Investigating *de novo* mutations in severe idiopathic male infertility

Hannah Elizabeth Smith

BSc, MRes.

A thesis submitted to Newcastle University for the degree of Doctor of Philosophy

Biosciences Institute

Faculty of Medical Sciences

Newcastle University

March 2024

Abstract

Infertility affects 1 in 6 couples worldwide and in approximately 50 % of these cases are attributed to a male factor. Whilst causes such as Klinefelter's and Y chromosome microdeletions have been well established, the genetic causes behind severe spermatogenic failure are largely unknown, with around 40 % of all male infertility cases remaining idiopathic. *De novo* mutations (DNMs) are known to play a prominent role in many sporadic disorders with reduced fitness. These mutations, however, are vastly understudied in the field of male infertility due to the difficulty of obtaining parental samples. The foundation of this thesis is formed around the hypothesis that DNMs play a vital role in male infertility and explain a significant fraction of the genetic causes of this understudied disorder. To test this hypothesis, next generation sequencing (NGS) techniques were utilised to sequence the DNA of males with idiopathic azoospermia or oligozoospermia and their unaffected parents. An initial study was conducted, utilising a targeted 54-gene panel NGS assay to identify DNMs in 75 patient parent trios. However, this proved to be too narrow a gene pool with no possibly causative DNMs identified across the cohort. A further whole exome sequencing (WES) study was then performed in a cohort of 185 males suffering from idiopathic male infertility and their parents. In total, 145 rare protein-altering *de novo* SNVs were identified. Following a systematic analysis assessing mutational impact and protein function, 29 DNMs were classified as possibly causative. A significant enrichment of loss-of-function DNMs in loss-of-function-intolerant genes (p-value = 1.00×10^{-5}) was seen in infertile men compared to controls. A significant increase was also identified in predicted pathogenic *de novo* missense mutations affecting missense-intolerant genes (p-value = 5.01×10^{-4}). One such gene containing a pathogenic missense DNM, *RBM5*, is an essential regulator of male germ cell pre-mRNA splicing and has been previously implicated in male infertility in mice. A follow-up study in a cohort of 2,506 infertile males identified six rare heterozygous pathogenic missense mutations affecting *RBM5,* whilst no such mutations were present in a cohort of 5,784 fertile men (p-value = 0.03). Results from a minigene splicing assay suggested that mutations in *RBM5* may potentially disrupt its function as a pre-MRNA splicing factor. The work presented in this thesis provides evidence for the role of *de novo* mutations in severe idiopathic male infertility, highlights several potential candidate genes causing infertility and suggests potential biological processes disturbed by these genes.

Acknowledgements

I would like to firstly thank my supervisor Professor Joris Veltman for giving me this incredible opportunity and for providing encouragement, guidance, and support throughout the entirety of my PhD. I would also like to thank my supervisor Professor David Elliott for all your help with the functional aspect of my studies when my initial knowledge on the subject was limited.

To Dr Miguel Xavier, I would like to say thank you for being there consistently throughout my PhD, for always being willing to help me no matter how big or small the issue and for all your assistance with the bioinformatic side of my work.

I want to thank all of the members of the Newcastle male infertility research group, both past and present for their friendship and support. In particular, I want to extend my thanks to Dr Bilal Alobadi who has taught me so much. Your support over the last 5 years has been invaluable and your kindness and friendship is something I am incredibly grateful for. To Dr Harsh Sheth, I would like to say thank you for your encouragement in applying for this PhD and your guidance during the first year of my PhD. To the members of my progress panel, Professor Mary Herbert, and Professor Heather Cordell, thank you for your feedback and guidance during my progress meetings. I would like to thank the Genomics Core Facility in Newcastle University for sequencing of all WES samples, and in particular, Rafiqul Hussain for helping me when it came to performing my own sequencing run. I would also like to thank the Bioinformatic Support Unit for processing the initial sequencing data.

