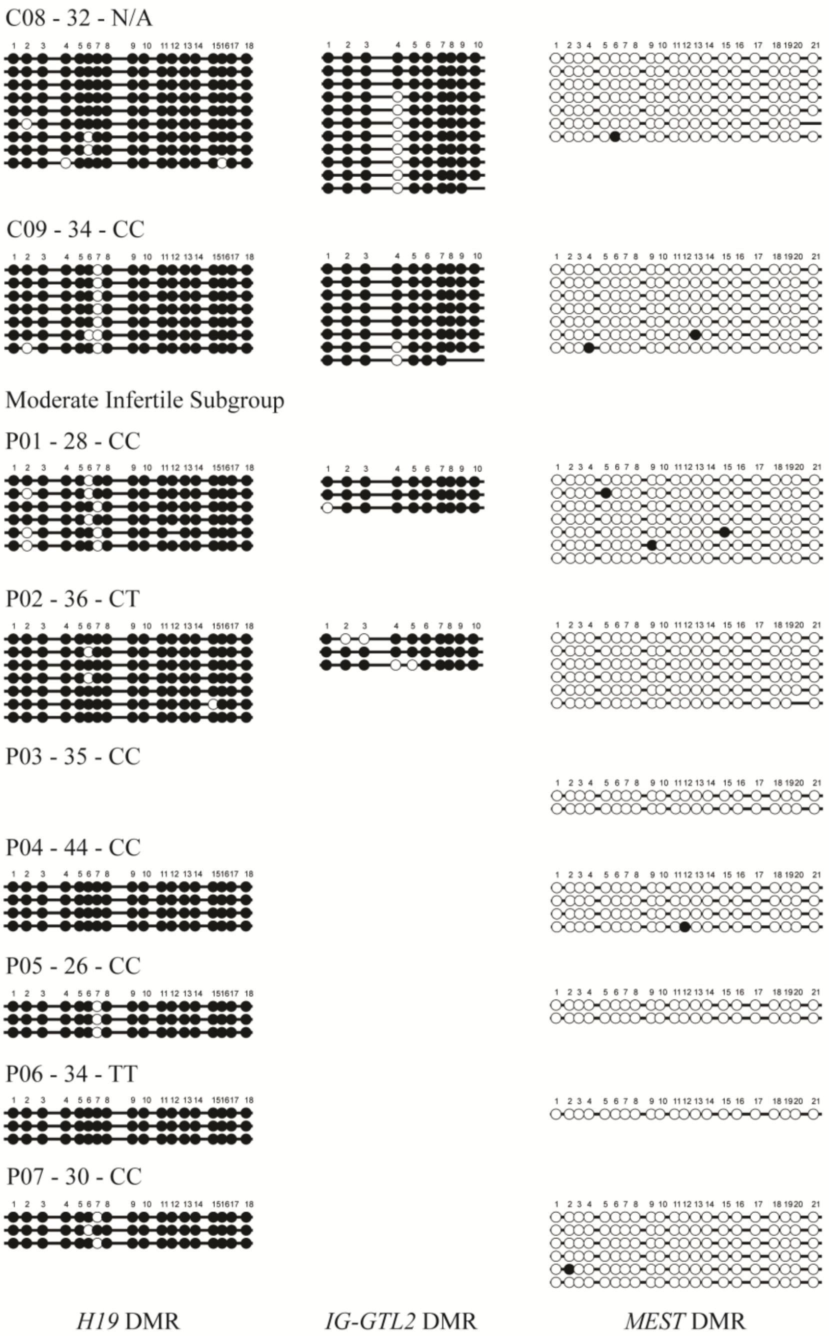
C03 - 35 - CC

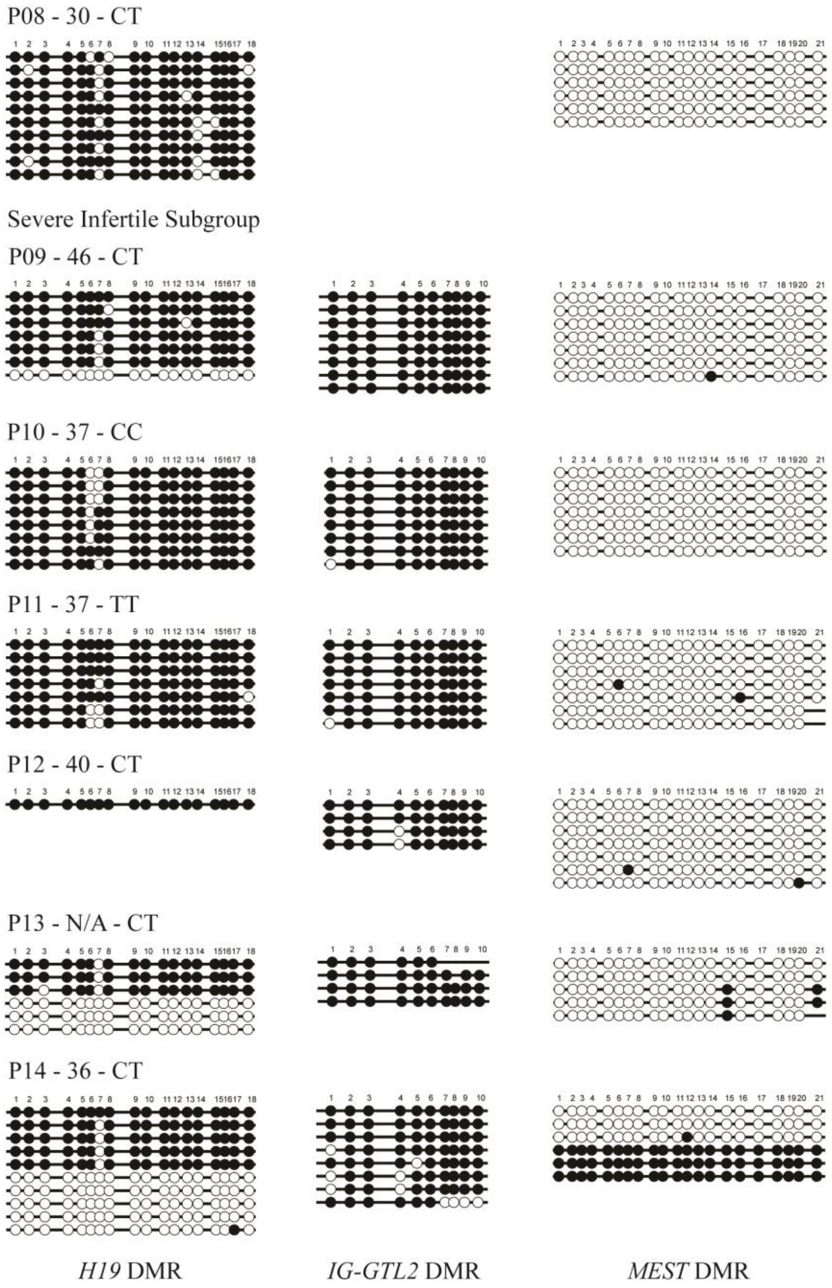
C04 - 30 - TT

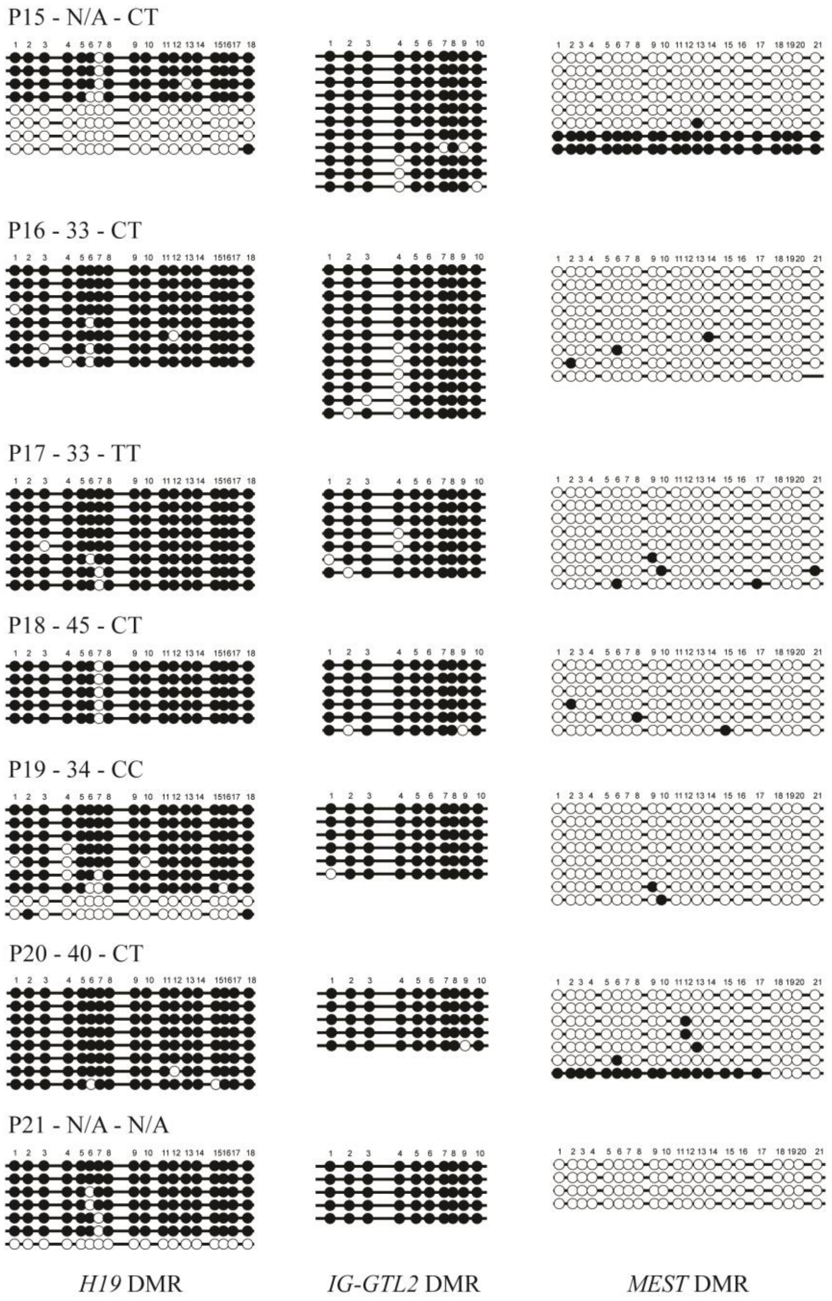
C05 - 33 - CT

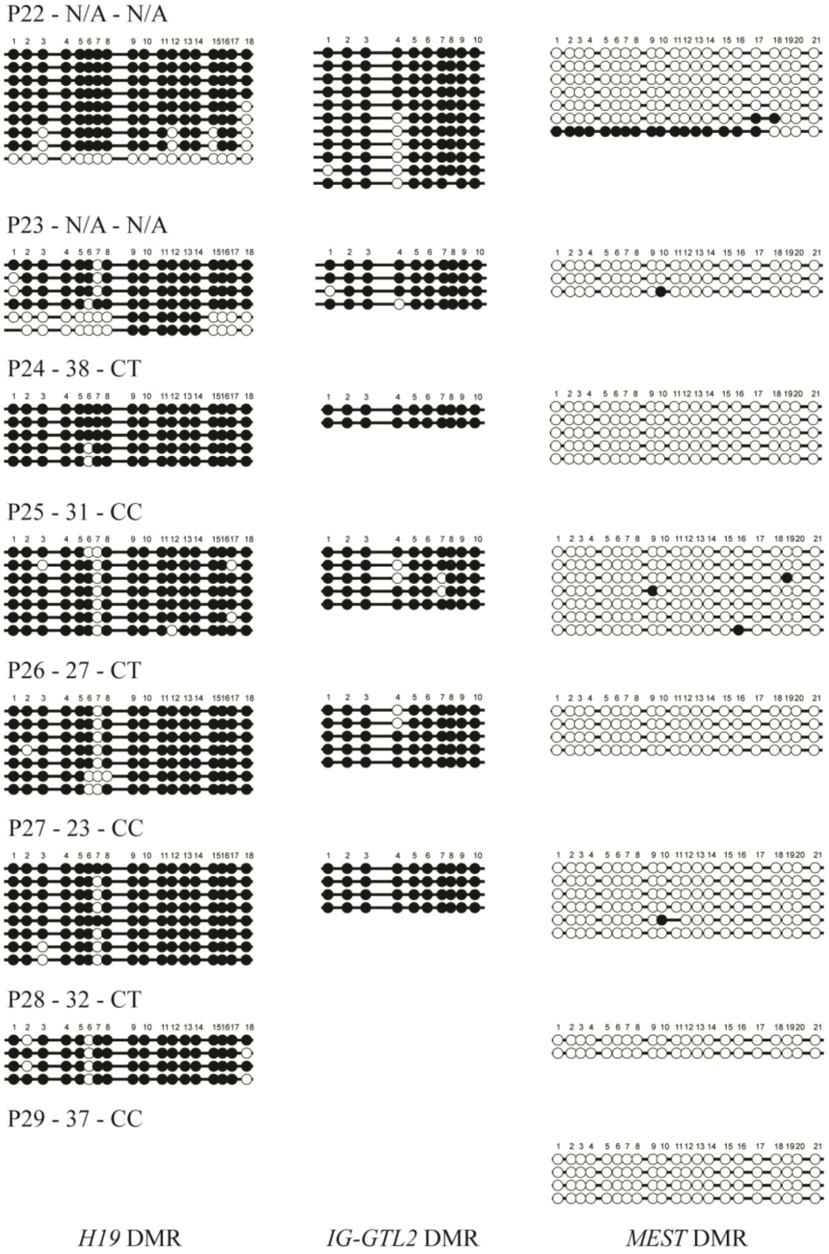
C06 - 33 - CC

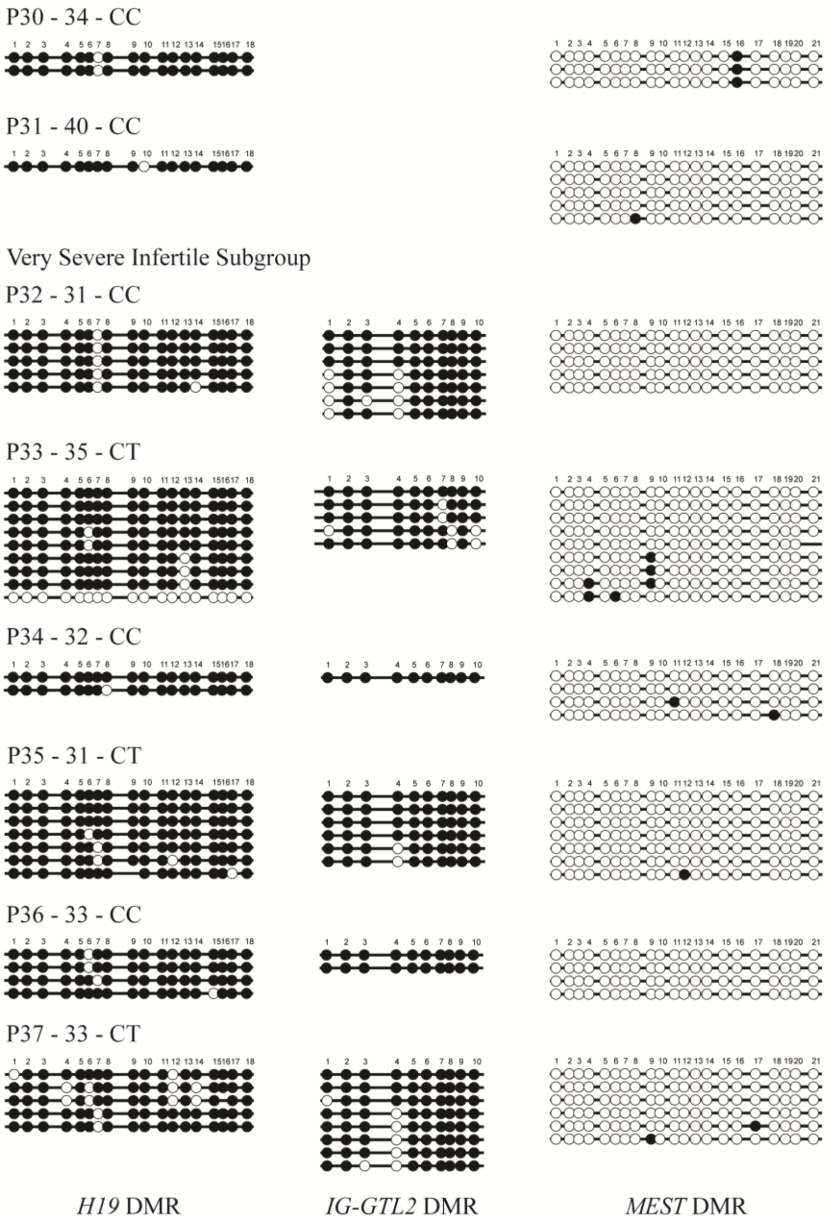
C07 - 36 - CT


C08 - 32 - N/A
C09 - 34 - CC
Moderate Infertile Subgroup
P01 - 28 - CC
P02 - 36 - CT
P03 - 35 - CC
P04 - 44 - CC
P05 - 26 - CC
P06 - 34 - TT
P07 - 30 - CC
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
H19 DMR
IG-GTL2 DMR
MEST DMR


P08 - 30 - CT
Severe Infertile Subgroup
P09 - 46 - CT
P10 - 37 - CC
P11 - 37 - TT
P12 - 40 - CT
P13 - N/A - CT
P14 - 36 - CT
H19 DMR
IG-GTL2 DMR
MEST DMR


P15 - N/A - CT
P16 - 33 - CT
P17 - 33 - TT
P18 - 45 - CT
P19 - 34 - CC
P20 - 40 - CT
P21 - N/A - N/A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
H19 DMR
IG-GTL2 DMR
MEST DMR


P22 - N/A - N/A
P23 - N/A - N/A
P24 - 38 - CT
P25 - 31 - CC
P26 - 27 - CT
P27 - 23 - CC
P28 - 32 - CT
P29 - 37 - CC
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
H19 DMR
IG-GTL2 DMR
MEST DMR


P30 - 34 - CC
P31 - 40 - CC
Very Severe Infertile Subgroup
P32 - 31 - CC
P33 - 35 - CT
P34 - 32 - CC
P35 - 31 - CT
P36 - 33 - CC
P37 - 33 - CT
H19 DMR
IG-GTL2 DMR
MEST DMR

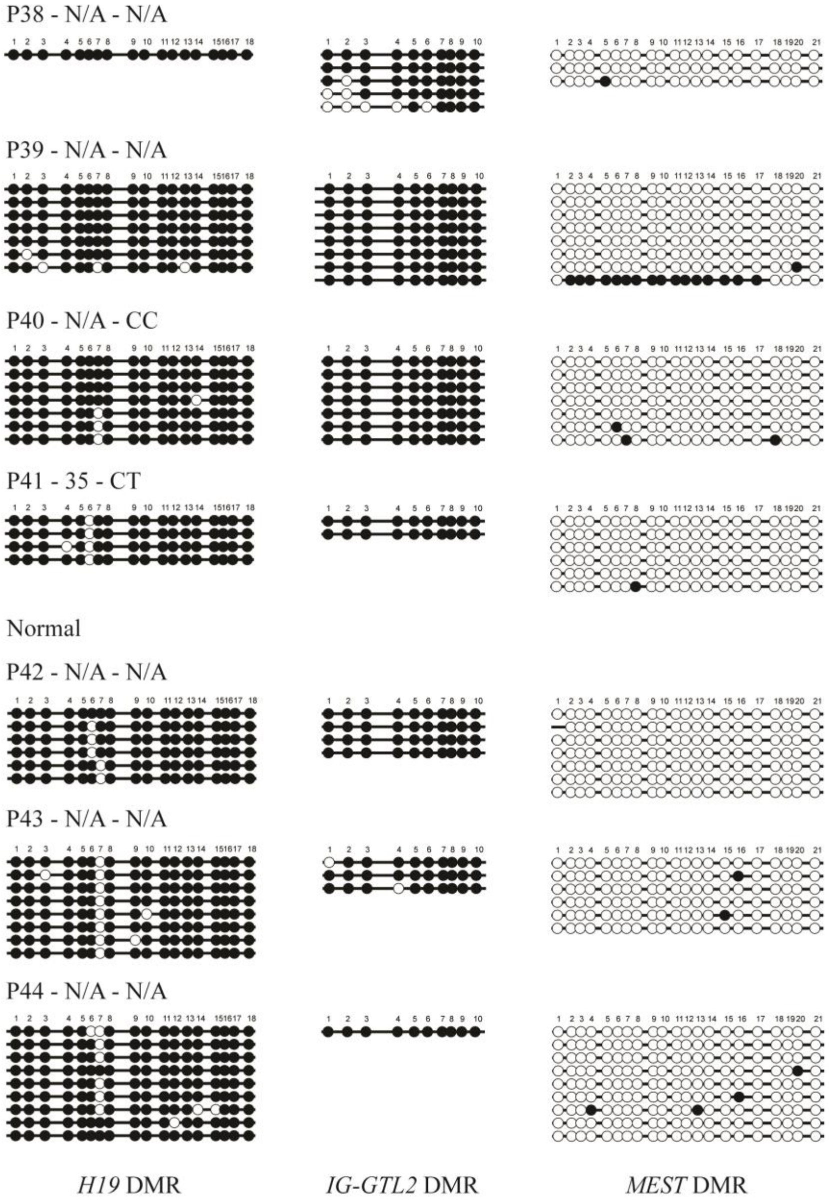

P38 - N/A - N/A
P39 - N/A - N/A
P40 - N/A - CC
P41 - 35 - CT
Normal
P42 - N/A - N/A
P43 - N/A - N/A
P44 - N/A - N/A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
H19 DMR
IG-GTL2 DMR
MEST DMR

Figure 6. Bead-on-a-string diagrams representing sperm clones from infertile and fertile men. Every string represents a unique clone investigated for the individual. Beads represent the CpG sites from the sequence analyzed. Dark beads represent methylated CpGs while, white beads are non-methylated. Missing beads are CpG sites that were not measured. Spaces represent proportional distances between CpG sites along the sequence analyzed. Numbers above the beads indicate the CpG number. Fertile control men consist of proven fertile men with normal semen parameters; moderate infertile subgroup consists of men with a sperm concentration of 5-15 million sperm/mL; severe consists of 1-5 million sperm/mL; very severe consist of <1 million sperm/mL; and normal consists of >15 million sperm/mL. The male's age and MTHFR C677T SNP genotype are reported above each diagram.

Based on the clone analysis, a total of 3 men were classified as being an altered case (P13, P14, P15) (Figure 6). All 3 of these men carried at least 50% of their clones studied at the *H19* DMR to be altered (3/6, 5/10, 4/8 clones, respectively). A total of 1 male (P14) was also found to be altered based on clone analysis at the *MEST* DMR (3/6 clones). All 3 men belonged to the severe subgroup (13% of subgroup); altered men were not found in the normal (0/3), moderate (0/8), very severe (0/10) infertile subgroups, and fertile controls (0/9). The incidence of altered status in severe oligozoospermic men was not significantly higher as compared to other subgroups (*P*=0.54).

2.4.3 Methylation analysis at imprinted genes in association with *MTHFR* C677T SNP genotype

Of the 53 men who donated a semen sample for DNA methylation evaluation, peripheral blood was collected and investigated for the *MTHFR* C677T genotype from 44 men (Table 9). Of these men, 5 CT genotype (26% of subgroup; P09, P13, P14, P15, P33) and 1 CC genotype (5% of subgroup; P19) were found to have altered *H19* clones (1/8, 3/6, 5/10, 4/8, 1/9, and 2/9 clones, respectively) (Figure 6). A total of 3 CT genotype men (9% of subgroup; P14, P15, P20) were found to have altered *MEST* clones (3/6, 2/8, and 1/7 clones, respectively). A total of 2 men

(P14, P15) were found to have altered clones at both the *H19* and *MEST* DMRs, both of which were CT in genotype. There were no altered clones identified among the TT genotype men.