I would like to thank the team at Radboud University including Dr Manon Oud who performed the initial studies which founded the basis for my PhD and Dr Godfried van der Heijden for all his support and interesting discussions on patient samples and histology. It has been a pleasure working with you both. I would also to extend my thanks to all our collaborators across the globe in particular, the members of the IMIGC. I would like to thank my funding body 'The Barbour Foundation' as well as all the patients and families who have participated in our studies, without whom my PhD would not have been possible.

I would like to thank my friends and family for all their love and patience over the course of my PhD, and for always believing in me.

Finally, I want to express my unending gratitude to my partner, Ethan, who has been by my side every step of the way throughout the many ups and downs that this PhD has had to offer. I will forever be grateful for the love and support you have shown.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABP	Androgen Binding Protein
ACMG	American College of Medical Genetics
AD	Autosomal Dominant
AF	Allele Frequency
AIS	Androgen Insensitivity Syndrome
ALS	Amyotrophic Lateral Sclerosis
AR	Autosomal Recessive
ART	Assisted Reproductive Techniques
AS	Alternative Splicing
ASD	Autism Spectrum Disorder
AZF	Azoospermia Factor
BAM	Binary Alignment Map
BED	Browser Extensible File
BPS	Branch Point Sequence
BSU	Bioinformatic Support Unit
CAIS	Complete Androgen Insensitivity Syndrome
CBAVD	Congenital Bilateral Absence of the *Vas Deferens*
CHD	Congenital Heart Disease
CHH	Congenital Hypogonadotropic Hypogonadism
CNV	Copy Number Variant
CS	Conserved Spermatogenesis
DD	Developmental Delay
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic Acid
DNM	De Novo Mutation
DSD	Disorder of Sex Development
EJC	Exon Junction Complex
FBS	Fetal Bovine Serum
FSH	Follicle Stimulating Hormone
GCA	Germ Cell Arrest
GCF	Genomics Core Facility

GEMINI	Genetics of Male Infertility Initiative
gnomAD	Genome Aggregation Database
GnRH	Gonadotropin-Releasing Hormone
GQ	Genotype Probabilities
HGMD	Human Gene Mutation Database
HH	Hypogonadotropic Hypogonadism
HPA	Hypothalamic Pituitary Axis
HPG	Hypothalamic-Pituitary-Gonadal
HS	Hypospermatogenesis
ICSI	Intracytoplasmic Sperm Injection
ID	Intellectual Disability
IDT	Integrated DNA Technologies
IGV	Integrated Genome Viewer
ILS	Intron Lariat Spliceosome
IMIGC	International Male Infertility Genomics Consortium
IVF	In Vitro Fertilisation
KS	Klinefelter Syndrome
LB	Luria-Bertani
LH	Luteinising Hormone
LHR	Luteinizing Hormone Receptors
LOEUF	Loss-of-function Observed over Expected upper bound Fraction
LoF	Loss of Function
MAF	Minor Allele Frequency
MAIS	Mild Androgen Insensitivity Syndrome
MeiA	Meiotic Arrest
MERGE	German Male Reproductive Genomics
MI	Male Infertility
MIP	Molecular Inversion Probe
MMAF	Multiple Morphological Abnormalities of the Flagella
NCL	Newcastle
NGS	Next Generation Sequencing
NIJ	Nijmegen
NOA	Non-Obstructive Azoospermia

NTC	NineTeen Complex
OA	Obstructive Azoospermia
PAIS	Partial Androgen Insensitivity Syndrome
PBS	Phosphate Buffered Saline
PCD	Primary Ciliary Dyskinesia
PCR	Polymerase Chain Reaction
pLI	Loss Intolerance probability
PNK	Polynucleotide Kinase
PSI	Percentage Splice Inclusion
RF	Read Fraction
RGC	Regeneron Genetics Center
RRM	RNA Recognition Motif
rsMF	Maturation Failure at the round spermatid stage
RT	Robertsonian Translocation
SAP	Shrimp Alkaline Phosphatase
scMF	Maturation Failure at the spermatocyte stage
SCO	Sertoli Cell Only (Syndrome)
SDS-PAGE	SDS Polyacrylamide Gel Electrophoresis
SIFT	Sorting Intolerant From Tolerant
smMIP	single-molecule Molecular Inversion Probe
SNP	Single Nucleotide Polymorphism
snRNA	small nuclear Ribonucleic Acid
snRNP	small nuclear Ribonucleoproteins
STRING	Search Tool for the Retrieval of Interacting Genes
SV	Structural Variants
TESE	Testicular Sperm Extraction
UMB	Unique Molecular Barcode
UCSC	University of California, Santa Cruz
VCF	Variant Call File
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organisation
WT	Wild Type
YCMD	Y Chromosome Microdeletions