Table 9. *MTHFR* C677T genotypes stratified by infertility subgroup

Subgroup	CC	CT	TT	Total
Fertile [a]	5	2	1	8
Normal [b]	0	0	0	0
Moderate	5	2	1	8
Severe	7	11	2	20
Very Severe	4	4	0	8
Total	21	19	4	44

[a] Fertile group consists of proven fertile men with normal semen parameters
[b] Infertile men with a sperm concentration of >15 million sperm/mL (normal), 5-15 million sperm/mL (moderate), 1-5 million sperm/mL (severe), and <1 million sperm/mL (very severe)

From the previous section, a total of 3 men (P13, P14, P15) met the threshold to be categorized as altered, i.e. have ≥50% of their clones altered. These 3 men were identified to have CT genotypes at the *MTHFR* C677T SNP. This incidence was not, however, significantly higher than CC genotype men (*P*=0.098). When combining both infertility and genotype subgroupings, the incidence of altered men within the severe oligozoospermic subgroup and the CT genotype was significantly higher than non-severe and/or non-CT (*P*=0.012). This significance, however, was lost after correction for multiple group comparisons.

2.5 Discussion

In this study, we investigated the DNA methylation at imprinted DMRs in the sperm of infertile men, stratified by sperm concentration, and fertile control men. We did not find a significant increase in the incidence of altered DNA methylation in the sperm between the different subgroups of oligozoospermic men nor with fertile control men. We also investigated whether the *MTHFR* C677T SNP was associated with the incidence of altered DNA methylation via the genotyping of the men in this study. We also did not find a significant increase in the incidence of altered DNA methylation in the sperm between different genotype men. However, we did find a significant increase of alteration within the subset of men who had severe oligozoospermia (1-5 million sperm/mL) and the CT genotype. The significance was lost after correction for multiple correction; however, this trend may be worthwhile to further investigate with larger sample sizes.

The strengths of our methods in determining the methylation in the sperm includes the investigation of single sperm as represented as unique clones – each of which were derived from different PCR amplification reactions. Although the methods were tedious, including more than 10 PCR amplification reactions per gene per patient, mostly due to failed reactions due to the limited DNA in the samples derived from only 200 sperm, this approach allowed for reduced amplification bias and thus reducing the probability of misrepresenting the patient’s true sperm distribution in their samples. On top of the meticulous and tedious process of picking sperm individually for most samples to avoid contamination of somatic cells, the primers selected in this study to amplify out the DMR regions allowed for the reduced bias in preferentially amplifying one parental allele. Overall, the clone analysis of methylation in this study provides a fair representation of both parental alleles and, with high confidence, from different single sperm.

Previous studies have identified increasing severity of oligozoospermia with methylation alterations (Marques et al., 2008; Boissonnas et al., 2010). The discrepancy between the previous studies and ours may be due to our stringent criteria used in defining an “altered” clone or case.

This in fact is a strength of our study in that we applied a more biologically relevant threshold for clones in that it is more likely methylation will have real biological effects over a greater region of CpGs as opposed to a single CpG site. An interesting trend that we noticed, however, was that altered methylation in the clones was found only in men with <5 million sperm/mL. While we expected the most alterations to be in the very severe group, we noticed that the majority were within the severe, i.e. 1-5 million sperm/mL. This finding is consistent with one other study that when they studied the *H19* DMR in the sperm, severe oligozoospermic men had the highest proportion of clones with alterations (Marques et al., 2008). While we acknowledge that there were fewer cases studied in the very severe group and that they were much harder in the picking process for sperm thus possibly biasing in studying on cases with better sperm production, it is plausible that the severe and very severe cases have different mechanisms of pathophysiology. The severe oligozoospermic men may have etiologies related to methylation while the very severe cases do not. The current analysis is limited in power to make strong conclusions due to the overall small sample sizes, which is unfortunately a result of the difficulty in patient recruitment and sperm extraction from very low sperm count samples. Furthermore, the age of the severe group was higher than the very severe group therefore confounding the results; the current power analysis is limited to about 10% in detection for differences in age between the two groups. Recent studies have identified correlations between age and methylation (Atsem et al., 2016). Future studies should verify these findings in larger sample sizes matched for age.

The *MTHFR* C677T genotype was evaluated in this study to potentially provide a mechanism for the frequent findings of altered methylation in the sperm of oligozoospermic men (Marques et al., 2008; Laurentino et al., 2015). In the current analysis, we did not observe that the SNP was associated with a higher risk for methylation alterations in the sperm. This may, however, be due to the limited number of cases investigated. A typical genetic association study has 10-fold more cases due to the relatively low frequencies of SNPs. Given that we identified a trend for higher T alleles among the altered cases, it is still possible to consider that the T allele may be associated with methylation alterations in the sperm. This is further confirmed by the finding that the majority of alterations were within the subset of infertile men, the severe subgroup, that also had the CT genotype. This might be expanded to suggest that infertile men with lower sperm count (1-5 million sperm/mL) who also have a T allele at the *MTHFR* C677T SNP may be at a higher risk for alterations. This suggests that the cause of infertility in these men may be further exacerbated by the *MTHFR* SNP, resulting in methylation alteration in the sperm. Larger sample size studies are needed to verify this finding.

Two previous studies also evaluated this SNP in relation to methylation in the sperm in infertile men (Camprubí et al., 2013; Arabi et al., 2015). Both studies also did not identify significant differences in methylation at imprinted DMRs in the sperm of infertile men between the *MTHFR* C677T genotypes. The first study by Camprubí et al. investigated a cohort of normozoospermic idiopathic infertile men. This is consistent with our results as our normal subgroup did not find any alterations regardless of genotype. The second study investigated a heterogenous group of infertile men which makes the interpretation of such results difficult as we narrowed our patients to oligozoospermic men. Nevertheless, the results are comparable with ours but may still be affected by the same limitation of sample size in our study.

There are considerations for the results in this study. Previous epidemiological studies identified that the *MTHFR* C677T SNP was not associated with male infertility in all ethnic populations, where only certain populations were affected (Gupta et al., 2011; Naqvi et al., 2014; Gong et al., 2015). As mentioned earlier, alterations during the window of epigenetic reprogramming during gametogenesis may sustain permeant changes that affect spermatogenesis for the lineage of germ cells. Therefore, any alterations during pregnancy incurred by the mother including her folate diet may affect the son’s fertility. It is even plausible that the mother’s *MTHFR* C677T SNP may increase the risk for altered methylation in the germ cells. Collection of such data may improve the interpretation of the results.

A shortcoming in our analysis is the limited number of clones studied per patient, thereby, reducing the power of the analyses. However, we did penalize our findings by setting minimum technical and biological thresholds in clone categorization to limit our type I errors. Our thresholds are quite conservative given that we did not deem a single altered clone finding in a male as an “altered” case, but this in fact may be significant in the clinical setting given that it only takes a single altered sperm during ART for the offspring to inherit an imprinting disorder. Thus, our study may be underreporting the incidence of infertile men with alterations. To improve the power of this study, deeper sequencing techniques are needed (Laurentino et al., 2015). Furthermore, the patient subgroups were unbalanced, with fewer men studied in the severe and very severe subgroups, primarily due to the difficulties in recruitment and isolation of sperm from these limited samples. Lastly, we were unable to collect blood from all cases thus reducing our power to investigate the SNP in this study.

Overall, through our analysis using bisulfite sequencing of clones, we analyzed single sperm at multiple CpG sites at base pair resolution and found that with a stringent criterion for defining an altered case, we found no association between the *MTHFR* C677T SNP and the risk for altered methylation in sperm. However, it is important to acknowledge the limited power in this study due to sample size. To that point, we found a trend in that men with 1-5 million sperm/mL who also had the SNP was more likely to have altered methylation in their sperm. Further studies are needed to confirm this potential synergistic effect.

CHAPTER 3: INVESTIGATION OF GENOME-WIDE METHYLATION IN THE TESTES OF SEVERE INFERTILITY MEN

3.1 Background

Azoospermia is the absence of sperm from at least two centrifugated total ejaculate semen analyses within 3 months (WHO). The prevalence of azoospermia is 1% in the general population and 10-20% among infertile men (Bhasin et al., 1994). The main types of azoospermia include OA and NOA. The pathology of OA arises from an obstruction at some position along the urogenital tract. About 40% of all azoospermic cases are OA. Common causes of OA include inflammation, vasectomy, or CBAVD. Despite the lack of sperm in the semen, these men usually have normal spermatogenesis in their testis. In contrast, NOA men corresponds to failure or dysfunction of spermatogenesis. Known causes of NOA include Klinefelter's syndrome, Y chromosome microdeletions, and chemotherapy. However, up to 18% of NOA men are idiopathic, i.e. the cause of spermatogenic dysfunction in these men are not known (Punab et al., 2016).

Spermatogenesis in the post-pubertal male requires the coordination of testicular gene expression during the sequential phases of mitosis, meiosis, and spermiogenesis (Eddy, 2002). The mitotic maintenance of the pool of spermatogonia or the progression of germ cells from quiescent progenitor cells into meiosis are all required for the proper completion of mature sperm. The genes expressed are male germ cell specific and regulated via a highly conserved genetic program. Therefore, the regulation of this highly specific gene expression program is also vital for the progression of germ cells to become mature sperm. One mechanism of gene expression regulation is methylation, an epigenetic modification, involving the stable and heritable addition of methyl groups to the 5' position of cytosines. Methylation has been shown to be a crucial part of spermatogenesis in the adult, but also playing a role during PGC development in the fetus.

The formation of gametes begins during fetal development and as early as day seventeen post fertilization. A hallmark of PGC development is the epigenome reprogramming in these cells. The function of this epigenetic program is not entirely understood; however, it is suggested to reset the inherited epigenetic marks, e.g. methylation, so as to give rise to germ cells that are sex and individual specific. In fact, a unique and persistent expression program dictates this reprogramming event, and is seen across different mammals, suggesting the importance of this event. The reprogramming, which includes erasure and remethylation, is thought to continue after birth and during the adult life (Kerjean et al., 2000). The formation of mature sperm thus requires specific methylation marks until it is finally through meiosis and packaged into a transcriptionally silenced cell capable of fertilization and passing on such marks onto the next generation. The loss of these methylation marks may play a role in male infertility as it may potentially disrupt the normal gene expression program. Indeed, the exposure of 5-aza-2'-deoxycytidine to adult male mice results in dose-dependent decreases in testicular weight, altered testicular histology, reduced sperm production, and general infertility. (Kelly et al., 2003). 5-aza-2'-deoxycytidine is a cytosine analogue which incorporates into the genome in replacement of cytosine and binds DNMTS covalently resulting in hypomethylation of the genome. Therefore, altered methylation may be a cause of spermatogenic failure; however, few studies have investigated the methylation in the testis of infertile men.

Methylation investigations in azoospermia have suggested epigenetic misregulations at testis specific genes as a potential etiology for NOA (Khazamipour et al., 2009; Ferfouri et al., 2013; Ramasamy et al., 2014). However, these studies can be improved by the investigation of a broader region of the genome and in conjunction with multivariate and non-parametric analyses. The Illumina Infinium methylation platform is a microarray capable of investigating the methylation at >485000 CpG locations across the genome with a spread across various genomic elements including promoters, enhancers, gene bodies, and the open genomic sea. However, the analysis of this vast amount of data has proven difficult (Wilhelm-Benartzi et al., 2013). Majority of studies apply regression models for the identification of DMPs. However, other studies have shown that CpG sites are not independent from nearby sites thus the assumptions associated with linear models may not be entirely valid. Furthermore, methylation of CpG sites may not be linearly related to the disease studied. Due to the high cost of such arrays, many studies have much fewer samples as there are genes studied, resulting in numerous biases. Thus, most studies perform a genome-wide DNA methylation analysis as a screening to narrow down genes to validate further via other methods (Laqqan et al., 2017). In other research fields such as cancer (Pirooznia et al., 2008; Saeys et al., 2007; Guyon et al., 2002; Statnikov et al., 2008), ML algorithms such as SVMs and decision tree based models, e.g. RF, are applied aiming to identify a set of genes indicative of disease. These methods provide a non-parametric and empirical approaches to data analysis that may supplement or improve on the results from large datasets.

The results of such methods have been successful in accurately classifying patients of specific tumour types and identifying a small subset of genes capable of such classification.

3.2 Rationale

Spermatogenic failure in NOA men presents a human model for the investigation of spermatogenesis. Given that spermatogenic dysfunction is a common trait among infertile men across various severities, it is plausible that the findings from studying azoospermic men may be applicable to other types of infertile men. Since methylation is the regulation of gene expression and that spermatogenesis requires a highly specific genetic program in both somatic and germ cells, alterations of such methylation marks may result in the disruption of spermatogenesis and thus spermatogenic arrest or failure. This is further supported by the fact that PGC development heavily encompasses a specific methylation program and that the disruption of such program during fetal development may result in acquired and stable alterations also leading to an NOA phenotype. Although majority of studies have focused on the methylation in the sperm partly as a result of less difficulty in obtaining research samples, the evaluation of the sperm may not be highly informative of epigenetically regulated spermatogenesis pathways given that the sperm is the final stage of spermatogenesis and is transcriptionally inactive. Testicular tissue, although much more difficult to obtain for research, would be more informative of actual methylation regulated pathways linked to the active transcription of genes in the germ or somatic cells.

The aim of this study was to determine whether there are any differences in testicular methylation between NOA, OA, and vasectomy reversal (VR) fertile control men. The primary objective was to evaluate the methylation in a genome-wide approach in testicular seminiferous tissue of these men. The secondary objective, and more importantly, was that if there were differences in methylation, to identify a set of genes that are differentially methylated between the fertility groups. The primary outcomes are anticipated to identify pathways that may provide insight into the epigenetically regulated mechanisms driving azoospermia. The value of this work would be to provide biomarkers and diagnostic techniques that may be easily translated to the clinical setting. Given the high adoption of ML algorithms for feature selection in other fields of research yet none to our knowledge has been applied to the fields of epigenetics in infertility, we presented results from a classic linear model in addition to the implementation of ML algorithms.

3.3 Methods

3.3.1 Patient recruitment

Patient recruitment practices were the same as described in 2.3.1. Testicular biopsies were collected from azoospermic men undergoing pathological assessment or sperm retrieval for ART, and from control fertile men undergoing VR. VR fertile controls were included only if they had at least one child prior to vasectomy. Clinical information was obtained through patient charts provided by the urologist. Medical charts were not available for VR cases. Ethical approval was obtained from the University of British Columbia Ethics Committee prior to commencing study.

3.3.2 Sample processing

On the day of surgery and from the same biopsy tissue used for ART, a piece was cut into a mHTF in a 15 mL falcon tube (Fisher Scientific, Ottawa, ON, Canada) and immediately incubated in a 37°C water bath for this study. Within 1 hour, the sample was transferred into a 37°C incubator briefly until preparation. The biopsy was transferred to a 60 x 15 mm petri dish filled with mHTF and the testicular biopsy was cut into 3 mm to 5 mm segments. Segmented tubules were incubated at 37°C for 45 min to 60 min in freshly prepared hypo-extraction buffer [30 mM Tris, 50 mM sucrose, 17 mM citric acid, 5 mM EDTA, 0.5 mM DTT, and 0.1 mM phenylmethylsulphonyl fluoride (PMSF); pH 8.4]. Minced biopsy tissue was stored in a 0.5 mL centrifuge tube at -20°C until DNA extraction.

3.3.3 DNA extraction

Samples were thawed on ice and using a 100 µL pipette tip, the tissue fibers were manually transferred to a 1.5 mL centrifuge tube. The extraction of DNA followed the Puregene Gentra Tissue kit (Qiagen, Mississauga, ON, Canada) according to manufacturer's protocol. DNA concentration and quality was evaluated on the Nanodrop spectrophotometer (Thermo Fisher).

3.3.4 Microarray

DNA samples were sent to microarray processing facilities for quality control assurance and bisulfite conversion via the EZ DNA Methylation Gold Kit (Zymo Research, Orange, CA). A minimum of 500 ng in 15 µL of the extracted testicular tissue DNA was sent to either the McGill University and Génome Québec Innovation Centre (Montreal, QC, Canada) or The Centre for Applied Genomics (TCAG) at the Sickkids Hospital (Toronto, ON, Canada). The first batch was sent to McGill with 12 samples (9 NOA, 2 VR, and 1 OA), while the second batch sent to TCAG one year later with also 12 samples (1 NOA, 6 VR, 5 OA). For both batches, the

Illumina Human Methylation450 BeadChip was used for the detection and evaluation of methylation (see section 1.4.2 for more details).

3.3.5 Data quality control and processing

Raw IDAT files were imported into R using the minfi package (version 1.22.1; Aryee et al., 2014). Samples were normalized with the subset quantile normalization, i.e. stratified by regions e.g. CpG island, shore, etc., for the correction of background adjustments and differences in probe types and dyes (Aryee et al., 2014). Microarray annotations and genomic feature data was imported from the R Bioconductor library packages (IlluminaHumanMethylation450kmanifest 0.4.0; Hansen and Aryee, 2012; IlluminaHumanMethylation450kanno.ilmn12.hg19 0.6.0, Hansen 2016).

3.3.6 Identification of differentially methylated positions and regions

In addition to regression methods such as linear modelling, ML algorithms were used for the identification of DMPs. As opposed to traditional statistics, feature selection by ML algorithms are trained in a supervised manner. These methods apply a cost reduction function in the form of an accuracy score from the prediction of cases, i.e. classifying cases as NOA, OA, or VR. This score is checked and iteratively optimized via the continual removal of genes until the highest score is achieved. These methods are termed wrapper methods or when in combination with a filtering method (i.e. with a t statistic), are called embedded methods. The resulting final gene set with the highest score in predicting cases was considered to harbour genes linked to azoospermia.

3.3.6.1 Empirical Bayesian moderated t-test

A statistical regression model was used for the identification of DMPs as tested previously (Ramasamy et al., 2014; Ferfouri et al., 2013) using the limma R package (3.32.10; Ritchie et al., 2015). This method estimates the differences in methylation with standard errors by fitting a linear model for each gene with smoothening to the standard errors under empirical Bayesian statistics (eBayes). A DMP was deemed significance if the *P*-value is below a false discovery rate (FDR) of 5%.

3.3.6.2 PCA filtering for gene selection

A modified algorithm based on PCA was used for the selection of genes with high variance associated with clinical covariates. This filtering algorithm was partially adopted from (Roden et al., 2006). Singular value decomposition (SVD) of the covariance matrix of the raw methylation M-values (*n* samples x *m* probes) was used to determine the PCs and the transformed PC scores. Each PC score was iteratively tested for significant association with known clinical covariates (Kruskal-Wallis for categorical data and Spearman's rank correlation for numerical data; $P<0.05$). For each PC found to be significantly associated with a clinical covariate, the top probes were selected as genes associated with that covariate. Top genes were selected based on outliers classified outside of 1.5 x the interquartile range (IQR) of the PC's principal directions or axes. These outliers were considered to contribute the most weighting to the PC's directions.

3.3.6.3 Machine learning embedded methods

Two established and one new embedded method together were used for gene selection. The SVM-RFE method using a linear kernel was applied (Guyon et al., 2002). Instead of a single probe removed per iteration, 1% of probes were removed each time for quicker implementation. The source code for the algorithm based on the original work from Guyon was adapted from Dejong, 2016. The BIRF embedded method was also used for gene selection (Anaissi et al., 2015). A RF was performed using the R package randomForest (4.6-12; Liaw and Wiener, 2002) using a cutoff value used for the compensation of imbalanced classes was equal to the proportion of cases for each group. The sampling parameter selected was half of the cases for each class. We implemented a new embedded method based on the gene sets derived from eBayes (section 3.3.6.1). A novel method was used in this study which was a combination of conducting an eBayes linear model with either an SVM or RF classifier to determine the FDR with the highest accuracy.

3.3.6.4 Machine learning classification validation

Two ML algorithms were used for validation of gene sets in this study: SVM with a Gaussian radial kernel (e1071 R package; 1.6-8; Meyer et al., 2017) and balanced RF using a cutoff parameter equal to the proportional of cases for each class and a sample size parameter equal to half the cases for each class. This approach enables a more balanced approach in the building of decision trees during the RF algorithm. Given the imbalance in the classes, i.e. more NOA cases than OA and VR, without forcing sample size and cutoff parameters, many trees will be built with mostly one class leading to poor decision models. Due to the problem with technical batch in this dataset, we selected “hard or batch” cases. These are the few cases that are of different fertility that were on opposite microarray batches. For example, on the first batch which was predominantly NOA cases, there were also 2 VR and 1 OA case. These last 2 VR and 1 OA cases are considered the hard or batch cases. Correct classification of these cases may indicate a gene set and ML model that has adjusted for batch effects. Therefore, preference for models will be given to those that predict the batch cases correctly. The leave-one-out-cross-validation (LOOCV) was used for the testing of cases. This approach trained a machine learning classifier with N -1 cases and used the excluded case as a test case. This is repeated N times until each case is tested exactly once. The total accuracy score is determined by the number of correct cases / N cases. Permutation testing was further used to test for the significance of the LOOCV above random chance. Test sets (n = 1000) were generated with the case labels randomly shuffled. The proportions of simulations with >= to the highest LOOCV accuracy score was used as the P-value.

3.3.6.5 Bump hunting

The bumphunter algorithm included in the R minfi package is a bump hunting algorithm (Jaffe et al., 2012). Instead of looking for associations at single genomic locations with a phenotype of interest, i.e. DMP, the bump hunting algorithm seeks out genomic regions. The algorithm defines clusters of probes on the 450k array that are not separated more than a set distance. Next, the algorithm identifies a t-statistic at each genomic location or probe followed by defining a candidate region to be a cluster of probes where all the t-statistics exceed a predefined threshold. Significance of the region is determined via permutations or bootstrapping, depending if there are covariates to control for. In this study, the bumphunter function was applied using a threshold difference of beta methylation at 0.20, i.e. a 20% difference between groups, and a bootstrap value of 1000.

3.3.7 Gene enrichment analysis

The RefGene names associated with gene sets were parsed from the Illumina annotation file and were submitted for GO term enrichment analysis (Ashburner et al., 2000; GO Consortium, 2017; Mi et al., 2017) using the PANTHER Overrepresentation Test (Released 20171205) with the GO Ontology database (Released 20171227) via the Gene Ontology Consortium web interface. The determination of expression of candidate genes in testis was cross-checked with the Human Protein Atlas database (Uhlén et al., 2010; Uhlén et al., 2015). The database was accessed from with R using the hpar Bioconductor package (1.18.1; Gatto, 2017).

3.3.8 Statistical analysis

The determination of significant association between gene sets with a specific genomic feature was achieved using a two-tailed Fisher's exact test with Bonferroni correction for multiple comparisons. $P<0.05$ was considered significant. Group comparisons of known clinical covariates between infertility groups were achieved using two-tailed Fisher's exact test, ANOVA, or Kruskal-Wallis – depending on data type and distribution, where significance was chosen at a level of $P<0.05$ after Bonferroni correction for multiple group and covariate comparisons.

3.4 Results

3.4.1 Patient demographics

A total of twenty-four men were recruited for this study: 10 NOA, 6 OA, and 8 VR controls (Table 10). The mean age between the NOA, OA, and VR groups were not significantly different (*P*=0.24). There is, however, a noticeable trend of decreasing age moving from VR, to OA, and finally the lowest mean age in NOA cases. With the limited number of cases with age information, caution should be used when interpreting this statistic as the provided sample size of known cases has <80% power to detect a small effect size with significance (*P*<0.05). The epigenetic ages of the cases were calculated using 353 CpG probes found on the 450k microarray known as the Horvath Clock (Horvath, 2013; Table 10). In accordance with the real age statistic, the mean epigenetic ages were not significantly different between groups (NOA 35.4 ± 4.8 years; OA 34.8 ± 4.0; VR 37.2 ± 3.9; *P*=0.55). The epigenetic age and real age, as expected, show a borderline significant correlation (Pearson's product moment 0.63; *P*=0.05) further providing confidence in the epigenetic age estimations.

Table 10. Patient demographics

		Age	Horvath Clock epigenetic age		LH		FSH			Ethnicity
Fertility Group	Mean ± SD	Total (n)	Mean ± SD	Normal (n)	Borderline (n)	Normal (n)	Borderline (n)	Caucasian (n)	Chinese (n)	Indian (n)
NOA	39.0 ± 5.7	2	35.4 ± 4.8	2	0	1	1[1]	2	0	1
OA	35.3 ± 5.6	4	34.8 ± 4.0	3	1	4	1	3	1	1
VR	46.8 ± 11.6	4	37.2 ± 3.9	-	-	-	-	7	0	0

[1] Case had an elevated FSH level

The LH and FSH hormone levels were abstracted from patient charts. The normal range of LH and FSH are 1-10 mIU/mL and 1.4-18.1 mIU/mL, respectively (Gordetsky et al., 2012). A total of three cases had borderline or elevated hormone levels (Table 10); however, no case had borderline/elevated levels in both FSH and LH. There were no significant differences in LH and FSH levels between the NOA and OA men (*P*=1.0 and *P*=0.52, respectively); however, caution should be used when interpreting these values with the given sample size. There was also no significant difference in the proportion of ethnicities between the three groups (*P*=0.17). A single VR case was found to have varicocele, i.e. enlargement of the veins in the scrotum, and a single OA case was affected by orchitis in the past, i.e. inflammation of the testis.

3.4.2 Microarray preprocessing and quality control

Of the 485,512 probes on the 450k microarray, 703 probes were excluded due to poor detection above background intensity in 10% or more samples. No samples were excluded on the criterion of poor detection of probe intensities in >5% of probes. In addition, no samples were dropped due to poor quality as measured via the built-in 450k control probes, which can be confirmed via the characteristic dual peaks in the beta distributions associated with methylation data (Figure 7). A total of 17,515 additional probes were dropped due to the presence of SNPs in the probe sequence, or in the single base extension sequence, or on the CpG site itself. A further 25,817 were dropped due to previously described non-specific binding (Pidsley et al., 2016). Overall, 441,479 probes in each of 24 samples were retained for further analyses.

3.4.3 Identification of differentially methylated probes between fertility groups

To determine whether there were DMPs between NOA, OA, and VR men, multiple methods were used in the analysis of the 450k data. An overview of the methodological procedure is outlined in Figure 7. An ensemble of methods was used to mine for methylation gene sets associated with azoospermia.

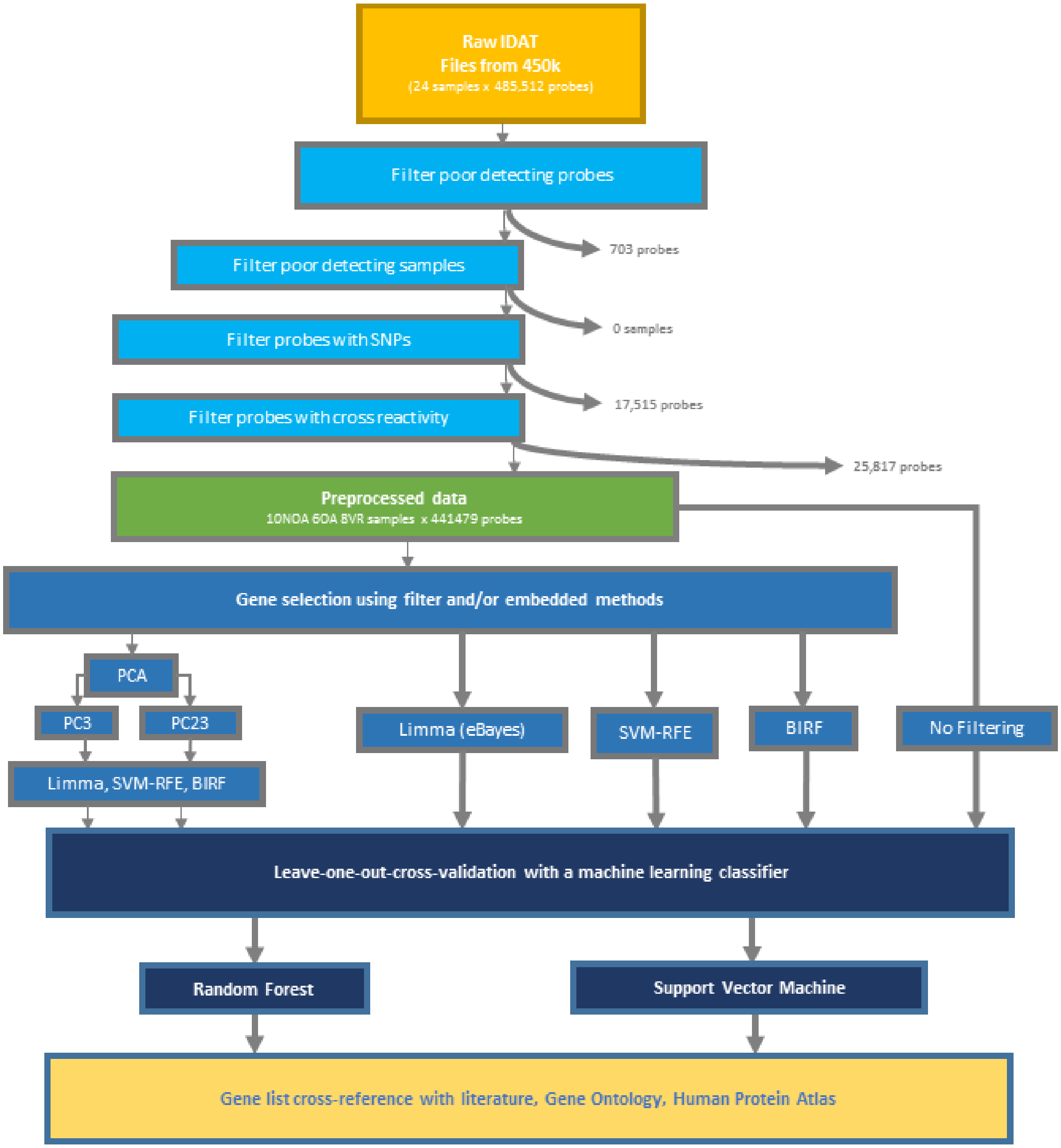

Raw IDAT
Files from 450k
(24 samples x 485,512 probes)
Filter poor detecting probes
703 probes
Filter poor detecting samples
0 samples
Filter probes with SNPs
17,515 probes
Filter probes with cross reactivity
25,817 probes
Preprocessed data
10NOA 6OA 8VR samples x 441479 probes
Gene selection using filter and/or embedded methods
PCA
PC3
PC23
Limma, SVM-RFE, BIRF
Limma (eBayes)
SVM-RFE
BIRF
No Filtering
Leave-one-out-cross-validation with a machine learning classifier
Random Forest
Support Vector Machine
Gene list cross-reference with literature, Gene Ontology, Human Protein Atlas

Figure 7. Analytical pipeline for 450k methylation microarray.
Raw data is first preprocessed and filtered. Gene selection (i.e. selection of DMPs) was achieved using various filtering and embedded methods. All gene selection methods resulted in a gene set. All gene sets were validated using a leave-one-out-cross-validation approach in combination with a machine learning classifier. The gene set with the highest accuracy scores were further evaluated using enrichment analysis via online resources.

A total of 10 gene sets were derived from the 4 feature selection methods (i.e. linear model, PCA, BIRF, SVM-RFE) individually or in combination with more than 1 other method.

PCA identified 2 PCs that were significantly associated with fertility grouping of cases thereby producing 2 gene sets (PC3 accounting for 3.5% of the variance *P*=0.019; PC23 for 1.2% of variance *P*=0.046). Each gene set was used as training data for SVM and RF supervised learning. A LOOCV approach was used to produce accuracy scores for each model (Table 11).

Table 11. Machine learning classification validation of gene sets from filtering and embedded methods

Gene selection method	LOOCV[1] accuracy scores			
	RF[2]		SVM[3]	
	All cases	Batch cases	All cases	Batch cases
No filtering	54.2%	0%	62.5%	0%
PC3[4]	58.3%	0%	62.5%	0%
PC23	66.7%	0%	54.2%	0%
limma[5] 22% FDR	79.2%	75%	83.3%	50%
PC3 + limma 1% FDR	79.2%	75%	83.3%	75%
PC23 + limma	91.7% (42% FDR)	75%	**95.8% (43% FDR)**	**75%**
SVM-RFE[6]	83.3%	25%	70.8%	25%
PC3 + SVM-RFE	70.8%	0%	62.5%	0%
PC23 + SVM-RFE	66.7%	0%	62.5%	0%
BIRF[7]	83.3%	0%	62.5%	0%
PC3 + BIRF	83.3%	50%	66.7%	25%
PC23 + BIRF	**91.7%**	**75%**	79.2%	25%

[1] Leave-one-out-cross-validation, i.e. with *n* samples a ML model is trained using *n* – 1 samples and prediction on the left-out sample. This is repeated *n* times without replacement.

[2] Balanced random forest classifier; cutoff set to 0.42, 0.25, 0.33 for NOA, OA, VR respectively; sampsize set to 5,3,4 for NOA, OA, VR respectively

[3] Support vector machine using a Gaussian radial kernel

[4] Principal component analysis filtered genes selecting the top weighted scores; numbered PC is the PC found associated with fertility

[5] Empirical Bayesian moderated t-test

[6] Support vector machine with recursive feature elimination set with a 5% fraction of feature elimination per iteration

[7] Balanced iterative random forest; cutoff set to 0.42, 0.25, 0.33 for NOA, OA, VR respectively; sampsize set to 5,3,4 for NOA, OA, VR respectively

From a total of 24 validation tests, the highest accuracy achieved was 95.8% (i.e., 23/24 correct classifications) with 75% scoring on the batch cases (i.e., the single misclassified case was a batch case). This was achieved by training an SVM model using the PC23 gene set filtered with an eBayes linear model with an FDR cutoff of 43% (i.e. PC23 + limma + SVM). The single misclassified case was an OA case predicated as VR. This test set achieved a *P*-value = 0.001

after permutation testing. A total of 60 probes were identified in this gene set (Table 12). Of these 60 probes, 28 of which bind to 24 known genes expressed in testicular tubule cells or Leydig cells. The Transmembrane protein 120B (*TMEM120B*) and nuclear distribution protein NudE homolog 1 (*NDE1*) probes identified >10% lower methylation among NOA men as compared to VR men.

Table 12. Significant probes associated with azoospermia as identified via PC23 + limma + SVM model

Probe	Nearest gene	Comparison	Mean ± SD[1] (%) Group 1	Mean ± SD (%) Group 2	Diff (%)[2]	*P*-value[3]	Testis expression[4]
cg18776021	*GPM6B*	NOA vs. VR	19.81 ± 2.74	22.59 ± 2.73	2.79	0.03	N
cg01232331		NOA vs. OA	85.09 ± 2.29	66.21 ± 16.85	-18.88	0.24	-
cg00026909	*DAB1*	NOA vs. VR	64.93 ± 5.42	75.16 ± 2.8	10.23	0.30	N
cg00055603		NOA vs. VR	92.89 ± 0.53	90.38 ± 1.39	-2.52	0.30	-
cg01567615	*RASA3*	NOA vs. VR	92.21 ± 3.39	84.74 ± 5.49	-7.47	0.30	Y
cg01901788	*MAP1LC3A*	NOA vs. VR	41.98 ± 6.11	35.12 ± 4.16	-6.86	0.30	Y
cg01971552		NOA vs. VR	19.25 ± 2.4	23.11 ± 4.05	3.86	0.30	-
cg03651613	*TMEM120B*	NOA vs. VR	14.68 ± 4.2	30.4 ± 7.18	15.72	0.30	Y
cg06083252	*RUFY2*	NOA vs. VR	89.88 ± 1.73	87.08 ± 1.96	-2.80	0.30	Y
cg06727198	*C1orf159*	NOA vs. VR	85.42 ± 4.19	79.41 ± 6.76	-6.01	0.30	Y
cg09662369	*NDE1*	NOA vs. VR	29.22 ± 4.05	41.89 ± 5.47	12.67	0.30	Y
cg16627148		NOA vs. VR	46.97 ± 10.64	33.66 ± 6.56	-13.31	0.30	-
cg17807683	*OR1D2*	NOA vs. VR	92.05 ± 1	90.44 ± 1.24	-1.61	0.30	N
cg19219655	*KLF11*	NOA vs. VR	64.81 ± 7.89	47.58 ± 3.55	-17.22	0.30	N
cg22832407	*CBR4*	NOA vs. VR	30.35 ± 4.46	32.1 ± 7.84	1.75	0.30	Y
cg23717725		NOA vs. VR	85.49 ± 2.21	88.34 ± 1.74	2.85	0.30	-
cg24431515	*ATHL1*	NOA vs. VR	92.04 ± 2.18	79.27 ± 9.53	-12.77	0.30	Y
cg25048701	*FOLR1*	NOA vs. VR	93.85 ± 0.9	92.3 ± 0.95	-1.56	0.30	N
cg27169685	*C1orf159*	NOA vs. VR	85.32 ± 2.52	81.51 ± 6.8	-3.81	0.30	Y
cg07194374	*DDX24*	NOA vs. VR	80.01 ± 4.14	87.42 ± 2.11	7.41	0.35	Y
cg03322334	*WBSCR27*	NOA vs. VR	11.29 ± 1.09	14.92 ± 2.74	3.63	0.36	N
cg03309180	*PRICKLE1*	NOA vs. OA	14.05 ± 3.95	17.88 ± 7.08	3.84	0.37	Y
cg14252222	*WDR37*	NOA vs. OA	86.06 ± 4.55	75 ± 8.19	-11.07	0.37	Y
cg16244374	*WDR37*	NOA vs. OA	91.32 ± 3.23	82.37 ± 7.55	-8.94	0.37	Y
cg17982504	*DDX28*	NOA vs. OA	6.94 ± 1.66	5.73 ± 2.3	-1.22	0.37	Y
cg06655062	*MGAT4B*	NOA vs. VR	87.03 ± 2.38	78.84 ± 7.19	-8.19	0.39	N
cg10472759	*PKIB*	NOA vs. VR	6.3 ± 1.35	11.22 ± 6.06	4.92	0.39	Y
cg25255293		NOA vs. VR	17.59 ± 3.96	20.28 ± 6.29	2.69	0.39	-
cg04361852	*UCP2*	NOA vs. OA	15.44 ± 1.46	20.4 ± 2.59	4.96	0.40	Y
cg22832407	*CBR4*	NOA vs. OA	30.35 ± 4.46	31.38 ± 7.92	1.03	0.40	Y
cg25571060	*PJA1*	NOA vs. OA	12.54 ± 2.38	10.32 ± 0.7	-2.22	0.40	Y
cg27591450		NOA vs. OA	15.76 ± 3.51	9.41 ± 1.99	-6.35	0.40	-
cg00579694		NOA vs. VR	86.04 ± 1.73	80.55 ± 3.35	-5.49	0.41	-
cg01244346	*TET3*	NOA vs. VR	88.71 ± 1.79	90.83 ± 1.58	2.12	0.41	Y
cg01761662	*POU2F3*	NOA vs. VR	10.21 ± 1.68	14.52 ± 2.49	4.31	0.41	N
cg04813240		NOA vs. VR	87.17 ± 4.22	80.83 ± 6.32	-6.34	0.41	-
cg05676450	*PITPNM3*	NOA vs. VR	7.83 ± 1.66	12.32 ± 1.17	4.50	0.41	N

cg06271956	*METT5D1*	NOA vs. VR	10.49 ± 1.16	14.27 ± 1.75	3.79	0.41	N
cg06366585	*RPRD2*	NOA vs. VR	26.63 ± 4.7	33.4 ± 6.33	6.76	0.41	N
cg07662121	*MPV17L*	NOA vs. VR	17.48 ± 5.05	15.51 ± 2.61	-1.97	0.41	Y
cg07813275	*NPL*	NOA vs. VR	78.71 ± 2.43	72.38 ± 4.63	-6.33	0.41	N
cg10682155	*SIM2*	NOA vs. VR	16.79 ± 2.43	13.43 ± 1.33	-3.36	0.41	Y
cg10782633	*C1orf159*	NOA vs. VR	90.39 ± 4.59	84.22 ± 6.76	-6.17	0.41	Y
cg11392877	*PDE8A*	NOA vs. VR	6.38 ± 1.26	9.42 ± 2.49	3.03	0.41	Y
cg13534734		NOA vs. VR	14.91 ± 1.92	19.74 ± 1.82	4.83	0.41	-
cg14813603	*FGF14*	NOA vs. VR	17.67 ± 2.17	22.65 ± 5.25	4.97	0.41	N
cg15527554	*YSK4*	NOA vs. VR	91.02 ± 0.76	85.81 ± 8.23	-5.21	0.41	N
cg16743781	*RHOC*	NOA vs. VR	7.96 ± 1.92	10.33 ± 2.23	2.37	0.41	N
cg17235766		NOA vs. VR	72.29 ± 2.69	78.3 ± 2.25	6.02	0.41	-
cg17895363	*GLRX2*	NOA vs. VR	90.84 ± 1.39	89.23 ± 3.29	-1.61	0.41	Y
cg18435505	*FAM47A*	NOA vs. VR	65 ± 3.46	66.05 ± 6.07	1.05	0.41	N
cg18747378	*HLA-C*	NOA vs. VR	26.32 ± 10.87	13.45 ± 5.02	-12.87	0.41	N
cg19089394	*RCE1*	NOA vs. VR	8.04 ± 1.04	7.34 ± 1.19	-0.70	0.41	N
cg20712178	*CBX5*	NOA vs. VR	92.1 ± 1.2	90.29 ± 2.78	-1.82	0.41	Y
cg22852353	*LMX1A*	NOA vs. VR	89.71 ± 1.81	85.13 ± 8.76	-4.59	0.41	N
cg23300732	*STK32B*	NOA vs. VR	9.04 ± 1.12	9.1 ± 2.09	0.06	0.41	N
cg23606922	*NFKB2*	NOA vs. VR	19.88 ± 1.94	21.22 ± 2.66	1.34	0.41	Y
cg26647332		NOA vs. VR	15.68 ± 1.93	19.82 ± 1.31	4.14	0.41	-
cg26687830	*EGFL8*	NOA vs. VR	87.85 ± 4.45	84.2 ± 6.09	-3.65	0.41	Y
cg01407235	*ARHGAP25*	NOA vs. VR	13.34 ± 2.4	9.89 ± 2.35	-3.46	0.42	N
cg08939371	*UNC45A*	NOA vs. VR	23.12 ± 6.85	19.04 ± 2.61	-4.08	0.42	Y

[1] Standard deviation

[2] The difference in mean methylation (group 2 – group 1)

[3] *P*-value derived from t-test moderated under a Bayesian framework with adjustments for batch

[4] As indicated from the Human Protein Atlas database interfaced via the R package hpar, version 1.18.1

A second model was also investigated which achieved a 91.7% accuracy (i.e., 22/24 correct classifications) and 75% on batch cases. Permutation testing identified a *P*-value = 0.001. This was achieved by training a RF model using the PC23 gene set filtered using the BIRF embedded method (i.e. PC23 + BIRF + RF). The misclassified cases were different from the PC23 + limma + SVM misclassification. The cases were a VR and NOA both predicted to be OA by the RF. A total of 299 were in the final gene set (data not shown). Of the probes in this gene set, 120 localized to 115 unique genes found to be expressed in the testis. A total of 19 probes in this gene set also had >10% methylation difference in at least 1 group comparison (Table 13). Lastly, 11 probes were also in the PC23 + limma + SVM gene set (cg03651613, cg00026909, cg19219655, cg24431515, cg06655062, cg00579694, cg15527554, cg05676450, cg18747378, cg13534734, cg26647332), which included the probe binding to *TMEM120B*.

Table 13. Biological significant probes associated with azoospermia as identified via PC23 + BIRF + RF model

Probe	Nearest gene	NOA mean methylation ± SD (%)	OA mean methylation ± SD (%)	VR mean methylation ± SD (%)
cg27506609	*TEKT2*	68.96 ± 10.38	57.63 ± 6.8	54.13 ± 5.95
cg22536808	*MMACHC*	10.12 ± 4.07	18.74 ± 2.31	24.89 ± 13.64
cg06529439	*RPIA*	27.69 ± 6.29	16.31 ± 7.32	14.75 ± 6.93
cg24918767	*HSD17B11*	32.65 ± 7.59	46.22 ± 7.04	36.69 ± 6.15
cg12065531	*RING1*	33.14 ± 8.75	17.36 ± 8.58	20.18 ± 9.71
cg04571327	*NDUFAF4*	31.9 ± 3.9	19.02 ± 10.45	18.78 ± 9.41
cg08742262	*PTPRN2*	38.85 ± 6.95	47.97 ± 8.61	36.71 ± 8.15
cg06565184	*MCPH1*	82.83 ± 9.71	83.38 ± 12.64	71.85 ± 5.78
cg01089095	*CHCHD1*	10.84 ± 7.33	20.98 ± 6.23	21.54 ± 9.6
cg16674860	*HPS1*	38.7 ± 5.56	26.1 ± 7.51	29.29 ± 11.77
cg21762589	*BNIP3*	62.27 ± 15.84	50.04 ± 2.61	51.96 ± 9.22
cg24431515	*ATHL1*	92.04 ± 2.18	83.82 ± 4.97	79.27 ± 9.53
cg04075973	*BCL9L*	15.59 ± 3.79	28.66 ± 9.08	26.22 ± 6.68
cg11332441	*THYN1*	35.12 ± 4.56	25.05 ± 6.78	23.67 ± 8.35
cg03651613	*TMEM120B*	14.68 ± 4.2	21.57 ± 4.09	30.4 ± 7.18
cg00469240	*CACNA1H*	83.93 ± 9.59	77.76 ± 8.58	73.35 ± 9.52
cg05358814	*COX11*	9.91 ± 3.47	20.18 ± 7.12	16.58 ± 5.28
cg01069823	*MCM5*	13.86 ± 1.07	11.86 ± 1.3	22.59 ± 18.47
cg13649400	*FAM120C*	27.38 ± 6.19	13.82 ± 9.61	16.75 ± 9.34

3.4.4 Gene enrichment analysis

To determine whether the significant gene sets carried probes localizing to genes that are over enriched in known biological pathways, the gene sets derived from PC23 + limma + SVM and PC23 + BIRF + RF were submitted to GO term analysis. Of the 299 probes mapping to 232 known RefGene names from the PC23 + BIRF + RF gene set, a total of 91 genes (Appendix A.1) was significantly associated with a single GO pathway, *localization* (GO:0051179; 1.59-fold enrichment; *P*=0.0154). Of the 91 probes, 53 were shown previously to be expressed in the testis. Using these 53 genes only, a total of 114 GO pathways were found to be significantly associated (Appendix A.2). Furthermore, of the 53 probes associated with localization, a total of 6 had

>10% methylation between at groups in at least 1 comparison (*COX11*, *TEKT2*, *PTPRN2*, *BNIP3*, *CACNA1H*, *MCPH1*; Table 13). Notable pathways included microtubule cytoskeleton organization involved in mitosis (GO:1902850), regulation of G2/M transition of mitotic cell cycle (GO:0010389), regulation of stem cell differentiation (GO:2000736), and positive regulation of cytoskeleton organization (GO:0051495). The 60 probes mapping to 45 known gene names from the PC23 + limma + SVM gene set was not significantly associated with any GO term.

3.4.5 Identification of differentially methylated regions

To determine whether there were regions of the genome with broad methylation changes, the bumphunter algorithm was used to iteratively group probes together to calculate regional methylation. With fertility grouping as the variable of interest and batch as the covariate identified, a total of 43 bumps or regions were identified with a beta difference between groups >20%. Of these 43 bumps, a bootstrap test (*n* = 1000) identified 4 DMRs with a *P*<0.05 (Figure 8). The first DMR identified (Figure 8) is a 60 bp region composed of 2 probes on chromosome 6. NOA cases were hypermethylated as compared to OA, which in turn were higher in methylation versus VR men (*P*=0.0098). The DMR is in a region without any known Refseq genes. The remaining 3 significant DMRs were each single probes (Table 14). A single probe (cg01232331) was found to also match with the PC23 + limma + SVM gene set.

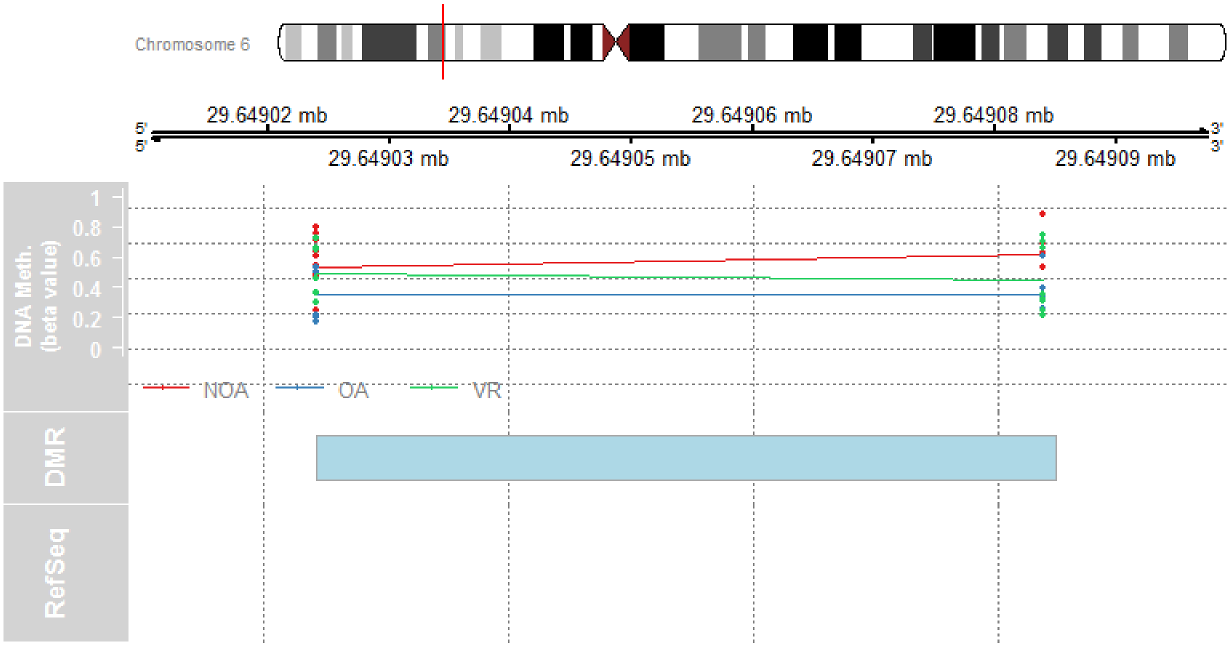


Figure 8. Differentially methylated regions between NOA, OA, and VR men
The upper track is the position of the DMR in context to the chromosome. The middle track represents the scale of the bp location of the region. The lower track shows the spread of the case methylation for the probes identified in the region. The bottom track is a visualization of the DMR block.

Table 14. Differentially methylated regions/probes between NOA, OA, and VR men as identified from bumphunting

Probe(s)	Nearest gene	Chr	Start (bp)	End (bp)	Beta Difference (%)	*P*-value	FWER
cg24100841, cg19636627	-	6	29649024	29649084	-0.29	0.009762	0.370
cg22459517	*EPS8L1*	19	55587193	55587193	0.38	0.010585	0.595
cg01232331	-	2	52759434	52759434	-0.32	0.036973	0.868
cg21158431	-	6	167786059	167786059	0.31	0.045213	0.886

3.4.6 Analysis of global methylation features between fertility groups

To determine whether any broad genomic feature was associated with azoospermia, the proportions of the 2 DMP gene sets were tested for significant differences with non-significant probes. The 450k microarray has annotation data on probes including the chromosome to which they bind, whether they are associated with CpG islands, relationship to CpG islands such as 0-2 kb upstream (N_Shore), downstream (S_Shore), 2-4 kb upstream (N_Shelf), downstream (S_Shelf), dissociated from CpG islands >4kb (open sea), DNase I hypersensitive sites, associated with promoters, DMRs, reprogramming-specific (r)DMRs, cancer-specific (c)DMRs, imprinted genes, Hidden Markov Model (HMM) predicted CpG islands, FANTOM 4 promoters.

The 299 probes in the PC23 + BIRF + RF gene set was significantly associated with a total of 5 genomic features. There was a higher proportion of probes binding to CpG sites associated with promoters (29.8% vs. 20.1%; *P*=0.0034), CpG islands (48.2% vs. 32.0%; *P*=4.2x10^{-7}), DNAase hypersensitive sites (20.1% vs. 12.7%; *P*=0.018), FANTOM 4 promoters (12.4% vs. 6.9%; *P*=0.43), and HMM predicted CpG islands (63.2% vs. 50.8%; *P*=0.00089) among the PC23 + BIRF + RF gene set probes as compared to the remaining probes on the array. The 60 probes in the PC23 + limma + SVM gene set were not associated with any genomic feature.

3.5 Discussion

In this study, we surveyed for novel genes or pathways associated with differential methylation between NOA, OA, and VR men using genome-wide methylation microarrays. We implemented multiple analytical pipelines, including numerous novel methods, to identify DMPs, which resulted in 2 gene sets totaling 359 probes capable of classifying fertility with >90% accuracy in our cases using trained ML models. In the first gene set, we identified significantly lower methylation (>10% difference between groups) at the *TMEM120B* and *NDE1* genes among NOA men when compared with VR men. Both probes localize to the promoter regions within 200 bp of the TSS of their respective genes as well as within CpG islands (although the probe binding to *NDE1* is further away >4kb). Thus, the methylation at these probes are likely to have a mechanistic role in the regulation of the expression of these genes. The *TMEM120B* gene produces nuclear envelope transmembrane proteins that are not very well studied but has been shown to be linked to efficient adipogenesis. NDE1, however, is required for centrosome duplication and formation, and function of the mitotic spindle. It is highly expressed during mitosis (Kim et al., 2011). The NUDE_C domain allows interaction of NDE1 with centromere protein F (CENPF), which associates with the centromere-kinetochore complex (Vergnolle et al., 2007), and the dynein complex. In zebrafish, the depletion of Nde1 show suppression of cell division and reduction in the number of cells (Feng et al., 2004). Human research has shown NDE1's role is important for neuron and brain development through the regulation of mitosis progression of neurons (Alkuraya et al., 2011). Homozygous mutations of this gene have been shown to be linked to microcephaly, i.e. head circumference is smaller than normal, with profound mental retardation due to the lack of division of neurons (Bakircioglu et al., 2011). However, little is known about NDE1's role in the testis. *NDE1* is expressed in low amounts in the cells of the seminiferous ducts as well as by Leydig cells (Uhlén et al., 2010; Uhlén et al., 2015). Therefore, it is plausible that NDE1 may be involved with the mitotic progression of spermatogonia in the testis. The lower methylation near the promoter of *NDE1* in NOA men in this study might suggest higher expression of *NDE1*, perhaps as an overcompensation mechanism for the reduced spermatogenesis in these men. This overexpression may also be in the Leydig cells to compensate for hormonal levels; however, we did not find a significant difference in LH and FSH between NOA and VR men in this study. Our lab and others have shown that proper synapsis and alignment of chromosomes with the aid of microtubules during recombination is essential for proper segregation and thus progression of spermatogenesis (Ferguson et al., 2007; Ren et al., 2016). Given the role of NDE1 as an organizer of microtubule structure during mitosis, this gene may also play an important role during meiosis and thus progression of spermatogenesis. In fact, the centromere is well regarded as important for fertility (Sathananthan et al., 1998; Simerly et al., 1995). Alternatively, the lower *NDE1* expression might instead suggest an alteration among VR men. Given the elective obstruction of sperm leaving the testis, higher methylation and thus potentially lower *NDE1* expression in VR men may be a mechanism to reduce spermatogonia division. It has been shown previously that VR men carry epigenetic differences in their sperm which suggests that there may be pathways altered in response to the obstruction in these men (Minor et al., 2011). Future studies should focus on investigating the methylation of the NDE1 promoter and the expression of this gene in the testis of hypospermatogenesis or meiotic arrest NOA men.

Of note, the methylation at *TET3* was also identified to have significantly lower methylation among NOA men as compared to VR men. The probe localizes to >4kb from the nearest CpG island but within 1.5 kb of the TSS. *TET3* is involved with demethylation pathways and is highly expressed in the testis. The lower methylation and thus potentially higher expression in the testis of NOA men may suggest a potential mechanism for hypomethylation in the sperm as observed from previous studies, including in Chapter 2 of this thesis. Only a single study has been conducted investigating TET enzymes in male infertility, which showed that *TET1-3* expression is positively associated with sperm concentration and progressive motility. Oligozoospermia and asthenozoospermia males showed significantly reduced *TET3* expression (Ni et al., 2016). This is opposite to our finding; however, the methylation values in our cases are quite comparable between groups despite the slight increase in NOA men. Nevertheless, further investigation into the methylation of this gene may provide another potential pathway involved with male infertility.

The second highest accuracy model, PC23 + BIRF + RF, only misclassified 2/24 cases, 1 of which was the batch case. This gene set consisted of more genes than the former set, totaling 299 probes. Interestingly, a subset of these probes was enriched in pathways linked to the GO term for localization. When filtering these probes for only those expressed in the testis, numerous pathways were identified that were linked to mitosis and microtubule formation in sperm flagella. The probe localizing to the Tektin protein 2 (*TEKT2*) gene indicated higher methylation in NOA men as compared to both OA and VR men. This gene has been shown in mice and mammals to be involved with the stability and structural complexity of flagella (Yan et al., 2009). In humans, *TETK2* is involved with the stability of the axoneme stability and structure thus involved with ciliary movement of the flagella. Asthenozoospermia patient screening identified that a heterozygous mutation in this gene is linked to the lack of motility in sperm (Hwang et al., 2010). Our finding of higher methylation at 0-2kb from a CpG island near the TSS might suggest lower expression in NOA men and therefore is in concurrence with previous genetic studies. The lower expression in NOA but not OA men may be indicative that OA men have normal spermatogenesis in their testis producing motile sperm. Therefore, the hypermethylation at *TEKT2* in the testicular cells in NOA men may be an etiology for asthenozoospermia. The probe binding to the gene body of microcephalin (*MCPH1*) was also found to be significantly higher in methylation in NOA and OA men as compared to VR men. *MCPH1* codes for a DNA damage protein that plays a role in the G2/M checkpoint and thus progression into mitosis. Mice studies have shown that deficiency of Mcph1 removes the localization of Chk1 to centrosomes, resulting in premature Cdk1 activation and early mitotic entry thereby uncoupling the mitosis and centrosome cycle (Gruber et al., 2011). This disrupts the mitotic spindle alignment resulting in failure to progress through mitosis. In addition, Mcph1 knockout mice were found to be infertile with homologous recombination impaired during meiosis (Liang et al., 2010). The spermatocytes show failure to synapsis during meiosis resulting in meiotic arrest at late zygotene of prophase I and subsequent apoptosis. MCPH1 is suggested to recruit RAD51/BRCA2 to DNA damaged sites for repair including during recombination. Thus, the higher methylation at the *MCPH1* gene in NOA and OA men may be involved with reduced MCPH1 protein resulting in mitotic or meiotic arrest, both of which may be etiologies in spermatogenic dysfunction. Indeed, our previous work have suggested that NOA men have reduced rates of recombination (Ferguson et al., 2007). This is yet another candidate along with *NDE1* and *TEKT2* to be investigated further in relation to the epigenetically regulated meiotic pathways associated with male infertility.

Only 3 studies have previously investigated the testicular methylation in azoospermic men (Khazamipour et al., 2009; Ferfouri et al., 2013; Ramasamy et al., 2014). The first study using methylation specific PCR (i.e. the evaluation of a single CpG site) identified altered methylation at the *MTHFR* promoter, suggesting that misregulation of MTHFR in the testis may result in altered methyl supply in the germ cells during spermatogenesis. This provides a direct rationale for the observed methylation errors in previous sperm studies including in Chapter 2. This inherently suggests that methylation in the germ cells may be important for the progression of spermatogenesis and that the altered methylation in the sperm of infertile men is a residual artifact of an altered spermatogenesis cycle. Ferfouri et al. identified 14 testis-specific genes with methylation differences between well-defined azoospermic men histological evidence for pathology. They used the Illumina HumanMethylation27 BeadChip, which is similar to the platform in this study but surveys a much fewer 27,000 CpG sites. Their pipeline included only the use of an F-test with moderation with a Bayesian framework. Of the 14 genes from their study, 3 were also identified from our PC23 + BIRF + RF model (*RBM24*, *BPI*, *MCM5*). Despite differences in analytical methodology, the replication of a portion of their results further supports our findings. Ramasamy et al. identified approximately 10,000 DMPs using the Illumina 450k microarray between 16 NOA and 5 proven fertile men. Of the 10,000, they narrowed down to 20 CpG sites with >30% differential methylation between groups. They selected a total of 6 genes which had known high expression in the testis for validation using RT-qPCR. Only discoidin domain receptor 1 (*DDR1*) had both altered methylation and expression in NOA men. In our study, we did not find *DDR1* as having altered methylation. This may be a result of heterogeneity between patient cohorts. Furthermore, Ramasamy et al. classified a case as NOA via ejaculate analysis whereas we had confirmation from histological analysis. They also included stringent criteria for their inclusion of significant DMPs thereby reducing their number of final hits. Nevertheless, with only 3 studies including this one having investigated the methylation in the testis of azoospermic men, strong conclusions should be refrained until further studies can verify our findings.

The use machine learning classification models in this study was a first for the investigation of methylation in male infertility. However, these methods have been used for over a decade in other fields of research such as cancer (Pirooznia et al., 2008; Saeys et al., 2007; Guyon et al., 2002; Statnikov et al., 2008). These studies have shown success using ML algorithms for the classification of patients with rare or specific types of cancer. The reason for this is due to the numerous issues have been cited regarding the use of “traditional” applied statistics, i.e. linear regression models, when working with microarrays. First, most studies have few samples due to the cost of such arrays, therefore resulting in a study design with few samples and high number of features. This is generally an experimental design not appropriate for a linear model. Second, linear models have numerous assumptions that if violated may not produce accurate inferences. Traditional statistical methods including t-test may not be entirely accurate given that such techniques rely on an underlying assumption of a normal distribution and independence. From functional studies, it has been shown that methylation at neighboring CpG sites are not independent from one another. With these issues in mind, we applied ML algorithms in identifying DMPs from microarrays. The non-parametric nature of ML algorithms, i.e. not making assumptions of the underlying distribution of the methylation data, was used because of the complexity that exists within epigenomic data. A further novel aspect of this study is the use of multiple analytical methods which is analogous to having multiple ‘expert opinions’ on the analysis of the methylation data (Peng et al., 2006; Tan et al., 2003; Liu et al., 2004). Indeed, we identified at least 2 potential pathways epigenetically linked to mitosis/microtubule structure. Overall, our results indicate that these methods can indeed supplement the analysis of the 450k and the newer 850k methylation microarrays.

We developed a novel PCA filtering method which was partially adopted from a previous study (Roden et al., 2006). The study suggested that low ranking PCs despite describing little variance to the total variance in the data, may still provide insight to the biological mechanisms of the dataset. Indeed, the use of PCA was critical in the finding of our gene sets. PCA is a linear technique that identifies orthogonal, i.e. independent from each direction, directions of variance in the data set, each of which are linear combinations of the methylation of each gene. We identified 2 PCs in our results as significantly associated with our fertility grouping. However, PC3 and PC23 had few overlapping genes, only 1472 probes were overlapping between gene sets, suggesting that these PC described different aspects of the variance in methylation. PC3 was found to show better separation between NOA and VR, while PC23 showed better separation between VR and OA (data not shown). Indeed, the PC23 gene set was the base for the highest accuracy models due to difficulty by most of our models in classifying between OA and VR cases. This may indeed be a biological effect as OA and VR are very similar in phenotype, only differing in their etiology in pathology. This may also be a batch effect issue as majority of OA and VR cases were from within the same batch while NOA were on a different batch. Thus, PC23 may have abstracted from the batch effects while batch may have been more linked to PC3. This technique also provides advantages when dealing with imbalances in groups between batches as well as other technical variances such as cell type heterogeneity. Indeed, the top PC accounting for the majority of the variance in the dataset was not associated with any of our known clinical information. We predict this to be cell type differences between sample as suggested by other studies (Jaffe et al., 2014). Future studies may obtain clearer results when controlling for batch and cell type at the experimental design stage instead of using post-experiment statistical techniques.

Some caution should be taken when interpreting our PC23 + limma + SVM model as the embedded method optimized for the FDR cutoff, as opposed to using the standard 5%. The optimal FDR was 43%, which is normally very high as it could be interpreted as 43% of the probes may be false positives. However, the more relaxed FDR presented with more probes to further investigate. When the 60 probes were cross-referenced with the literature, expression databases, and GO, we were able to narrow down to 2 probes that were expressed in the testis and had a biological rationale linked to spermatogenesis. This novel embedded method that we have presented here may be an optimal combination between regression statistics and the use of empirical evidence from ML models. Using ML to validate the probe set may be a suitable method for choosing an FDR. The high scoring on the batch cases (i.e. cases that were difficult to predict due to batch bias in the group distribution) suggest that both PCA and the linear adjustment for batch were capable of removing the majority of batch effects.

There were limitations in our study. As with most other microarray studies, our sample size is limited. This was a result of the difficulty in recruiting azoospermic cases. Second, clinical data was not available for all cases which resulted in fewer adjustments that could have been made in linear modelling. Lastly, technical or batch variation was an important aspect we took into consideration when designing our analytical pipeline. Due to the severe imbalance of groups between batches (i.e. most NOA were on one batch while most OA and VR were on the other), there was a high risk for confoundment of the results due to batch. However, we applied a non-conventional criterion in evaluating our gene sets in that the classification of the "batch or hard" cases took priority in evaluating the accuracy scores. Furthermore, we applied direct adjustment in the linear modelling as well as selecting PCs that did not show correlation with batch. However, despite these methods, the risk for batch effects may still be there as NOA was always easiest to categorize, while OA and VR were difficult. This could mean that it was easy for the ML classifiers to differentiate cases between batches. This might also inherently be a biological effect VR is a type of OA with the only difference in their etiology and that both groups should have normal spermatogenesis. Another limitation or caution is that with the numerous methods we applied, we run the risk of overfitting our models to our data. However, we did find similar hits from the Ferfouri et al. study suggesting that our results may be reproducible.

Overall, using newly developed gene selection methods, in our investigation of genome-wide methylation in the testis of NOA, OA, and VR men, we identified 359 CpG sites as candidates for DMPs. The training of machine learning models with these hits allowed for the prediction of the phenotype of our cases with >90% accuracy. Of the significant hits, 3 with >10% differential in methylation between groups present strong biological rationale for azoospermia. The genes *NDE1*, *TEKT2*, and *MCPH1* have been shown in other fields of research to be involved with mitotic progression, meiotic recombination, and spindle/microtubule regulation at centromeres. The aberrant methylation of such genes are potential etiologies to the dysfunctional spermatogenesis in infertile men. The work done in this study may have diagnostic clinical applicability given the ease of implementing our novel bioinformatic methods. The results herein present novel biomarkers for male infertility and an epigenetic view on potential etiologies of male infertility, spermatogenic failure, and methylation defects in the sperm.

CHAPTER 4: INTEGRATED INVESTIGATION OF GENOME-WIDE SNPS AND METHYLATION ASSOCIATED WITH INFERTILE MEN

4.1 Background

Nearly 7% of all men experience the inability to achieve a pregnancy after regular unprotected sexual intercourse (i.e. male infertility). In up to 15-30% of male infertility cases, a genetic cause can be attributed to the condition (Krausz et al., 2015). Furthermore, despite the heterogeneity of male infertility, genetic causes can be identified in all major categories (Krausz et al., 2011). This perhaps is due to the fact that spermatogenesis and male fertility health is a complex process involving the proper expression and coordination of hundreds of genes, many of which are still unknown (Aston, 2014). Known genetic causes include microdeletions on the Y chromosome within the AZF region, mutations within the *CFTR* gene, whole chromosome abnormalities such as Klinefelter syndrome, and structural chromosome abnormalities such as translocations. Overall, genetic causes of male infertility are typically found to affect genes involved with spermatogenesis and expressed in the testis or chromosome structural changes which affect the progress of meiosis.

Emergence of population specific studies have identified >314 SNPs, i.e. common mutation variants, associated with male infertility from 123 genes (Carrell and Aston, 2011; Krausz et al., 2015). However, very few have been verified by independent sequencing studies. Only single studies are available for the majority of the SNPs (n = 269). Furthermore, meta-analyses and literature reviews have identified only few genetic factors truly significantly associated with impaired spermatogenesis (Tüttelmann et al., 2007; Krausz et al., 2015). There is a large discordant in results between GWAS studies (Aston and Carrell, 2009; Aston et al., 2010; Dalgaard et al., 2012; Hu et al., 2012; Zhao et al., 2012). Ethnicity/geographic origin seems to play an important role in the effect of a SNP in the manifestation of the disease. Furthermore, many of the SNPs identified only have a modest risk, i.e. non-rare variants, for male infertility despite their significant association (Wei et al., 2012). Thus, unless a rare variant is identified, it may be plausible that common SNPs alone may not be single drivers for male infertility.

As demonstrated from Chapter 2, altered methylation in the sperm of infertile men were identified in men with the *MTHFR* C677T SNP. In that proposed model, the methylation of imprinted genes may be affected by reduced performance of MTHFR due to the SNP (Louie et al., 2016). Indeed, the combinatorial effect of methylation and SNP at seemingly distinct loci may in fact be involved with male infertility. The significant association of methylation and SNPs have been identified in other studies unrelated to male infertility (Tsaprouni et al., 2014; Soto-Ramíreze et al., 2013). Thus, it is plausible that given the numerous and separate SNP/methylation studies, many of these hits may in fact be conditionally related to each other. The investigation of both methylation and SNP concordantly may presents a novel analysis for the identification of genetic/epigenetic driver genes related to male infertility.

4.2 Rationale

Of the 2300 genes estimated to be involved with spermatogenesis (Schultz et al., 2003), only 139 have been validated by re-sequencing. This is less than 1% of the protein-coding genes in the genome. Furthermore, given the rise of non-coding protein regions of the genome presenting regulatory effects on distal regions of the genome, the actual effort thus far is estimated to only represent 0.01% of the potential genes linked to male infertility (Aston, 2014). Thus, the identification of potential SNPs associated with male infertility is far from complete. The advancement of microarray technologies to evaluate >900,000 SNPs in a single assay has provided an unprecedented ability to identify rare variant SNPs involved with diseases.

Given that a significant portion of the genome has yet to be evaluated for genetic causes of male infertility, yet up to 50% of male infertility cases are idiopathic, the likelihood of additional genetic variants contributing to this disease is plausible. The aim of this study was to identify novel SNPs associated with dysfunctional spermatogenesis. The primary objective was to evaluate SNPs in a genome-wide approach in the peripheral blood of NOA men and fertile controls. The secondary objective was to conduct a novel integrated analysis by combining both the SNP data with the methylation data (Chapter 3), for cases that have both, into a single dataset for analysis.

4.3 Methods

4.3.1 Patient recruitment

Patient recruitment practices were the same as described in 2.3.1 and 3.3.1. Ethical approval was obtained from the University of British Columbia Ethics Committee prior to commencing study. The control men group includes men who had undergone VR.

4.3.2 Sample processing

Peripheral blood samples were processed as described in 2.3.10. Extracted DNA samples were further concentrated if the DNA concentration was below the minimum 50 ng/µL as required for microarray application (Amicon Ultra, Millipore, ON, Canada). DNA samples were sent to TCAG (Toronto, Canada) for the application to the Affymetrix Genome-Wide Human SNP Array 6.0 (SNP 6.0).

4.3.3 Microarray analysis

Raw CEL files were imported into R using the crlmm R Bioconductor package (1.34.0; Carvalho et al., 2010; Ritchie et al., 2009; Scharpf et al., 2011). Normalization of the raw fluorescence intensities to remove array-to-array variability and subsequently summarization of replicate probes was conducted using the robust multichip average (RMA; Irizarry et al., 2003; Carvalho et al., 2006). Genotyping of SNPs were achieved using the corrected robust linear mixture model (CRLMM) algorithm (Carvalho et al., 2006; Carvalho et al., 2009). The identification of informative SNPs was achieved with a balanced iterative random forest (BIRF; see section 3.3.6.1). The leave-one-out-cross-validation (LOOCV) was used for the evaluation of the minimal set of SNPs after BIRF. Briefly, a support vector machine (SVM) classifier was trained with N -1 cases and used the excluded case as a test case. This is repeated N times until each case is tested exactly once. The total accuracy score is determined by the number of correct cases / N cases. Permutation testing was further used to test for the probability of the LOOCV result above random chance. Briefly, test sets (n = 1000) were generated with the case labels randomly shuffled. The proportions of simulations with >= to the highest LOOCV accuracy score was used as the P-value. The selected SNPs were further filtered using a two-tailed

Fisher’s exact test where significance was considered with a *P*-value below a 5% FDR. The identification of genes associated with RS SNP ids was achieved using the biomaRt (2.32.1) R Bioconductor package (Durnck et al., 2005 and 2009). The identification of enriched biological pathways followed the same methodology as described in section 3.3.7. The determination of expression of candidate genes in testis was cross-checked with the Human Protein Atlas database (Uhlén et al., 2010; Uhlén et al., 2015). The database was accessed from with R using the hpar (1.18.1) Bioconductor package (Gatto, 2017).

4.3.4 Integrated analysis

The analysis of both methylation and SNP calls from the same cases was achieved using an integrated analysis. Each of the respective raw datasets were processed separately as described in previous sections (e.g. normalization and probe filtering). The datasets were combined into a single 6 x 1,348,079 matrix where gene selection was achieved using a BIRF with the following hyperparameters: *ntree* = 3000, *sampsize* = 3 for NOA and 1 for VR, *cutoff* = 4/6 for NOA and 2/6 for VR. The minimal set of methylation and SNP probes with the least out-of-the-box error after BIRF was validated using a RF LOOCV. Permutation test of the LOOCV was conducted as described in section 4.3.3. Permutation feature importance (PFI) was used to determine ranking of features (Docs.microsoft.com, 2018). Briefly, PFI randomly shuffles a feature’s values and reevaluates the model accuracy. This is repeated for each feature once. The feature which results in the greatest accuracy drop is deemed to be the most important.

4.4 Results

4.4.1 Patient Summary

A total of seven NOA, 13 control cases (five fertile controls and eight VR), and one unknown case were retrospectively studied. Of the cases, six (four NOA, two fertile controls) overlapped with those that were studied for genome-wide methylation in Chapter 3. Clinical information was not available for these cases.

4.4.2 Identification of informative SNPs associated with infertility

A total of 120 SNPs were identified in the minimal set of probes with the lowest error rate via BIRF. SVM LOOCV achieved a prediction accuracy of 100%. Permutation testing identified a mean accuracy of 64% for the simulations with the highest achieved accuracy being 90%. Therefore, an accuracy of 100% in the SVM LOOCV achieves a *P*-value = 0.001. Of the 120 SNPs, 39 were found to have significantly different proportions of NOA and C groups (Table 15). Only 23 were mapped to 22 known genes. These 22 genes were not significantly enriched in any GO pathway, and did not overlap with the genes that had differential methylation between NOA and VR cases (Chapter 3). Furthermore, none of the identified SNPs were within exon regions. One SNP probe localized to a ncRNA region and the remainder to intronic regions. Of the 39 SNPs, none would be considered a rare variant as the highest global minor allele frequency identified was 0.079.

Table 15. Significant SNPs associated with NOA vs. control men

dbSNP[1] RS ID	Gene	Chr	Global MAF[2]	Allele A	Allele B	NOA SNP Proportions	Control SNP Proportion	Adj. *P*-value[3]
rs11954272	*NREP*	5	0.249	C	G	AA = 0.14; AB = 0.86; BB = 0	AA = 1; AB = 0; BB = 0	0.0217
rs10757087	*SLC24A2*	9	0.3808	A	G	AA = 0.29; AB = 0.71; BB = 0	AA = 0; AB = 0.15; BB = 0.85	0.0279
rs11993028	*CSMD1*	8	0.2686	A	T	AA = 0; AB = 0; BB = 1	AA = 0.31; AB = 0.54; BB = 0.15	0.0287
rs6045762		20	0.4151	C	T	AA = 0.71; AB = 0.29; BB = 0	AA = 0; AB = 0.54; BB = 0.46	0.0287
rs1504772	*CSMD1*	8	0.2674	A	G	AA = 1; AB = 0; BB = 0	AA = 0.15; AB = 0.54; BB = 0.31	0.0287
rs11171979		12	0.2268	C	G	AA = 0.86; AB = 0; BB = 0.14	AA = 0.15; AB = 0.77; BB = 0.08	0.0287
rs7589808	*LOC105373597*	2	0.4842	C	T	AA = 0.43; AB = 0.57; BB = 0	AA = 0.08; AB = 0.15; BB = 0.77	0.0287
rs16961948		16	0.3337	A	T	AA = 0.14; AB = 0.57; BB = 0.29	AA = 0.08; AB = 0; BB = 0.92	0.0292
rs10831936	*LOC105376557*	11	0.36	A	G	AA = 0.57; AB = 0.29; BB = 0.14	AA = 0; AB = 0.23; BB = 0.77	0.0292
rs4471922	*TTN*	2	0.4425	A	C	AA = 0.14; AB = 0.86; BB = 0	AA = 0.69; AB = 0.08; BB = 0.23	0.0292
rs10017153		4	0.2578	A	G	AA = 0.14; AB = 0.14; BB = 0.71	AA = 0.08; AB = 0.85; BB = 0.08	0.0292
rs694935		1	0.495	A	T	AA = 0; AB = 0.29; BB = 0.71	AA = 0.62; AB = 0.31; BB = 0.08	0.0292
rs17153375	*LOC105379160*	5	0.2522	A	C	AA = 1; AB = 0; BB = 0	AA = 0.23; AB = 0.69; BB = 0.08	0.0292
rs2385509	*MTR*	1	0.3271	C	G	AA = 0; AB = 0; BB = 1	AA = 0.23; AB = 0.54; BB = 0.23	0.0304
rs7714765	*PDE4D*	5	0.4197	A	G	AA = 0.29; AB = 0; BB = 0.71	AA = 0.15; AB = 0.69; BB = 0.15	0.0304
rs8017481	*NIN*	14	0.3279	A	G	AA = 0; AB = 0.86; BB = 0.14	AA = 0.08; AB = 0.15; BB = 0.77	0.0304
rs888064		14	0.4938	A	G	AA = 0; AB = 0.43; BB = 0.57	AA = 0.46; AB = 0.54; BB = 0	0.0304
rs3829168	*LOC105378305*	10	0.1595	C	T	AA = 0.14; AB = 0.71; BB = 0.14	AA = 0; AB = 0.15; BB = 0.85	0.0304
rs12947636	*SLC39A11*	17	0.4639	C	T	AA = 0; AB = 0.14; BB = 0.86	AA = 0.08; AB = 0.77; BB = 0.15	0.0304
rs574993	*SPPL3*	12	0.0785	C	T	AA = 0.43; AB = 0.57; BB = 0	AA = 1; AB = 0; BB = 0	0.0321
rs11652547		17	0.3476	C	G	AA = 0; AB = 0; BB = 1	AA = 0.15; AB = 0.54; BB = 0.31	0.0321
rs564225		9	0.2909	C	G	AA = 0.14; AB = 0.43; BB = 0.43	AA = 0; AB = 0; BB = 1	0.0321
rs10762164	*CTNNA3*	10	0.4046	C	T	AA = 0.14; AB = 0.57; BB = 0.29	AA = 0; AB = 0.08; BB = 0.92	0.0321
rs4600084		1	0.4437	A	C	AA = 0; AB = 0.43; BB = 0.57	AA = 0.62; AB = 0.31; BB = 0.08	0.0321
rs16851854		1	0.2724	C	T	AA = 0; AB = 0.57; BB = 0.43	AA = 0; AB = 0; BB = 1	0.0321
rs7282770	*DOPEY2*	21	0.3101	A	C	AA = 0.86; AB = 0; BB = 0.14	AA = 0.31; AB = 0.69; BB = 0	0.0321
rs13179066		5	0.2812	C	T	AA = 0.29; AB = 0.71; BB = 0	AA = 0.08; AB = 0.23; BB = 0.69	0.0321
rs13075697	*LINC00693*	3	0.3832	A	G	AA = 0.14; AB = 0.86; BB = 0	AA = 0.08; AB = 0.23; BB = 0.69	0.0365
rs12571712	*LOC105378305*	10	0.4347	A	G	AA = 0.86; AB = 0.14; BB = 0	AA = 0.15; AB = 0.46; BB = 0.38	0.0369
rs7907957		`	0.3554	A	G	AA = 0.14; AB = 0.86; BB = 0	AA = 0.46; AB = 0.15; BB = 0.38	0.0369
rs8065893		17	0.3047	A	G	AA = 0.14; AB = 0.43; BB = 0.43	AA = 0.69; AB = 0.31; BB = 0	0.0408
rs7576705	*LOC105373951*	2	0.4712	C	T	AA = 0.43; AB = 0.43; BB = 0.14	AA = 0; AB = 0.31; BB = 0.69	0.0408
rs6759703		2	0.3365	A	G	AA = 0.14; AB = 0.57; BB = 0.29	AA = 0.77; AB = 0.23; BB = 0	0.0408
rs950217		1	0.4563	C	T	AA = 0; AB = 0.14; BB = 0.86	AA = 0.23; AB = 0.62; BB = 0.15	0.0441
rs17782554	*XKR6*	8	0.1747	C	G	AA = 0.86; AB = 0.14; BB = 0	AA = 0.15; AB = 0.62; BB = 0.23	0.0441

rs1396848	*LUZP2*	11	0.382	A	C	AA = 0.57; AB = 0.29; BB = 0.14	AA = 0; AB = 0.46; BB = 0.54	0.0495
rs1845257	*CLVS1*	8	0.4828	A	G	AA = 0; AB = 0.43; BB = 0.57	AA = 0.54; AB = 0.38; BB = 0.08	0.0495
rs10781180		9	0.4573	C	T	AA = 0.14; AB = 0.86; BB = 0	AA = 0.08; AB = 0.31; BB = 0.62	0.0495
rs1679804	*LOC105370512*	14	0.2817	C	T	AA = 0.86; AB = 0.14; BB = 0	AA = 0.15; AB = 0.54; BB = 0.31	0.0495

[1] dbSNP Build 151

[2] Global minor allele frequencies; from 1000 Genomes Project data

[3] Two-tailed Fisher's exact test with the Benjamini-Hochberg procedure for the correction of multiple comparisons.

4.4.3 Integrated analysis of SNP and methylation

A total of six cases (four NOA and two VR) had both 450k methylation and SNP 6.0 array data. To determine whether there are methylation and SNP combinations that were associated with NOA, an integrated analysis using both datasets was conducted. A total of 925 probes (698 methylation and 227 SNP probes) were found to predict cases in a RF LOOCV with 100% accuracy. The top 50 features ranked by their importance is presented in Table 16. Permutation testing identified that 6.5% of the simulations also achieved 100% accuracy (i.e. *P*=0.065). However, all 6.5% were found to be variations of the actual case labels, thus signifying that no other simulations achieved 100% accuracy. The next highest permutation simulation achieved an accuracy of 67%. Enrichment analysis of pathways identified no pathways linked to the top 50 probes with known genes expressed in the testis.

Table 16. CpG and SNPs identified in an integrated analysis associated with NOA

Probe	Gene	Importance[1]	Testis expressed
cg16264537	-	-31.60	
SNP_A-8430911	*SPATA22*	-31.46	
cg18124009	*MTA1*	-30.95	True
cg12298241	-	-30.44	
SNP_A-1930168	-	-30.39	
cg01641514	-	-30.24	
cg00223715	*FAM19A5*	-30.14	True
cg03016446	*CBARA1*	-29.97	
SNP_A-2100736	*PAK5*	-29.97	
SNP_A-2314162	-	-29.90	
SNP_A-2030850	-	-29.88	
cg13242661	*ADRBK2*	-29.70	True
cg14053001	*PFN4*	-29.66	
cg26419378	*TRIO*	-29.62	True
cg16333846	*TIAM2*	-29.53	True
SNP_A-8419713	-	-29.53	
cg02991799	*ZNF302*	-29.51	
cg24076474	*RGL3*	-29.50	True
SNP_A-8444181	*CSMD1*	-29.40	
cg26611723	-	-29.37	
cg20838586	*RBMXL2*	-29.36	True

SNP_A-8562352	*DIRC3*	-29.30	
cg05882522	*EHD3*	-29.26	True
cg02233493	*FLJ30058*	-29.25	
cg20386586	*TGIF1*	-29.23	
cg17826956	-	-29.18	
cg01312546	*ZC3H3*	-29.17	True
cg10754935	*FBXL15*	-29.16	
cg26707646	*WIPF1*	-29.13	
SNP_A-2026963	*NFXL1*	-29.05	True
cg04219700	*TRIM11*	-29.03	True
cg06728437	*KLF7*	-29.00	True
cg19883472	-	-28.96	
cg09445799	*NFATC2IP*	-28.95	
cg02970425	*RNF114*	-28.94	True
cg03469484	*TTC15*	-28.86	
cg09174205	*ZNF516*	-28.81	True
SNP_A-8608210	*KCNMA1*	-28.79	True
cg01996723	-	-28.77	
cg15727409	*SRPX*	-28.76	
SNP_A-8632470	*PLCE1*	-28.74	True
cg11630920	*CPZ*	-28.73	
cg14554743	*MRPL30*	-28.71	True
cg24159891	*GGT1*	-28.71	
cg09735014	-	-28.71	
cg17960934	*AUTS2*	-28.69	True
cg00531745	-	-28.67	
SNP_A-2053449	*LINC01795*	-28.66	
SNP_A-4209800	*RSRC1*	-28.59	True
SNP_A-8445268	*PECAM1*	-28.56	

[1] Permutation feature importance score: baseline accuracy of model – model accuracy after shuffling feature values.

4.5 Discussion

Previous studies have identified >300 SNPs associated with male infertility, many of which have been recently validated in separate cohorts (Aston, 2014). The goal of this study was to further identify novel SNPs or potentially confirm a previously identified SNP. In this study, we investigated the genotypes at >900,000 SNPs across the genome from NOA and fertile control men. We identified 39 candidate SNPs associated NOA men and thus potentially associated with impaired spermatogenesis. We further presented a novel investigation of mining for genes that may be epigenetically and genetically linked and associated with male infertility. We identified 925 potential candidates in this regard.

Despite the numerous GWAS studies identifying SNPs involved with impaired spermatogenesis, we too included a GWAS investigation. We, however, included novel methods in identifying SNPs whereas previous methods included traditional methods. We implemented machine learning classifiers achieving 100% accuracy in the trained prediction models. Of the 39 SNPs evaluated as associated with NOA, two have biological and literature precedence. The 5-Methyltetrahydrofolate-Homocysteine Methyltransferase (*MTR*) gene encodes for the methionine synthase enzyme. As with *MTHFR*, this gene is involved with the folate cycle, which is a source of methyl donors for cellular methylation reactions. The *MTR* gene has been previously associated with male infertility in certain populations (Kurzawski et al., 2015). All of the NOA men in our study had the GG genotype, which was only found in 25% of controls. The role for this gene in male infertility is unclear; however, as with *MTHFR* it is suggested to be associated with the proper of methylation in germ cells. Ninein (*NIN*) is a testis expressed gene which produces a protein associated with positioning and anchoring microtubules minus-ends in epithelial cells (Mogensen et al., 2000; Delgehyr et al., 2005). This protein localizes to the centrosome and is potentially important for chromosome segregation during mitosis (Chen et al., 2003). We identified that 86% of NOA men in our study had the A allele at the SNP as opposed to only 23% of controls. Thus, it is plausible that a SNP within the *NIN* gene may have adverse consequences for mitotic progression in testicular cells thereby contributing to NOA. However, this gene has yet to be investigated in the context of male infertility.

Given the spurious findings of SNPs and methylation at genes associated with male infertility in recent GWAS studies, we conducted an integrated analysis to combine methylation and SNP data. We identified a >900 CpG and SNP combination that predicted our cases with a significant 100% accuracy. Surprisingly, we found that conducting individual non-parametric testing / proportion tests of each of the >900 hits identified that none were significant (results not shown). Indeed, checking each gene individually, we noticed a trend towards NOA cases, but not significantly associated. This is perhaps concordant with the hypothesis that individual CpG sites or SNPs may not be sufficient to cause male infertility. Instead, a combination of such genes is required for perhaps cause NOA, as inferred from the machine learning algorithm results. Although a thorough investigation of combinations of the >900 hits has yet to be conducted, our preliminary investigation of the top 50 most important probes suggest interesting results. The methylation at the *RBMXL2* gene and SNP at the *RSRC1* gene is perhaps one combination to investigate further. Both genes are expressed in the testis and associated with pre-mRNAs in the nucleus. The arginine and serine rich coiled-coil 1 (*RSRC1*) gene is involved with alternative and constitutive splicing of mRNA. The RNA binding motif protein, X-linked 2 (*RBMXL2*) is a heterogenous nuclear ribonucleoprotein associated with the processing of pre-mRNAs. Thus, a specific SNP genotype of *RSRC1* in addition to the altered methylation of *RBMXL2* in NOA men may together push beyond a threshold which affects the pre-mRNA processing resulting in altered downstream pathways resulting in NOA. Further investigation of this pathway as well as other combinations from this dataset may be warranted.

A clear limitation of this study is the sample size, thus, caution must be taken in interpreting our results. This pilot study laid the foundation for the experimental design and analytical pipeline for continued future work for both a GWAS and integrated analysis of methylation and SNPs in infertile men. However, the current sample size is undoubtedly underpowered to identify even moderate effect sizes between groups as most previous studies have used larger sample sizes yet found false positives at a high rate (Aston, 2014). The continual addition of cases and annotation of phenotypes following this design will improve the ability to find results with higher confidence. Given that ethnicity and origin of geolocation of cases have been shown to affect the risk of SNPs for male infertility, the additional annotation of clinical information of our cases was sorely lacking. In a pilot study to circumvent this issue, we conducted an integrated analysis by combining methylation data in a subset of cases. Methylation data has been previously linked to ethnic origin and age of individuals, therefore the addition of methylation data as a proxy for clinical data may have improved our results. The integrated analysis design of combining both methylation and SNP calls and conducting a predictive and non-parametric approach to filtering genes may be a novel method in identifying the complex intricacies between SNP and methylation. The standard approach to integrated analysis is currently to conduct each dataset separately and identify overlapping genes identified in both. However, an alternative approach to interpreting this combined dataset may be to identify significant probes that target the same pathways. By building machine learning models which train using both datasets, higher accuracies may be identified with a combination of data types as shown in our pilot study. Furthermore, interesting relationships between genes may be presented that are not very obvious. These genes may then be studied in enrichment analyses, identifying pathways that are common. Currently, we have identified >900 which is a large number of hits. The continued addition of further samples may refine this number to more feasible investigations given that the current analysis is limited with only six cases.

CHAPTER 5: CLOSING REMARKS AND FUTURE DIRECTIONS

5.1 Summary and conclusions of studies

Infertility affects an estimated 1 in 6 couples worldwide (WHO), and approximately half of these cases can be attributed to the male partner. However, given that up to 58% of male infertility cases are idiopathic (Hamada et al., 2012), there is still much to learn about the causes of this disease of the reproductive system. With the development of advanced screening technologies for identifying the methylation at both targeted and genome-wide locations and the curation of previous work into public databases, it has now become possible to predict with a fair degree of confidence the genes or pathways potentially linked to male infertility. The major findings of this thesis suggest that methylation defects and point mutations/polymorphisms associated with meiotic/mitotic mechanisms in testicular cells may be a contributor to defective spermatogenesis. Moreover, we present evidence to suggest that methylation and SNP defects may work together to contribute to this disease. In Chapter 2, we examined the methylation at three imprinted DMRs -*H19, IG-GTL2, MEST* – in the sperm of 44 oligozoospermic infertile men and nine proven fertile controls. Of the infertile cases, 6.8% (3/44) displayed methylation defects at the *H19* and *MEST* DMRs, confirming the work from other research groups. The three cases belonged to the subgroup of infertile men with sperm concentrations ranging from 1-5 million sperm/mL, and not from the most severe subgroup. However, the same three cases were also found to be carriers of the T SNP at the *MTHFR* C677T genomic loci. This may suggest that infertile men who are severe oligozoospermia and also carriers of the *MTHFR* SNP may be at a higher risk for methylation defects in their sperm. Thus, the regulation of methyl supply from the folate cycle may be a mechanism associated with dysfunctional spermatogenesis. In addition, SNPs affecting the function of enzymes in the folate cycle may therefore be useful predictors of male infertility or methylation defects in the sperm.

In our study of genome-wide methylation in testicular biopsies from 10 NOA, six OA, and eight VR men, we observed 359 differential methylation signatures that when trained with machine learning models were capable of predicting the fertility of cases with up to 95.8% accuracy. Using novel gene selection algorithms developed in this study, two prospective gene sets were identified describing 60 and 299 CpG sites, respectively, with methylation signals discriminative for NOA cases. The methylation in the promoter of *NDE1* was 12.6% lower in NOA men as compared to VR men. NDE1 is a regulator of mitotic progression due to an interactive protein domain with mitotic spindle machinery and centromeres. Thus, the reduced methylation and anticipated higher expression of *NDE1* may indicate the activation of mitotic progression pathways in spermatogonia cells. This may be a compensation mechanism for the reduced or arrested sperm production indicative of NOA men. The *MCPH1* gene methylation was found to be >10.98% higher in NOA and OA men versus VR controls. Related to *NDE1*, *MCPH1* codes for DNA damage response genes and regulates the G2/M checkpoint. The presumed reduced expression of *MCPH1* in NOA and OA men might indicate a deficiency in MCPH1 resulting in an asynchronous entry into mitosis. This early entry decouples the centrosome and mitotic cycles resulting in arrest of the cells. MCPH1 is also known to recruit DNA repair proteins to recombination sites during meiosis. The lack of MCPH1 may therefore prevent meiotic progression resulting in arrested spermatocytes. This was the first indication of *NDE1* and *MCPH1* in male infertility, and suggest that the methylation of cell division genes may be an underlying cause in spermatogenic dysfunction. Other notable genes include a 2.12%

methylation reduction at the *TET3*, and 14.83% higher methylation at the *TEKT2* genes in NOA men. *TET3* is involved with the regulation methylation in germ cells. *TEKT2* regulates the microtubule structure in the sperm flagella. Both genes present candidate etiologies for the methylation defects observed in Chapter 2, and asthenozoospermia, respectively. Our study was also the first use of machine learning embedded algorithms for the selection of genes with methylation defects in male infertility, showing that such non-traditional approaches may be sufficient or improve upon traditional regression methods for epigenomic data mining.

In Chapter 4, we examined the genotype differences in a genome-wide fashion in seven NOA and 13 proven fertile control men. We identified 39 candidate SNPs whose genotypes were significantly different between groups. When machine learning classifiers were trained with these candidate SNPs, the cases in the study could be predicted with 100% accuracy. Although many of the SNP's functions are not known especially in the context of male infertility, noteworthy candidates with biological precedence was observed. We found that 100% of our NOA cases had the GG genotype at a SNP within the *MTR* gene (rs2385509). Control cases were observed to have this genotype in only 25% of the cases. The *MTR* gene codes for methionine synthase, which is involved with the folate cycle. As with the *MTHFR* C677T SNP, *MTR* SNPs have been previously implicated in male infertility in certain ethnic populations. The presumed role of an *MTR* SNP is also similar, which involves a reduction in enzyme efficiency thereby backing up the folate cycle and reducing the supply of methyl donors. This may result in reduced methylation reactions in testicular cells. From work within this thesis and other publications, methylation appears to be strongly associated with male fertility. However, the exact mechanisms remain unclear. In addition to *MTR*, we observed that 86% of NOA men in our study had the A allele at a SNP within the *NIN* gene (rs8017481), while only 23% of controls had this genotype. Ninein is suggested to be a testis protein involved with the centrosome complex and plays a role in the proper segregation of chromosomes during mitosis. Although *NIN* SNPs have not been previously associated with male infertility, the inferred disruption of the function of this protein is a plausible explanation for the arrest of cells during mitosis. As with other genome-wide association studies, the major limiting factor to our study was the sample size. As cases are added to the those presented in Chapter 4, the accuracy of the trained machine learning models will be improved. Nevertheless, resequencing studies in separate cohorts will be necessary to validate the list of SNPs from our study. Overall, compiling the results from Chapters 3 and 4, there appears to be evidence now to suggest that methylation and/or SNPs in meiotic/mitotic related genes that are expressed in the testis may be involved with male infertility.

In the latter part of Chapter 4, a pilot integrated analysis was conducted in four NOA and two VR cases overlapping from Chapters 3 and 4. By combining both methylation and SNP data, we identified a candidate list of 925 CpG/SNP loci that was capable of predicting the (limited number of) cases with 100% accuracy. An advantage of this integrated analysis includes the ability to proxy clinical data that may be latent or unavailable to the researcher given that methylation has been suggested to model characteristics such as age and ethnicity. This also provided the opportunity to identify methylation defects together with SNPs linked to male infertility. From Chapter 2, we found that methylation of a specific gene may be altered with a conditional probability of the presence of a SNP within the *MTHFR* gene. This suggested that numerous genes altered by methylation and/or SNPs together may be needed to sufficiently cause male infertility. A candidate pathway from our preliminary results suggest that the methylation at the *RBMXL2* gene and a SNP within the *RSRC1* gene may be an example of such a novel combination of genes/pathway, given that both genes are involved with pre-mRNA processing. As more samples are added to this integrated analysis, the predictive models may be able to identify a smaller yet more robust set of genes linked to male infertility. From there, novel pathways associated with defective spermatogenesis may be catalogued.

5.2 Closing remarks and future directions

The work presented in this thesis has only begun to investigate the relationship between methylation, SNPs, and male infertility. However, the finding of genes potentially involved with meiosis is exciting as it seems to be the most intuitive answer. There is yet a study having investigated the methylation at meiotic genes in relation to male infertility. We have presented novel candidate genes involved with mitosis or meiosis progression that may be epigenetically regulated. The next steps with such genes is to validate with targeted sequencing of the promoter regions to fully uncover the methylation and SNPs at these genes in infertile men with spermatogenic dysfunction. Such studies will confirm the role of these genes in male infertility.

We have presented a handful of novel feature selection algorithms to select genes. However, these algorithms are not perfect and will continue to improve with larger datasets. To further improve on the 450k methylation analysis, the addition of testicular cell composition methylation profiles can provide an estimation of cell type proportions in our study. The addition of such information directly into our models may further improve on the methylation signals due to true biological effects and thus improve on our confidence in our results. However, at the time of writing, these datasets do not exist. An additional opportunity that was also not yet available is to cross-reference our gene sets with Hi-C data. Hi-C is a chromosome conformation capture technique capable of identifying the genome-wide interactions between genomic loci. These studies have developed a 3D model of the genome in multiple cell types. Genes that may not be involved in the same pathway may still regulated by the same mechanisms due to close proximal locations in the 3D structure in the genome – pockets known as topologically associated chromatin domains (TADs) (Ong et al., 2014). Therefore, although the genes identified in our gene sets may not be associated together in biological pathways, their physical location within the genome may be together associated in topologically associating domains in the testicular cells. Thus, they may be mis/regulated together providing a topological etiology of misregulation between genes. However, at the time of writing, there are no testis Hi-C datasets available. When available in future work, the genes identified in our gene sets should be cross matched with Hi-C data to determine whether any are within the same TAD.

We also presented a novel pilot study using an integrated analysis combining both methylation and SNP data. At the time of writing, there is yet a study using such an approach in male infertility. The value of such a method is the ability to train ML models using both data and potentially identify a subset of features, either methylation or SNP, that may be capable of predicting fertility. This has implications for further understanding the relationship between genetics and epigenetics in diseases as shown in Chapter 2. The addition of further samples to this pilot study will refine and stabilize the gene set. As of this thesis, the number of potential features is >900, which is infeasible to manually sift through.

The clinically relevant significance of our work is the establishment of analytical pipelines potentially ready for clinical diagnosis. As microarray costs become cheaper, individuals may opt for investigation of their genome. By training ML models with a continued growing set of infertility cases, prediction accuracy will improve and the identification of underlying genes may be further refined. Thus, the identification of biomarkers readily available for clinical diagnosis can help shape future fertility practice and decision making by patients.

Bibliography

1. Aarabi M, San Gabriel MC, Chan D, Behan NA, Caron M, Pastinen T, Bourque G, MacFarlane AJ, Zini A, Trasler J. High-dose folic acid supplementation alters the human sperm methylome and is influenced by the MTHFR C677T polymorphism. Human molecular genetics. 2015 Aug 24;24(22):6301-13.
2. Abu-Halima M, Hammadeh M, Backes C, Fischer U, Leidinger P, Lubbad AM, Keller A, Meese E. Panel of five microRNAs as potential biomarkers for the diagnosis and assessment of male infertility. Fertility and sterility. 2014 Oct 31;102(4):989-97.
3. Adelman CA, Petrini JH. ZIP4H (TEX11) deficiency in the mouse impairs meiotic double strand break repair and the regulation of crossing over. PLoS genetics. 2008 Mar 28;4(3):e1000042.
4. Agarwal A, Mulgund A, Hamada A, Chyatte MR. A unique view on male infertility around the globe. Reproductive Biology and Endocrinology. 2015 Dec 1;13(1):37.
5. Albrecht KH, Eicher EM. Evidence that Sry is expressed in pre-Sertoli cells and Sertoli and granulosa cells have a common precursor. Developmental biology. 2001 Dec 1;240(1):92-107.
6. Alkuraya FS, Cai X, Emery C, Mochida GH, Al-Dosari MS, Felie JM, Hill RS, Barry BJ, Partlow JN, Gascon GG, Kentab A. Human mutations in NDE1 cause extreme microcephaly with lissencephaly. The American Journal of Human Genetics. 2011 May 13;88(5):536-47.
7. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR. A balanced iterative random forest for gene selection from microarray data. BMC bioinformatics. 2013 Aug 27;14(1):261.
8. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014 Jan 28;30(10):1363-9.

9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. Gene Ontology: tool for the unification of biology. Nature genetics. 2000 May 1;25(1):25-9.
10. Aston KI, Carrell DT. Genome-wide study of single-nucleotide polymorphisms associated with azoospermia and severe oligozoospermia. Journal of andrology. 2009 Nov 12;30(6):711-25.
11. Aston KI, Krausz C, Laface I, Ruiz-Castane E, Carrell DT. Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of European descent. Human Reproduction. 2010 Apr 8;25(6):1383-97.
12. Aston KI. Genetic susceptibility to male infertility: news from genome-wide association studies. Andrology. 2014 May 1;2(3):315-21.
13. Atsem S, Reichenbach J, Potabattula R, Dittrich M, Nava C, Depienne C, Böhm L, Rost S, Hahn T, Schorsch M, Haaf T. Paternal age effects on sperm FOXK1 and KCNA7 methylation and transmission into the next generation. Human molecular genetics. 2016 Nov 15;25(22):4996-5005.
14. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nature Reviews Genetics. 2013 Feb 1;14(2):113-24.
15. Bakircioglu M, Carvalho OP, Khurshid M, Cox JJ, Tuysuz B, Barak T, Yilmaz S, Caglayan O, Dincer A, Nicholas AK, Quarrell O. The essential role of centrosomal NDE1 in human cerebral cortex neurogenesis. The American Journal of Human Genetics. 2011 May 13;88(5):523-35.
16. Bamezai S, Rawat VP, Buske C. Concise Review: The Piwi-piRNA Axis: Pivotal Beyond Transposon Silencing. Stem Cells. 2012 Dec 1;30(12):2603-11.
17. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. Cold Spring Harbor perspectives in biology. 2014 Feb 1;6(2):a018382.
18. Barton SC, Ferguson-Smith AC, Fundele RE, Surani MA. Influence of paternally imprinted genes on development. Development. 1991 Oct 1;113(2):679-87.

19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological). 1995 Jan 1:289-300.
20. Bhasin S, De Kretser DM, Baker HW. Clinical review 64: Pathophysiology and natural history of male infertility. The Journal of Clinical Endocrinology & Metabolism. 1994 Dec 1;79(6):1525-9.
21. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005).
22. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986 May 15;321(6067):209-13.
23. Boissonnas CC, El Abdalaoui H, Haelewyn V, Fauque P, Dupont JM, Gut I, Vaiman D, Jouannet P, Tost J, Jammes H. Specific epigenetic alterations of IGF2-H19 locus in spermatozoa from infertile men. European Journal of Human Genetics. 2010 Jan 1;18(1):73-80.
24. Boissonnas CC, El Abdalaoui H, Haelewyn V, Fauque P, Dupont JM, Gut I, Vaiman D, Jouannet P, Tost J, Jammes H. Specific epigenetic alterations of IGF2-H19 locus in spermatozoa from infertile men. European Journal of Human Genetics. 2010 Jan 1;18(1):73-80.
25. Bojesen A, Juul S, Gravholt CH. Prenatal and postnatal prevalence of Klinefelter syndrome: a national registry study. The Journal of Clinical Endocrinology & Metabolism. 2003 Feb 1;88(2):622-6.
26. Borgmann J, Tüttelmann F, Dworniczak B, Röpke A, Song HW, Kliesch S, Wilkinson MF, Laurentino S, Gromoll J. The human RHOX gene cluster: target genes and functional analysis of gene variants in infertile men. Human molecular genetics. 2016 Nov 15;25(22):4898-910.
27. Bowdin S, Allen C, Kirby G, Brueton L, Afnan M, Barratt C, Kirkman-Brown J, Harrison R, Maher ER, Reardon W. A survey of assisted reproductive technology births and imprinting disorders. Human Reproduction. 2007 Oct 5;22(12):3237-40.

28. Broen JC, Radstake TR. How birds of a feather flock together: genetics in autoimmune diseases. Expert review of clinical immunology. 2011 Mar 1;7(2):127-8.
29. Camprubí C, Pladevall M, Grossmann M, Garrido N, Pons MC & Blanco J. Lack of association of MTHFR rs1801133 polymorphism and CTCFL mutations with sperm methylation errors in infertile patients. 2013. J Assist Reprod Genet. 30, 1125–1131.
30. Camprubí C, Salas-Huetos A, Aiese-Cigliano R, Godo A, Pons MC, Castellano G, Grossmann M, Sanseverino W, Martin-Subero JI, Garrido N, Blanco J. Spermatozoa from infertile patients exhibit differences of DNA methylation associated with spermatogenesis-related processes: an array-based analysis. Reproductive biomedicine online. 2016 Dec 31;33(6):709-19.
31. Cariboni A, Pimpinelli F, Colamarino S, Zaninetti R, Piccolella M, Rumio C, Piva F, Rugarli EI, Maggi R. The product of X-linked Kallmann's syndrome gene (KAL1) affects the migratory activity of gonadotropin-releasing hormone (GnRH)-producing neurons. Human molecular genetics. 2004 Oct 7;13(22):2781-91.
32. Carrell DT, Aston KI. The search for SNPs, CNVs, and epigenetic variants associated with the complex disease of male infertility. Systems biology in reproductive medicine. 2011 Jan 1;57(1-2):17-26.
33. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell. 2009 Feb 20;136(4):642-55.
34. Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics. 2006 Dec 22;8(2):485-99.
35. Carvalho BS, Louis TA, Irizarry RA. Quantifying uncertainty in genotype calls. Bioinformatics. 2009 Nov 11;26(2):242-9.
36. Carvalho BS, Louis TA, Irizarry RA. Quantifying uncertainty in genotype calls. Bioinformatics. 2010 Jan 15;26(2):242-9.
37. Castro R, Rivera I, Ravasco P, Camilo ME, Jakobs C, Blom HJ, De Almeida IT. 5, 10-methylenetetrahydrofolate reductase (MTHFR) 677C→ T and 1298A→ C mutations are associated with DNA hypomethylation. Journal of medical genetics. 2004 Jun 1;41(6):454-8.

38. Charlesworth D, Charlesworth B. Sex chromosomes: evolution of the weird and wonderful. Current Biology. 2005 Feb 22;15(4):R129-31.
39. Chen CH, Howng SL, Cheng TS, Chou MH, Huang CY, Hong YR. Molecular characterization of human ninein protein: two distinct subdomains required for centrosomal targeting and regulating signals in cell cycle. Biochemical and biophysical research communications. 2003 Sep 5;308(4):975-83.
40. Chen H, Ruan YC, Xu WM, Chen J, Chan HC. Regulation of male fertility by CFTR and implications in male infertility. Human reproduction update. 2012 Jun 17;18(6):703-13.
41. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013 Feb 1;8(2):203-9.
42. Chen Z, Karaplis AC, Ackerman SL, Pogribny IP, Melnyk S, Lussier-Cacan S, Chen MF, Pai A, John SW, Smith RS, Bottiglieri T, Bagley P, Selhub J, Rudnicki MA, James SJ & Rozen R. (2001) Mice deficient in methylenetetrahydrofolate reductase exhibit hyperhomocysteinemia and decreased methylation capacity, with neuropathology and aortic lipid deposition. Hum Mol Genet 10, 433–443.
43. Chen, Z., Karaplis, A.C., Ackerman, S.L., Pogribny, I.P., Melnyk, S., Lussier-Cacan, S., Chen, M.F., Pai, A., John, S.W., Smith, R.S. and Bottiglieri, T., 2001. Mice deficient in methylenetetrahydrofolate reductase exhibit hyperhomocysteinemia and decreased methylation capacity, with neuropathology and aortic lipid deposition. Human molecular genetics, 10(5), pp.433-444.
44. Chianese C, Gunning AC, Giachini C, Daguin F, Balercia G, Ars E, Giacco DL, Ruiz-Castañé E, Forti G, Krausz C. X chromosome-linked CNVs in male infertility: discovery of overall duplication load and recurrent, patient-specific gains with potential clinical relevance. PLoS One. 2014 Jun 10;9(6):e97746.
45. Costa-Barbosa FA, Balasubramanian R, Keefe KW, Shaw ND, Al-Tassan N, Plummer L, Dwyer AA, Buck CL, Choi JH, Seminara SB, Quinton R. Prioritizing

genetic testing in patients with Kallmann syndrome using clinical phenotypes. The Journal of Clinical Endocrinology & Metabolism. 2013 Mar 26;98(5):E943-53.

46. Cousineau TM, Domar AD. Psychological impact of infertility. Best Practice & Research Clinical Obstetrics & Gynaecology. 2007 Apr 30;21(2):293-308.
47. Cutter AR, Hayes JJ. A brief review of nucleosome structure. FEBS letters. 2015 Oct 7;589(20PartA):2914-22.
48. Dalgaard MD, Weinhold N, Edsgärd D, Silver JD, Pers TH, Nielsen JE, Jørgensen N, Juul A, Gerds TA, Giwercman A, Giwercman YL. A genome-wide association study of men with symptoms of testicular dysgenesis syndrome and its network biology interpretation. Journal of medical genetics. 2011 Jan 1:jmedgenet-2011.
49. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071
50. Davis TL, Yang GJ, McCarrey JR, Bartolomei MS. The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. Human molecular genetics. 2000 Nov 22;9(19):2885-94.
51. de la Calle JF, Rachou E, le Martelot MT, Ducot B, Multigner L, Thonneau PF. Male infertility risk factors in a French military population. Human Reproduction. 2001 Mar 1;16(3):481-6.
52. Dejong J. SVM with recursive feature elimination in R [Internet]. 2016 [cited 29 December 2017]. Available from: https://johanndejong.wordpress.com/2016/01/17/svm-with-recursive-feature-elimination/
53. Delgehyr N, Sillibourne J, Bornens M. Microtubule nucleation and anchoring at the centrosome are independent processes linked by ninein function. J Cell Sci. 2005 Apr 15;118(8):1565-75.
54. Denomme MM, McCallie BR, Parks JC, Schoolcraft WB, Katz-Jaffe MG. Alterations in the sperm histone-retained epigenome are associated with unexplained male factor

infertility and poor blastocyst development in donor oocyte IVF cycles. Human Reproduction. 2017 Oct 26:1-3.

55. Dere E, Huse S, Hwang K, Sigman M, Boekelheide K. Intra-and inter-individual differences in human sperm Dna methylation. Andrology. 2016 Sep 1;4(5):832-42.
56. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006 Dec;7(1):3.
57. Doerksen T, Benoit G, Trasler JM. Deoxyribonucleic acid hypomethylation of male germ cells by mitotic and meiotic exposure to 5-azacytidine is associated with altered testicular histology. Endocrinology. 2000 Sep 1;141(9):3235-44.
58. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics. 2010 Nov 30;11(1):587.
59. Duda RO, Hart PE, Stork DG. Pattern classification. Wiley, New York; 1973 Nov.
60. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association. 2002 Mar 1;97(457):77-87.
61. Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavaré S, Ritchie ME. Statistical issues in the analysis of Illumina data. BMC bioinformatics. 2008 Feb 6;9(1):85.
62. Eddy EM. Male germ cell gene expression. Recent progress in hormone research. 2002 Jan 1;57(1):103-28.
63. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002 Jan 1;30(1):207-10.
64. El Hajj N, Zechner U, Schneider E, Tresch A, Gromoll J, Hahn T, Schorsch M, Haaf T. Methylation status of imprinted genes and repetitive elements in sperm DNA from infertile males. Sexual Development. 2011;5(2):60-9.
65. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology. 2011 Apr 19;9(4):e1001046.
66. Everson TM, Lyons G, Zhang H, Soto-Ramírez N, Lockett GA, Patil VK, Merid SK, Söderhäll C, Melén E, Holloway JW, Arshad SH. DNA methylation loci associated

with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. Genome medicine. 2015 Aug 21;7(1):89.

67. Feinberg JI, Bakulski KM, Jaffe AE, Tryggvadottir R, Brown SC, Goldman LR, Croen LA, Hertz-Picciotto I, Newschaffer CJ, Daniele Fallin M, Feinberg AP. Paternal sperm DNA methylation associated with early signs of autism risk in an autism-enriched cohort. International journal of epidemiology. 2015 Apr 14;44(4):1199-210.
68. Feng Y, Walsh CA. Mitotic spindle regulation by Nde1 controls cerebral cortical size. Neuron. 2004 Oct 14;44(2):279-93.
69. Ferfouri F, Boitrelle F, Ghout I, Albert M, Molina Gomes D, Wainer R, Bailly M, Selva J, Vialard F. A genome-wide DNA methylation study in azoospermia. Andrology. 2013 Nov 1;1(6):815-21.
70. Ferguson KA, Wong EC, Chow V, Nigro M, Ma S. Abnormal meiotic recombination in infertile men and its association with sperm aneuploidy. Human molecular genetics. 2007 Aug 29;16(23):2870-9.
71. Fodor SP. Massively parallel genomics. Science. 1997 Jul 18;277(5324):393.
72. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, Greenwood CM, Hansen KD. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome biology. 2014 Dec 3;15(11):503.
73. Fowden AL, Sibley C, Reik W, Constancia M. Imprinted genes, placental development and fetal growth. Hormone Research in Paediatrics. 2006;65(Suppl. 3):50-8.
74. Friso S, Choi SW, Girelli D, Mason JB, Dolnikowski GG, Bagley PJ, Olivieri O, Jacques PF, Rosenberg IH, Corrocher R, Selhub J. A common mutation in the 5, 10-methylenetetrahydrofolate reductase gene affects genomic DNA methylation through an interaction with folate status. Proceedings of the National Academy of Sciences. 2002 Apr 16;99(8):5606-11.
75. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJ, Den Heijer M, Kluijtmans LA, Van Den Heuve LP, Rozen R. A candidate genetic risk

factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nature genetics. 1995 May 1;10(1):111-3.

76. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000 Oct 1;16(10):906-14.
77. Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic acids research. 2017 Jan 4;45(D1):D331-8.
78. Gene Ontology Consortium. Gene ontology consortium: going forward. Nucleic acids research. 2015 Jan 28;43(D1):D1049-56.
79. Giacco DL, Chianese C, Ars E, Ruiz-Castañé E, Forti G, Krausz C. Recurrent X chromosome-linked deletions: discovery of new genetic factors in male infertility. Journal of medical genetics. 2014 Jan 11:jmedgenet-2013.
80. Gong M, Dong W, He T, Shi Z, Huang G, Ren R, Huang S, Qiu S, Yuan R. MTHFR 677C> T polymorphism increases the male infertility risk: a meta-analysis involving 26 studies. PloS one. 2015 Mar 20;10(3):e0121147.
81. Gordetsky J, van Wijngaarden E, O'brien J. Redefining abnormal follicle-stimulating hormone in the male infertility population. BJU international. 2012 Aug 1;110(4):568-72.
82. Goyette P, Frosst P, Rosenblatt DS, Rozen R. Seven novel mutations in the methylenetetrahydrofolate reductase gene and genotype/phenotype correlations in severe methylenetetrahydrofolate reductase deficiency. American journal of human genetics. 1995 May;56(5):1052.
83. Grafodatskaya D, Cytrynbaum C, Weksberg R. The health risks of ART. EMBO reports. 2013 Feb 1;14(2):129-35.
84. Griswold MD. The central role of Sertoli cells in spermatogenesis. InSeminars in cell & developmental biology 1998 Aug 1 (Vol. 9, No. 4, pp. 411-416). Academic Press.
85. Gruber R, Zhou Z, Sukchev M, Joerss T, Frappart PO, Wang ZQ. MCPH1 regulates the neuroprogenitor division mode by coupling the centrosomal cycle with mitotic entry through the Chk1-Cdc25 pathway. Nature cell biology. 2011 Nov 1;13(11):1325-34.

86. Gupta N, Gupta S, Dama M, David A, Khanna G, Khanna A, Rajender S. Strong association of 677 C> T substitution in the MTHFR gene with male infertility-a study on an Indian population and a meta-analysis. PloS one. 2011 Jul 20;6(7):e22277.
87. Guyon I, Matic N, Vapnik V. Discovering Informative Patterns and Data Cleaning. 1996. 181-203.
88. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002 Jan 1;46(1):389-422.
89. Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA, Surani MA. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science. 2013 Jan 25;339(6118):448-52.
90. Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, Walter J, Surani MA. Epigenetic reprogramming in mouse primordial germ cells. Mechanisms of development. 2002 Sep 30;117(1):15-23.
91. Hall MA. Correlation-based feature selection for machine learning.
92. Hamada A, Esteves SC, Nizza M, Agarwal A. Unexplained male infertility: diagnosis and management. International braz j urol. 2012 Oct;38(5):576-94.
93. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. Distinctive chromatin in human sperm packages genes for embryo development. Nature. 2009 Jul 23;460(7254):473-8.
94. Hanley NA, Hagan DM, Clement-Jones M, Ball SG, Strachan T, Salas-Cortes L, McElreavey K, Lindsay S, Robson S, Bullen P, Ostrer H. SRY, SOX9, and DAX1 expression patterns during human sex determination and gonadal development. Mechanisms of development. 2000 Mar 1;91(1):403-7.
95. Hansen KD and Aryee M. (2012). IlluminaHumanMethylation450kmanifest: Annotation for Illumina's 450k methylation arrays. R package version 0.4.0.
96. Hansen LK, Salamon P. Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence. 1990 Oct;12(10):993-1001.
97. Harper KN, Peters BA, Gamble MV. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. Cancer Epidemiology and Prevention Biomarkers. 2013 Jun 1;22(6):1052-60.

98. Hemberger M, Dean W, Reik W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nature reviews Molecular cell biology. 2009 Aug 1;10(8):526-37.

99. Hill PW, Amouroux R, Hajkova P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: an emerging complex story. Genomics. 2014 Nov 30;104(5):324-33.

100. Hiort O, Holterhus PM, Horter T, Schulze W, Kremke B, Bals-Pratsch M, Sinnecker GH, Kruse K. Significance of mutations in the androgen receptor gene in males with idiopathic infertility. The Journal of Clinical Endocrinology & Metabolism. 2000 Aug 1;85(8):2810-5.

101. Hiraoka M, Kagawa Y. Genetic polymorphisms and folate status. Congenital Anomalies. 2017 Jun 9.

102. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC. Integrative annotation of chromatin elements from ENCODE data. Nucleic acids research. 2012 Dec 5;41(2):827-41.

103. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013 Oct;14(10):3156.

104. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics. 2012 May 8;13(1):86.

105. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014 Jan 21;30(10):1431-9.

106. Hu Z, Xia Y, Guo X, Dai J, Li H, Hu H, Jiang Y, Lu F, Wu Y, Yang X, Li H. A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia. Nature genetics. 2012 Feb;44(2):183.

107. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2008 Nov 25;37(1):1-3.

108. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R. Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods. 2015 Feb 1;12(2):115-21.

109. Hughes VL, Randolph SE. Testosterone depresses innate and acquired resistance to ticks in natural rodent hosts: a force for aggregated distributions of parasites. Journal of Parasitology. 2001 Feb;87(1):49-54.

110. Hwang K, Yatsenko AN, Jorgez CJ, Mukherjee S, Nalam RL, Matzuk MM, Lamb DJ. Mendelian genetics of male infertility. Annals of the New York Academy of Sciences. 2010 Dec 1;1214(1).

111. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr 1;4(2):249-64.

112. Irvine DS. Epidemiology and aetiology of male infertility. Human reproduction. 1998 Apr 1;13(suppl_1):33-44.

113. Jabbari K, Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene. 2004 May 26;333:143-9.

114. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome biology. 2014 Feb 4;15(2):R31.

115. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International journal of epidemiology. 2012 Feb 1;41(1):200-9.

116. Jäger J, Sengupta R, Ruzzo WL. Improved gene selection for classification of microarrays. InPacific Symposium on Biocomputing 2002 Dec (Vol. 8, pp. 53-64).

117. Jameel T. Sperm swim-up: a simple and effective technique of semen processing for intrauterine insemination. JPMA. The Journal of the Pakistan Medical Association. 2008 Feb;58(2):71.

118. Jenkins TG, Carrell DT. The sperm epigenome and potential implications for the developing embryo. Reproduction. 2012 Jun 1;143(6):727-34.

119. Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature. 2007 Sep 13;449(7159):248-51.

120. Jirtle R. Imprinted Genes: by Species [Internet]. geneimprint. 2017 [cited 26 December 2017]. Available from: http://www.geneimprint.com/site/genes-by-species

121. Jones PA, Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation. Cell. 1980 May 1;20(1):85-93.

122. Jones PL, Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nature genetics. 1998 Jun 1;19(2):187-91.

123. Kafri T, Ariel M, Brandeis M, Shemer R, Urven L, McCarrey J, Cedar H, Razin A. Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes & development. 1992 May 1;6(5):705-14.

124. Kaludov NK, Wolffe AP. MeCP2 driven transcriptional repression in vitro: selectivity for methylated DNA, action at a distance and contacts with the basal transcription machinery. Nucleic acids research. 2000 May 1;28(9):1921-8.

125. Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, Sasaki H. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. nature. 2004 Jun 24;429(6994):900-3.

126. Kappil MA, Green BB, Armstrong DA, Sharp AJ, Lambertini L, Marsit CJ, Chen J. Placental expression profile of imprinted genes impacts birth weight. Epigenetics. 2015 Sep 2;10(9):842-9.

127. Kashimada K, Koopman P. Sry: the master switch in mammalian sex determination. Development. 2010 Dec 1;137(23):3921-30.

128. Katis VL, Lipp JJ, Imre R, Bogdanova A, Okaz E, Habermann B, Mechtler K, Nasmyth K, Zachariae W. Rec8 phosphorylation by casein kinase 1 and Cdc7-Dbf4 kinase regulates cohesin cleavage by separase during meiosis. Developmental cell. 2010 Mar 16;18(3):397-409.

129. Kelly TL, Li EN, Trasler JM. 5-Aza-2′-Deoxycytidine Induces Alterations in Murine Spermatogenesis and Pregnancy Outcome. Journal of andrology. 2003 Nov 12;24(6):822-30.

130. Kelly TL, Li EN, Trasler JM. 5-Aza-2′-Deoxycytidine Induces Alterations in Murine Spermatogenesis and Pregnancy Outcome. Journal of andrology. 2003 Nov 12;24(6):822-30.

131. Khazamipour N, Noruzinia M, Fatehmanesh P, Keyhanee M, Pujol P. MTHFR promoter hypermethylation in testicular biopsies of patients with non-obstructive azoospermia: the role of epigenetics in male infertility. Human reproduction. 2009 May 28;24(9):2361-4.

132. Khazamipour N, Noruzinia M, Fatehmanesh P, Keyhanee M, Pujol P. MTHFR promoter hypermethylation in testicular biopsies of patients with non-obstructive azoospermia: the role of epigenetics in male infertility. Human reproduction. 2009 May 28;24(9):2361-4.

133. Kim S, Zaghloul NA, Bubenshchikova E, Oh EC, Rankin S, Katsanis N, Obara T, Tsiokas L. Nde1-mediated inhibition of ciliogenesis affects cell cycle re-entry. Nature cell biology. 2011 Apr 1;13(4):351-60.

134. Kobayashi H, Hiura H, John RM, Sato A, Otsu E, Kobayashi N, Suzuki R, Suzuki F, Hayashi C, Utsunomiya T, Yaegashi N. DNA methylation errors at imprinted loci after assisted conception originate in the parental sperm. European journal of human genetics. 2009 Dec 1;17(12):1582-91.

135. Kobayashi H, Sato A, Otsu E, Hiura H, Tomatsu C, Utsunomiya T, Sasaki H, Yaegashi N, Arima T. Aberrant DNA methylation of imprinted loci in sperm from oligospermic patients. Human molecular genetics. 2007 Jul 17;16(21):2542-51.

136. Kouzarides T. Chromatin modifications and their function. Cell. 2007 Feb 23;128(4):693-705.

137. Krausz C, Chianese C. Genetic testing and counselling for male infertility. Current Opinion in Endocrinology, Diabetes and Obesity. 2014 Jun 1;21(3):244-50.

138. Krausz C, Escamilla AR, Chianese C. Genetics of male infertility: from research to clinic. Reproduction. 2015 Nov 1;150(5):R159-74.

139. Krausz C, Giachini C, Giacco DL, Daguin F, Chianese C, Ars E, Ruiz-Castane E, Forti G, Rossi E. High resolution X chromosome-specific array-CGH detects new CNVs in infertile males. PloS one. 2012 Oct 9;7(10):e44887.

140. Krausz C. Male infertility: pathogenesis and clinical diagnosis. Best practice & research Clinical endocrinology & metabolism. 2011 Apr 1;25(2):271-85.

141. Kurzawski M, Wajda A, Malinowski D, Kazienko A, Kurzawa R, Drozdzik M. Association study of folate-related enzymes (MTHFR, MTR, MTRR) genetic variants with non-obstructive male infertility in a Polish population. Genetics and molecular biology. 2015 Mar;38(1):42-7.

142. Langley P. Selection of relevant features in machine learning. InProceedings of the AAAI Fall symposium on relevance 1994 Nov 4 (Vol. 184, pp. 245-271).

143. Laqqan M, Hammadeh ME. Alterations in DNA methylation patterns and gene expression in spermatozoa of subfertile males. Andrologia. 2017.

144. Laqqan M, Tierling S, Alkhaled Y, Lo Porto C, Solomayer EF, Hammadeh M. Spermatozoa from males with reduced fecundity exhibit differential DNA methylation patterns. Andrology. 2017 May 23.

145. Laurent Gatto (2017). hpar: Human Protein Atlas in R. R package version 1.18.1.

146. Laurentino S, Beygo J, Nordhoff V, Kliesch S, Wistuba J, Borgmann J, Buiting K, Horsthemke B, Gromoll J. Epigenetic germline mosaicism in infertile men. Human molecular genetics. 2014 Oct 21;24(5):1295-304.

147. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, Zhang Y and Torres LC (2017). sva: Surrogate Variable Analysis. R package version 3.24.4.

148. Leonhardt H, Page AW, Weier HU, Bestor TH. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell. 1992 Nov 27;71(5):865-73.

149. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992 Jun 12;69(6):915-26.

150. Li J, Ching T, Huang S, Garmire LX. Using epigenomics data to predict gene expression in lung cancer. BMC bioinformatics. 2015 Mar 18;16(5):S10.

151. Liang Y, Gao H, Lin SY, Peng G, Huang X, Zhang P, Goss JA, Brunicardi FC, Multani AS, Chang S, Li K. BRIT1/MCPH1 is essential for mitotic and meiotic recombination DNA repair and maintaining genomic stability in mice. PLoS genetics. 2010 Jan 22;6(1):e1000826.

152. Liaw, A. and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

153. Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC bioinformatics. 2004 Sep 27;5(1):136.

154. Liu H, Rankin S, Yu H. Phosphorylation-enabled binding of SGO1–PP2A to cohesin protects sororin and centromeric cohesion during mitosis. Nature cell biology. 2013 Jan 1;15(1):40-9.

155. Long PM, Vega VB. Boosting and microarray data. Machine Learning. 2003 Jul 1;52(1):31-44.

156. Louie K, Minor A, Ng R, Poon K, Chow V, Ma S. Evaluation of DNA methylation at imprinted DMRs in the spermatozoa of oligozoospermic men in association with MTHFR C677T genotype. Andrology. 2016 Sep 1;4(5):825-31.

157. Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ. Computational and experimental identification of novel human imprinted genes. Genome research. 2007 Dec 1;17(12):1723-30.

158. Luo L, Chen H, Zirkin BR. Leydig Cell Aging: Steroidogenic Acute Regulatory Protein (StAR) and Cholesterol Side-Chain Cleavage Enzyme. Journal of andrology. 2001 Jan 2;22(1):149-56.

159. Lynch AG, Dunning MJ, Iddawela M, Barbosa-Morais NL, Ritchie ME. Considerations for the processing and analysis of GoldenGate-based two-colour Illumina platforms. Statistical methods in medical research. 2009 Oct;18(5):437-52.

160. Mackay DJ, Boonen SE, Clayton-Smith J, Goodship J, Hahnemann JM, Kant SG, Njølstad PR, Robin NH, Robinson DO, Siebert R, Shield JP. A maternal hypomethylation syndrome presenting as transient neonatal diabetes mellitus. Human genetics. 2006 Sep 1;120(2):262-9.

161. Mackay DJ, Callaway JL, Marks SM, White HE, Acerini CL, Boonen SE, Dayanikli P, Firth HV, Goodship JA, Haemers AP, Hahnemann JM. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. Nature genetics. 2008 Aug 1;40(8):949-51.

162. Maher ER, Afnan M, Barratt CL. Epigenetic risks related to assisted reproductive technologies: epigenetics, imprinting, ART and icebergs?. Human Reproduction. 2003 Dec 1;18(12):2508-11.

163. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome biology. 2012 Jun 15;13(6):R44.

164. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).

165. Marques CJ, Carvalho F, Sousa M, Barros A. Genomic imprinting in disruptive spermatogenesis. The lancet. 2004 May 22;363(9422):1700-2.

166. Marques CJ, Costa P, Vaz B, Carvalho F, Fernandes S, Barros A, Sousa M. Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia. MHR-Basic Science of Reproductive Medicine. 2008 Jan 4;14(2):67-74.

167. Marques CJ, Costa P, Vaz B, Carvalho F, Fernandes S, Barros A, Sousa M. Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia. MHR-Basic Science of Reproductive Medicine. 2008 Jan 4;14(2):67-74.

168. Marques CJ, Francisco T, Sousa S, Carvalho F, Barros A, Sousa M. Methylation defects of imprinted genes in human testicular spermatozoa. Fertility and sterility. 2010 Jul 31;94(2):585-94.

169. Meethal SV, Atwood CS. The role of hypothalamic-pituitary-gonadal hormones in the normal structure and functioning of the brain. Cell Mol Life Sci. 2005 Feb 1;62(3):257-70.

170. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008 Aug 7;454(7205):766-70.

171. Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB, Sidransky D. 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nature medicine. 1995 Jul 1;1(7):686-92.

172. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes & development. 2014 Apr 15;28(8):812-28.

173. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic acids research. 2017 Jan 4;45(D1):D183-9.

174. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic acids research. 2012 Nov 26;41(D1):D377-86.

175. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, Houseman EA, Izzi B, Kelsey KT, Meissner A, Milosavljevic A. Recommendations for the design and analysis of epigenome-wide association studies. Nature methods. 2013 Oct 1;10(10):949-55.

176. Minor A, Chow V, Ma S. Aberrant DNA methylation at imprinted genes in testicular sperm retrieved from men with obstructive azoospermia and undergoing vasectomy reversal. Reproduction. 2011 Jun 1;141(6):749-57.

177. Mogensen MM, Malik A, Piel M, Bouckson-Castaing V, Bornens M. Microtubule minus-end anchorage at centrosomal and non-centrosomal sites: the role of ninein. J Cell Sci. 2000 Sep 1;113(17):3013-23.

178. Moore T, Haig D. Genomic imprinting in mammalian development: a parental tug-of-war. Trends in Genetics. 1991 Feb 1;7(2):45-9.

179. Morris TJ, Beck S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. Methods. 2015 Jan 15;72:3-8.

180. Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. Independent specialization of the human and mouse X chromosomes for the male germ line. Nature genetics. 2013 Sep 1;45(9):1083-7.

181. Naqvi H, Hussain SR, Ahmad MK, Mahdi F, Jaiswar SP, Shankhwar SN, Mahdi AA. Role of 677C→ T polymorphism a single substitution in methylenetetrahydrofolate reductase (MTHFR) gene in North Indian infertile men. Molecular biology reports. 2014 Feb 1;41(2):573-9.

182. Newton CR, Sherrard W, Glavac I. The Fertility Problem Inventory: measuring perceived infertility-related stress. Fertility and sterility. 1999 Jul 31;72(1):54-62.

183. Ni K, Dansranjavin T, Rogenhofer N, Oeztuerk N, Deuker J, Bergmann M, Schuppe HC, Wagenlehner F, Weidner W, Steger K, Schagdarsurengin U. TET enzymes are successively expressed during human spermatogenesis and their expression level is pivotal for male fertility. Human Reproduction. 2016 May 1;31(7):1411-24.

184. Nordin M, Bergman D, Halje M, Engström W, Ward A. Epigenetic regulation of the Igf2/H19 gene cluster. Cell proliferation. 2014 Jun 1;47(3):189-99.

185. O’Donnell L. Mechanisms of spermiogenesis and spermiation and how they are disturbed. Spermatogenesis. 2014 Mar 4;4(2):e979623.

186. O'Brien KL, Varghese AC, Agarwal A. The genetic causes of male factor infertility: a review. Fertility and sterility. 2010 Jan 1;93(1):1-2.

187. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999 Oct 29;99(3):247-57.

188. Okutman O, Muller J, Baert Y, Serdarogullari M, Gultomruk M, Piton A, Rombaut C, Benkhalifa M, Teletin M, Skory V, Bakircioglu E. Exome sequencing reveals a nonsense mutation in TEX15 causing spermatogenic failure in a Turkish family. Human molecular genetics. 2015 Jul 21;24(19):5581-8.

189. Olek A, Walter J. The pre-implantation ontogeny of the H19 methylation imprint. Nature genetics. 1997 Nov 1;17(3):275-6.

190. Oliva A, Spira A, Multigner L. Contribution of environmental factors to the risk of male infertility. Human Reproduction. 2001 Aug 1;16(8):1768-76.

191. Oliva R, Margarit E, Ballescá JL, Carrió A, Sánchez A, Milà M, Jiménez L, Alvarez-Vijande JR, Ballesta F. Prevalence of Y chromosome microdeletions in oligospermic and azoospermic candidates for intracytoplasmic sperm injection. Fertility and sterility. 1998 Sep 30;70(3):506-10.

192. Oliva R. Protamines and male infertility. Human reproduction update. 2006 Mar 31;12(4):417-35.

193. O'Moore AM, O'Moore RR, Harrison RF, Murphy G, Carruthers ME. Psychosomatic aspects in idiopathic infertility: effects of treatment with autogenic training. Journal of Psychosomatic Research. 1983 Dec 31;27(2):145-51.

194. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nature reviews Genetics. 2014 Apr 1;15(4):234-46.

195. Ong ML, Holbrook JD. Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. Aging cell. 2014 Feb 1;13(1):142-55.

196. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 2007 Aug 9;448(7154):714-7.

197. Otter M, Schrander-Stumpel CT, Curfs LM. Triple X syndrome: a review of the literature. European Journal of Human Genetics. 2010 Mar 1;18(3):265-71.

198. Pan B, Li R, Chen Y, Tang Q, Wu W, Chen L, Lu C, Pan F, Ding H, Xia Y, Hu L. Genetic association between androgen receptor gene CAG repeat length polymorphism and male infertility: a meta-analysis. Medicine. 2016 Mar;95(10).

199. Papaioannou MD, Nef S. microRNAs in the testis: building up male fertility. Journal of andrology. 2010 Jan 2;31(1):26-33.

200. Peng Y. A novel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine. 2006 Jun 30;36(6):553-73.

201. Permutation Feature Importance - Azure Machine Learning Studio [Internet]. Docs.microsoft.com. 2018 [cited 1 February 2018]. Available from: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance

202. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome biology. 2016 Oct 7;17(1):208.

203. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC genomics. 2008 Mar 20;9(1):S13.

204. Poplinski A, Tüttelmann F, Kanber D, Horsthemke B, Gromoll J. Idiopathic male infertility is strongly associated with aberrant methylation of MEST and IGF2/H19 ICR1. International journal of andrology. 2010 Aug 1;33(4):642-9.

205. Price EM, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, Robinson WP, Kobor MS. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics & chromatin. 2013 Mar 3;6(1):4.

206. Punab M, Poolamets O, Paju P, Vihljajev V, Pomm K, Ladva R, Korrovits P, Laan M. Causes of male infertility: a 9-year prospective monocentre study on 1737 patients with reduced total sperm counts. Human reproduction. 2016 Dec 16;32(1):18-31.

207. Rajender S, Rajani V, Gupta NJ, Chakravarty B, Singh L, Thangaraj K. No association of androgen receptor GGN repeat length polymorphism with infertility in Indian men. Journal of andrology. 2006 Nov 12;27(6):785-9.

208. Ramasamy R, Ridgeway A, Lipshultz LI, Lamb DJ. Integrative DNA methylation and gene expression analysis identifies discoidin domain receptor 1 association with idiopathic nonobstructive azoospermia. Fertility and sterility. 2014 Oct 31;102(4):968-73.

209. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK. Global variation in copy number in the human genome. nature. 2006 Nov 23;444(7118):444-54.

210. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007 May 24;447(7143):425-32.

211. Ren H, Ferguson K, Kirkpatrick G, Vinning T, Chow V, Ma S. Altered crossover distribution and frequency in spermatocytes of infertile men with Azoospermia. PloS one. 2016 Jun 6;11(6):e0156817.

212. Ritchie ME, Carvalho BS, Hetrick KN, Tavar'e S. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. Bioinformatics. 2009 Oct 1;25(19):2621-3

213. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 43(7) e47.

214. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK. A comparison of background correction methods for two-colour microarrays. Bioinformatics. 2007 Aug 25;23(20):2700-7.

215. Robert B. Scharpf, Rafael A. Irizarry, Matthew E. Ritchie, Benilton Carvalho, Ingo Ruczinski (2011). Using the R Package crlmm for Genotyping and Copy Number Estimation. Journal of Statistical Software, 40(12), 1-32. URL http://www.jstatsoft.org/v40/i12/

216. Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE. Mining gene expression data by interpreting principal components. BMC bioinformatics. 2006 Apr 7;7(1):194.

217. Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE. Mining gene expression data by interpreting principal components. BMC bioinformatics. 2006 Apr 7;7(1):194.

218. Roseweir AK, Millar RP. The role of kisspeptin in the control of gonadotrophin secretion. Human reproduction update. 2008 Dec 24;15(2):203-12.

219. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, Frankish A. The DNA sequence of the human X chromosome. Nature. 2005 Mar 17;434(7031):325-37.

220. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007 Oct 1;23(19):2507-17.

221. Sakamoto H, Suzuki M, Abe T, Hosoyama T, Himeno E, Tanaka S, Greally JM, Hattori N, Yagi S, Shiota K. Cell type-specific methylation profiles occurring disproportionately in CpG-less regions that delineate developmental similarity. Genes to Cells. 2007 Oct 1;12(10):1123-32.

222. Sakian S, Louie K, Wong EC, Havelock J, Kashyap S, Rowe T, Taylor B, Ma S. Altered gene expression of H19 and IGF2 in placentas from ART pregnancies. Placenta. 2015 Oct 31;36(10):1100-5.

223. Sandovici I, Hoelle K, Angiolini E, Constância M. Placental adaptations to the maternal–fetal environment: implications for fetal growth and developmental programming. Reproductive biomedicine online. 2012 Jul 31;25(1):68-89.

224. Santi D, De Vincentis S, Magnani E, Spaggiari G. Impairment of sperm DNA methylation in male infertility: a meta-analytic study. Andrology. 2017 Jul 1;5(4):695-703.

225. Santi D, De Vincentis S, Magnani E, Spaggiari G. Impairment of sperm DNA methylation in male infertility: a meta-analytic study. Andrology. 2017 Jul 1;5(4):695-703.

226. Sathananthan AH. Paternal centrosomal dynamics in early human development and infertility. Journal of assisted reproduction and genetics. 1998 Mar 1;15(3):129-39.

227. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA A multilevel model to address batch effects in copy number estimation using SNP arrays Biostatistics. 2011.

228. Schultz N, Hamra FK, Garbers DL. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proceedings of the National Academy of Sciences. 2003 Oct 14;100(21):12201-6.

229. Shima Y, Morohashi KI. Leydig progenitor cells in fetal testis. Molecular and cellular endocrinology. 2017 Apr 15;445:55-64.

230. Simerly C, Wu GJ, Zoran S, Ord T, Rawlins R, Jones J, Navara C, Gerrity M, Rinehart J, Binor Z, Asch R. The paternal inheritance of the centrosome, the cell's microtubule-organizing center, in humans, and the implications for infertility. Nature medicine. 1995 Jan 1;1(1):47-52.

231. Soto-Ramírez N, Arshad SH, Holloway JW, Zhang H, Schauberger E, Ewart S, Patil V, Karmaus W. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. Clinical epigenetics. 2013 Dec;5(1):1.

232. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC bioinformatics. 2008 Jul 22;9(1):319.

233. Struhl K. Histone acetylation and transcriptional regulatory mechanisms. Genes & development. 1998 Mar 1;12(5):599-606.

234. Stuppia L, Franzago M, Ballerini P, Gatta V, Antonucci I. Epigenetics and male reproduction: the consequences of paternal lifestyle on fertility, embryo development, and children lifetime health. Clinical epigenetics. 2015 Nov 11;7(1):120.

235. Sutcliffe AG, Peters CJ, Bowdin S, Temple K, Reardon W, Wilson L, Clayton-Smith J, Brueton LA, Bannister W, Maher ER. Assisted reproductive therapies and imprinting disorders—a preliminary British survey. Human Reproduction. 2005 Dec 16;21(4):1009-11.

236. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS computational biology. 2010 Feb 5;6(2):e1000662.

237. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification.

238. Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, Hackett JA, Chinnery PF, Surani MA. A unique gene regulatory network resets the human germline epigenome for development. Cell. 2015 Jun 4;161(6):1453-67.

239. Tang WW, Kobayashi T, Irie N, Dietmann S, Surani MA. Specification and epigenetic programming of the human germ line. Nature Reviews Genetics. 2016 Aug 30.

240. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome research. 2003 Sep 1;13(9):2129-41.

241. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012 Jun;4(3):325-41.

242. Tremblay KD, Duran KL, Bartolomei MS. A 5'2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. Molecular and cellular biology. 1997 Aug 1;17(8):4322-9.

243. Triche Jr TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of illumina infinium dna methylation beadarrays. Nucleic acids research. 2013 Mar 9;41(7):e90-.

244. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, Viñuela A, Grundberg E, Nelson CP, Meduri E, Buil A. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014 Oct 3;9(10):1382-96.

245. Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers. Connection science. 1996 Dec 1;8(3-4):385-404.

246. Tüttelmann F, Rajpert-De Meyts E, Nieschlag E, Simoni M. Gene polymorphisms and male infertility–a meta-analysis and literature review. Reproductive biomedicine online. 2007 Jan 1;15(6):643-58.

247. Tüttelmann F, Rajpert-De Meyts E, Nieschlag E, Simoni M. Gene polymorphisms and male infertility–a meta-analysis and literature review. Reproductive biomedicine online. 2007 Jan 1;15(6):643-58.

248. Tüttelmann F, Röpke A. Genetics of Male Infertility. Endocrinology of the Testis and Male Reproduction. 2017:1-21.

249. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

250. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H. Towards a knowledge-based human protein atlas. Nature biotechnology. 2010 Dec 1;28(12):1248-50.

251. Urdinguio RG, Bayón GF, Dmitrijeva M, Toraño EG, Bravo C, Fraga MF, Bassas L, Larriba S, Fernández AF. Aberrant DNA methylation patterns of spermatozoa in men with unexplained infertility. Human reproduction. 2015 Mar 9;30(5):1014-28.

252. Van Assche E, Bonduelle M, Tournaye H, Joris H, Verheyen G, Devroey P, Van Steirteghem A, Liebaers I. Cytogenetics of infertile men. Human reproduction. 1996 Jan 1;11(suppl_4):1-26.

253. Vergnolle MA, Taylor SS. Cenp-F links kinetochores to Ndel1/Nde1/Lis1/dynein microtubule motor complexes. Current biology. 2007 Jul 3;17(13):1173-9.

254. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Court DS. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Science International: Genetics. 2017 May 31;28:225-36.

255. Vincent RN, Gooding LD, Louie K, Wong EC, Ma S. Altered DNA methylation and expression of PLAGL1 in cord blood from assisted reproductive technology pregnancies compared with natural conceptions. Fertility and sterility. 2016 Sep 1;106(3):739-48.

256. Vincent SD, Dunn NR, Sciammas R, Shapiro-Shalef M, Davis MM, Calame K, Bikoff EK, Robertson EJ. The zinc finger transcriptional repressor Blimp1/Prdm1 is dispensable for early axis formation but is required for specification of primordial germ cells in the mouse. Development. 2005 Mar 15;132(6):1315-25.

257. Vogt PH. Molecular genetic of human male infertility: from genes to new therapeutic perspectives. Current pharmaceutical design. 2004 Feb 1;10(5):471-500.

258. Walker DL, Bhagwate AV, Baheti S, Smalley RL, Hilker CA, Sun Z, Cunningham JM. DNA methylation profiling: comparison of genome-wide sequencing methods and the Infinium Human Methylation 450 Bead Chip. Epigenomics. 2015 Dec;7(8):1287-302.

259. Wallach EE, Moghissi KS. Unexplained infertility. Fertility and sterility. 1983 Jan 31;39(1):5-21.

260. Wang C, Yang C, Chen X, Yao B, Yang C, Zhu C, Li L, Wang J, Li X, Shao Y, Liu Y. Altered profile of seminal plasma microRNAs in the molecular diagnosis of male infertility. Clinical chemistry. 2011 Dec 1;57(12):1722-31.

261. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW. Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry. 2005 Feb 28;29(1):37-46.

262. Watanabe Y, Nurse P. Cohesin Rec8 is required for reductional chromosome segregation at meiosis. Nature. 1999 Jul 29;400(6743):461-4.

263. Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, Meachem SJ. Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. Proceedings of the National Academy of Sciences of the United States of America. 2005 Mar 15;102(11):4068-73.

264. Wei B, Xu Z, Ruan J, Zhu M, Jin K, Zhou D, Xu Z, Hu Q, Wang Q, Wang Z. MTHFR 677C> T and 1298A> C polymorphisms and male infertility risk: a meta-analysis. Molecular biology reports. 2012 Feb 1;39(2):1997-2002.

265. Whorton D, Krauss R, Marshall S, Milby T. Infertility in male pesticide workers. The Lancet. 1977 Dec 17;310(8051):1259-61.

266. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, Marsit CJ, Houseman EA, Brown R. Review of processing and analysis methods for DNA methylation array data. British journal of cancer. 2013 Sep 17;109(6):1394-402.

267. Wright ML, Dozmorov MG, Wolen AR, Jackson-Cook C, Starkweather AR, Lyon DE, York TP. Establishing an analytic pipeline for genome-wide DNA methylation. Clinical epigenetics. 2016 Apr 27;8(1):45.

268. Yamaguchi S, Hong K, Liu R, Inoue A, Shen L, Zhang K, Zhang Y. Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. Cell research. 2013 Mar 1;23(3):329-39.

269. Yaman R, Grandjean V. Timing of entry of meiosis depends on a mark generated by DNA methyltransferase 3a in testis. Molecular reproduction and development. 2006 Mar 1;73(3):390-7.

270. Yan W. Male infertility caused by spermiogenic defects: lessons from gene knockouts. Molecular and cellular endocrinology. 2009 Jul 10;306(1):24-32.

271. Yan W. Potential roles of noncoding RNAs in environmental epigenetic transgenerational inheritance. Molecular and cellular endocrinology. 2014 Dec 31;398(1):24-30.

272. Yao D, Yang J, Zhan X, Zhan X, Xie Z. A novel random forests-based feature selection method for microarray expression data analysis. International journal of data mining and bioinformatics. 2015;13(1):84-101.

273. Yatsenko AN, Georgiadis AP, Röpke A, Berman AJ, Jaffe T, Olszewska M, Westernströer B, Sanfilippo J, Kurpisz M, Rajkovic A, Yatsenko SA. X-linked TEX11 mutations, meiotic arrest, and azoospermia in infertile men. New England Journal of Medicine. 2015 May 28;372(22):2097-107.

274. Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. International neurourology journal. 2014 Jun;18(2):50.

275. Zeng Y, Yi R, Cullen BR. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. The EMBO journal. 2005 Jan 12;24(1):138-48.

276. Zhao H, Xu J, Zhang H, Sun J, Sun Y, Wang Z, Liu J, Ding Q, Lu S, Shi R, You L. A genome-wide association study reveals that variants within the HLA region are associated with risk for nonobstructive azoospermia. The American Journal of Human Genetics. 2012 May 4;90(5):900-6.

277. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nature methods. 2014 Mar 1;11(3):309-11.

Appendices

Appendix A

A.1 Genes associated with localization GO term

RefGene Name	Description
APOH	Beta-2-glycoprotein 1;APOH;ortholog
ARFGAP1	ADP-ribosylation factor GTPase-activating protein 1;ARFGAP1;ortholog
ARFGAP1	ADP-ribosylation factor GTPase-activating protein 3;ARFGAP3;ortholog
ARL13B	ADP-ribosylation factor-like protein 13B;ARL13B;ortholog
ATP6V0D2	V-type proton ATPase subunit d 2;ATP6V0D2;ortholog
BAG4	BAG family molecular chaperone regulator 4;BAG4;ortholog
BAIAP2	Brain-specific angiogenesis inhibitor 1-associated protein 2;BAIAP2;ortholog
BNIP3	BCL2/adenovirus E1B 19 kDa protein-interacting protein 3;BNIP3;ortholog
BPI	Bactericidal permeability-increasing protein;BPI;ortholog
CACNA1H	Voltage-dependent T-type calcium channel subunit alpha-1H;CACNA1H;ortholog
CACNA2D3	Voltage-dependent calcium channel subunit alpha-2/delta-3;CACNA2D3;ortholog
CD302	CD302 antigen;CD302;ortholog
CDC7	Cell division cycle 7-related protein kinase;CDC7;ortholog
CDK5RAP2	CDK5 regulatory subunit-associated protein 2;CDK5RAP2;ortholog
CEBPE	CCAAT/enhancer-binding protein epsilon;CEBPE;ortholog
CEP110	Centriolar coiled-coil protein of 110 kDa;CCP110;ortholog
CEP97	Centrosomal protein of 97 kDa;CEP97;ortholog
CHST11	Carbohydrate sulfotransferase 11;CHST11;ortholog
CKAP5	Cytoskeleton-associated protein 5;CKAP5;ortholog
COX11	Cytochrome c oxidase assembly protein COX11, mitochondrial;COX11;ortholog
CYB5R2	NADH-cytochrome b5 reductase 2;CYB5R2;ortholog
DAB1	Disabled homolog 1;DAB1;ortholog
EBP	3-beta-hydroxysteroid-Delta(8),Delta(7)-isomerase;EBP;ortholog
EFR3A	Protein EFR3 homolog A;EFR3A;ortholog
FBL	rRNA 2'-O-methyltransferase fibrillarin;FBL;ortholog
FLNA	Filamin-A;FLNA;ortholog
FXC1	Mitochondrial import inner membrane translocase subunit Tim10 B;TIMM10B;ortholog
FYTTD1	UAP56-interacting factor;FYTTD1;ortholog
GABRA6	Gamma-aminobutyric acid receptor subunit alpha-6;GABRA6;ortholog
GABRG3	Gamma-aminobutyric acid receptor subunit gamma-3;GABRG3;ortholog
GBF1	Golgi-specific brefeldin A-resistance guanine nucleotide exchange factor 1;GBF1;ortholog
GLI3	Transcriptional activator GLI3;GLI3;ortholog
GOLGA3	Golgin subfamily A member 3;GOLGA3;ortholog
GPLD1	Phosphatidylinositol-glycan-specific phospholipase D;GPLD1;ortholog
GPR177	Protein wntless homolog;WLS;ortholog
GRIN2C	Glutamate receptor ionotropic, NMDA 2C;GRIN2C;ortholog
GSDMD	Gasdermin-D;GSDMD;ortholog
HDLBP	Vigilin;HDLBP;ortholog
HGS	Hepatocyte growth factor-regulated tyrosine kinase substrate;HGS;ortholog
HHIPL1	HHIP-like protein 1;HHIPL1;ortholog
HLA-A	HLA class I histocompatibility antigen, A-68 alpha chain;HLA-A;ortholog
HLA-C	HLA class I histocompatibility antigen, Cw-17 alpha chain;HLA-C;ortholog
KCNJ5	G protein-activated inward rectifier potassium channel 4;KCNJ5;ortholog
KIF2B	Kinesin-like protein KIF2B;KIF2B;ortholog
KIF3B	Kinesin-like protein KIF3B;KIF3B;ortholog
MAGI2	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein

	2;MAGI2;ortholog
MAP6D1	MAP6 domain-containing protein 1;MAP6D1;ortholog
MBP	Bone marrow proteoglycan;PRG2;ortholog
MCPH1	Microcephalin;MCPH1;ortholog
METTL7A	Methyltransferase-like protein 7A;METTL7A;ortholog
MMP14	Matrix metalloproteinase-14;MMP14;ortholog
MYO1C	Unconventional myosin-Ie;MYO1E;ortholog
MYO1C	Unconventional myosin-Ic;MYO1C;ortholog
NEDD1	Protein NEDD1;NEDD1;ortholog
NFKBIA	NF-kappa-B inhibitor alpha;NFKBIA;ortholog
NODAL	Nodal homolog;NODAL;ortholog
NPTX1	Neuronal pentraxin-1;NPTX1;ortholog
NRBP2	Nuclear receptor-binding protein 2;NRBP2;ortholog
NRXN3	Neurexin-3;NRXN3;ortholog
NTRK3	NT-3 growth factor receptor;NTRK3;ortholog
OSBPL5	Oxysterol-binding protein-related protein 5;OSBPL5;ortholog
PHLDB2	Pleckstrin homology-like domain family B member 2;PHLDB2;ortholog
PITPNM3	Membrane-associated phosphatidylinositol transfer protein 3;PITPNM3;ortholog
PLAC8	Placenta-specific gene 8 protein;PLAC8;ortholog
PSG2	Pregnancy-specific beta-1-glycoprotein 2;PSG2;ortholog
PSMA6	Proteasome subunit alpha type-6;PSMA6;ortholog
PSMD1	26S proteasome non-ATPase regulatory subunit 1;PSMD1;ortholog
PSMD14	26S proteasome non-ATPase regulatory subunit 14;PSMD14;ortholog
PTPRN2	Receptor-type tyrosine-protein phosphatase N2;PTPRN2;ortholog
RAB11B	Ras-related protein Rab-11B;RAB11B;ortholog
SCAMP1	Secretory carrier-associated membrane protein 1;SCAMP1;ortholog
SFRS7	Serine/arginine-rich splicing factor 7;SRSF7;ortholog
SH3PXD2B	SH3 and PX domain-containing protein 2B;SH3PXD2B;ortholog
SLC12A6	Solute carrier family 12 member 6;SLC12A6;ortholog
SLC6A15	Sodium-dependent neutral amino acid transporter B(0)AT2;SLC6A15;ortholog
SLC7A10	Asc-type amino acid transporter 1;SLC7A10;ortholog
SLCO5A1	Solute carrier organic anion transporter family member 5A1;SLCO5A1;ortholog
STXBP3	Syntaxin-binding protein 3;STXBP3;ortholog
SUPT6H	Transcription elongation factor SPT6;SUPT6H;ortholog
TBC1D25	TBC1 domain family member 25;TBC1D25;ortholog
TBX5	T-box transcription factor TBX5;TBX5;ortholog
TEKT2	Tektin-2;TEKT2;ortholog
TMCO1	Calcium load-activated calcium channel;TMCO1;ortholog
TMPRSS13	Transmembrane protease serine 13;TMPRSS13;ortholog
TRPC3	Short transient receptor potential channel 3;TRPC3;ortholog
TRPM8	Transient receptor potential cation channel subfamily M member 8;TRPM8;ortholog
TSG101	Tumor susceptibility gene 101 protein;TSG101;ortholog
TSPAN5	Tetraspanin-5;TSPAN5;ortholog
VPS11	Vacuolar protein sorting-associated protein 11 homolog;VPS11;ortholog
WNT5A	Protein Wnt-5a;WNT5A;ortholog
ZNF593	Zinc finger protein 593;ZNF593;ortholog

A.2 GO term enrichment analysis of testis specific genes linked to localization

GO biological process complete	Matching genes	Fold enrichment	FDR adjusted *P*-value

formation of radial glial scaffolds (GO:0021943)	4	> 100	1.98E-02
cell differentiation involved in kidney development (GO:0061005)	43	27.19	3.42E-02
retrograde vesicle-mediated transport, Golgi to ER (GO:0006890)	81	24.05	1.45E-03
cranial skeletal system development (GO:1904888)	66	23.62	9.26E-03
microtubule cytoskeleton organization involved in mitosis (GO:1902850)	95	16.41	2.20E-02
ciliary basal body-plasma membrane docking (GO:0097711)	96	16.24	2.24E-02
Wnt signaling pathway, planar cell polarity pathway (GO:0060071)	101	15.43	2.53E-02
tumor necrosis factor-mediated signaling pathway (GO:0033209)	129	15.1	7.49E-03
organelle localization by membrane tethering (GO:0140056)	158	14.8	1.80E-03
membrane docking (GO:0022406)	167	14	2.31E-03
regulation of G2/M transition of mitotic cell cycle (GO:0010389)	197	13.85	7.85E-04
regulation of establishment of planar polarity (GO:0090175)	114	13.67	3.54E-02
cerebral cortex development (GO:0021987)	114	13.67	3.51E-02
regulation of stem cell differentiation (GO:2000736)	117	13.32	3.75E-02
regulation of cell cycle G2/M phase transition (GO:1902749)	215	12.69	1.12E-03
stimulatory C-type lectin receptor signaling pathway (GO:0002223)	123	12.67	4.31E-02
innate immune response activating cell surface receptor signaling pathway (GO:0002220)	126	12.37	4.63E-02
antigen processing and presentation of exogenous peptide antigen (GO:0002478)	175	11.13	1.97E-02
regulation of morphogenesis of an epithelium (GO:1905330)	182	10.71	2.18E-02
antigen processing and presentation of exogenous antigen (GO:0019884)	182	10.71	2.16E-02
antigen processing and presentation of peptide antigen (GO:0048002)	185	10.53	2.24E-02
positive regulation of cytoskeleton organization (GO:0051495)	200	9.74	2.84E-02
cilium assembly (GO:0060271)	328	9.5	1.25E-03
Fc receptor signaling pathway (GO:0038093)	249	9.39	1.19E-02
Golgi vesicle transport (GO:0048193)	334	9.33	1.32E-03
antigen processing and presentation (GO:0019882)	213	9.15	3.55E-02
cilium organization (GO:0044782)	341	9.14	1.44E-03
RNA localization (GO:0006403)	219	8.9	3.81E-02
plasma membrane bounded cell projection assembly (GO:0120031)	415	8.45	9.60E-04
cell projection assembly (GO:0030031)	420	8.35	1.00E-03
organelle localization (GO:0051640)	577	8.1	4.58E-05
regulation of actin filament-based process (GO:0032970)	347	7.86	9.30E-03
microtubule cytoskeleton organization (GO:0000226)	399	7.81	3.42E-03
regulation of actin cytoskeleton organization (GO:0032956)	305	7.67	2.40E-02
regulation of cytoskeleton organization (GO:0051493)	474	7.4	1.66E-03
establishment of organelle localization (GO:0051656)	385	7.08	1.47E-02
negative regulation of organelle organization (GO:0010639)	330	7.08	3.29E-02
regulation of protein catabolic process (GO:0042176)	406	6.72	1.81E-02
regulation of mitotic cell cycle phase transition (GO:1901990)	407	6.7	1.81E-02
protein localization to membrane (GO:0072657)	407	6.7	1.79E-02
regulation of vesicle-mediated transport (GO:0060627)	468	6.66	8.23E-03
regulation of mitotic cell cycle (GO:0007346)	653	6.56	7.50E-04
cellular component disassembly (GO:0022411)	419	6.51	2.00E-02
positive regulation of cellular component biogenesis (GO:0044089)	488	6.39	9.67E-03
organelle assembly (GO:0070925)	679	6.31	9.70E-04
regulation of cell cycle phase transition (GO:1901987)	439	6.21	2.36E-02
mitotic cell cycle process (GO:1903047)	653	5.97	2.53E-03
regulation of cellular protein localization (GO:1903827)	603	5.82	7.36E-03
mitotic cell cycle (GO:0000278)	711	5.48	4.75E-03
regulation of cell cycle process (GO:0010564)	714	5.46	4.80E-03
negative regulation of cell cycle (GO:0045786)	572	5.45	2.06E-02
regulation of cellular component biogenesis (GO:0044087)	858	5.45	1.10E-03

microtubule-based process (GO:0007017)	590	5.28	2.31E-02
cell cycle process (GO:0022402)	1027	4.93	1.07E-03
macromolecule localization (GO:0033036)	2258	4.83	4.66E-10
regulation of anatomical structure morphogenesis (GO:0022603)	980	4.77	2.60E-03
cellular protein localization (GO:0034613)	1332	4.68	2.54E-04
cellular macromolecule localization (GO:0070727)	1342	4.65	2.53E-04
cellular localization (GO:0051641)	2122	4.59	3.75E-08
regulation of protein transport (GO:0051223)	787	4.46	2.80E-02
regulation of organelle organization (GO:0033043)	1227	4.45	1.21E-03
regulation of cellular localization (GO:0060341)	877	4.44	1.69E-02
protein localization (GO:0008104)	1945	4.41	2.04E-06
vesicle-mediated transport (GO:0016192)	1787	4.36	1.69E-05
regulation of cell cycle (GO:0051726)	1164	4.35	2.52E-03
regulation of peptide transport (GO:0090087)	817	4.29	3.44E-02
regulation of establishment of protein localization (GO:0070201)	830	4.23	3.68E-02
regulation of catabolic process (GO:0009894)	924	4.22	2.15E-02
establishment of localization in cell (GO:0051649)	1494	4.17	7.57E-04
plasma membrane bounded cell projection organization (GO:0120036)	1043	4.11	1.45E-02
regulation of locomotion (GO:0040012)	855	4.1	4.29E-02
establishment of protein localization (GO:0045184)	1439	4.06	1.34E-03
cytoskeleton organization (GO:0007010)	974	4	2.78E-02
cell projection organization (GO:0030030)	1072	4	1.70E-02
positive regulation of transport (GO:0051050)	978	3.98	2.81E-02
localization (GO:0051179)	5433	3.8	5.17E-26
regulation of protein localization (GO:0032880)	1043	3.74	4.11E-02
cell cycle (GO:0007049)	1361	3.72	9.25E-03
protein transport (GO:0015031)	1362	3.72	9.15E-03
intracellular transport (GO:0046907)	1265	3.7	1.68E-02
peptide transport (GO:0015833)	1386	3.65	1.04E-02
positive regulation of cellular component organization (GO:0051130)	1182	3.63	2.87E-02
amide transport (GO:0042886)	1408	3.6	1.18E-02
regulation of cell differentiation (GO:0045595)	1672	3.5	5.16E-03
nitrogen compound transport (GO:0071705)	1672	3.5	5.05E-03
movement of cell or subcellular component (GO:0006928)	1462	3.46	1.58E-02
organic substance transport (GO:0071702)	2043	3.43	1.22E-03
establishment of localization (GO:0051234)	4442	3.42	1.08E-11
regulation of transport (GO:0051049)	1849	3.37	4.15E-03
regulation of cellular component organization (GO:0051128)	2343	3.33	7.82E-04
transport (GO:0006810)	4326	3.24	1.39E-09
regulation of multicellular organismal development (GO:2000026)	1831	3.19	1.09E-02
regulation of localization (GO:0032879)	2595	3.15	7.57E-04
regulation of developmental process (GO:0050793)	2379	2.95	5.39E-03
cellular component biogenesis (GO:0044085)	2610	2.84	5.04E-03
cellular component assembly (GO:0022607)	2346	2.82	1.28E-02
regulation of catalytic activity (GO:0050790)	2317	2.69	2.72E-02
organelle organization (GO:0006996)	3182	2.57	6.35E-03
regulation of multicellular organismal process (GO:0051239)	2801	2.5	2.38E-02
animal organ development (GO:0048513)	2988	2.48	1.95E-02
positive regulation of macromolecule metabolic process (GO:0010604)	3034	2.44	2.20E-02
positive regulation of metabolic process (GO:0009893)	3268	2.38	1.98E-02
regulation of molecular function (GO:0065009)	3391	2.3	3.51E-02
cellular component organization (GO:0016043)	5280	2.29	8.05E-04
cellular component organization or biogenesis (GO:0071840)	5505	2.27	5.38E-04
positive regulation of cellular process (GO:0048522)	4985	2.11	8.99E-03

negative regulation of cellular process (GO:0048523)	4457	2.1	2.51E-02
negative regulation of biological process (GO:0048519)	4952	1.97	4.02E-02
positive regulation of biological process (GO:0048518)	5607	1.88	4.04E-02
regulation of biological process (GO:0050789)	11425	1.57	1.50E-03
regulation of cellular process (GO:0050794)	10771	1.56	8.18E-03
biological regulation (GO:0065007)	12072	1.52	2.36E-03
cellular process (GO:0009987)	15086	1.32	1.28E-02
Biological process (GO:0008150)	17500	1.2	1.65E-02