Chapter 1: Introduction

1.1 Introduction to Male Infertility

Infertility is defined by the World Health Organisation (WHO) as the inability for a couple to achieve pregnancy after one year or more of unprotected sexual intercourse (World Health Organization, 2019). Approximately 10 % – 15 % of all couples worldwide are affected by some form of infertility and in around 50 % of these cases, a male factor is the cause (Sharma et al., 2021). According to WHO, normal sperm production (Normozoospermia) occurs when a male produces greater than 16 million sperm per ml of ejaculate with 42 % motility, and at least 4 % with a normal sperm morphology (World Health Organization, 2021).

Male infertility (MI) is a complex multifactorial disorder, presenting with highly heterogeneous phenotypes (Oud et al., 2019; Tüttelmann et al., 2018; Vockel et al., 2021; Krausz et al., 2020). Within this condition are four main aetiologies to which the majority of all cases can be attributed (Krausz & A Riera-Escamilla, 2018; Laan, 2019; Tournaye et al., 2017). Quantitative sperm defects, caused by primary testicular failure, are the most common of the aetiologies representing around 75 % of all MI cases (Krausz, 2011). Quantitative issues for MI begin when a males sperm count is reduced below the normal levels; with mild, moderate, severe and extreme oligozoospermia being defined as a sperm count of 10 - 15 million sperm per ml of ejaculate (mild), 5 - 10 million sperm per ml (moderate), 0.1 - 5 million sperm per ml (severe), 0 - 0.1 million sperm per ml (extreme) (Figure 1.1) (Choy & Amory, 2020; Koscinski et al., 2007). Azoospermia is the most severe form of MI with the absence of any sperm in the ejaculate. Azoospermia can then be further broken down into non-obstructive azoospermia (NOA), caused by issues during spermatogenesis itself as a result of primary or secondary testicular failure, and obstructive azoospermia (OA) (Cioppi et al., 2021). OA is caused by bilateral distal or proximal obstruction of the ejaculatory ducts which represents in part the next most common cause of infertility- ductal obstructions and dysfunction. This is followed by disruptions of the hypothalamic-pituitary-gonadal (HPG) axis, also known as secondary testicular failure. The final, and least common, aetiology is qualitative sperm defects. Qualitative sperm defects involve a wide range of conditions ranging from reduced sperm motility (asthenozoospermia), morphological issues referred to as teratozoospermia or a combination of the two (asthenoteratozoospermia) (Coutton et al., 2015). Other issues with sperm quality can be seen as globozoospermia, a form of teratozoospermia in which large numbers of sperm are round-headed, present with cytoskeleton defects and typically lack an acrosome (Fesahat et al., 2020). Along with globozoospermia, the remaining rarest types of MI are Primary Ciliary Dyskinesia (PCD), Multiple Morphological Abnormalities of the sperm Flagella (MMAF) and sperm vitality issues such as necrozoospermia. PCD is a disease caused

by the malfunction of motile cilia, typically resulting in chronic respiratory infections however in males it is often associated with infertility due to disruption to the motile cilia within the sperm axoneme (Jayasena & Sironen, 2021; Newman et al., 2023). Unlike PCD, MMAF is localised to sperm and is characterized by sperm cells with a variety of irregularities of the sperm tail including absent, shortened, or coiled flagella (Ben Khelifa et al., 2014). Necrozoospermia is found in around only 0.2 % – 0.4 % of all MI cases and whilst the exact threshold is still undetermined, it is represented by a majority of all sperm in the ejaculate being necrotic (Boursier et al., 2022; Brahem et al., 2012).

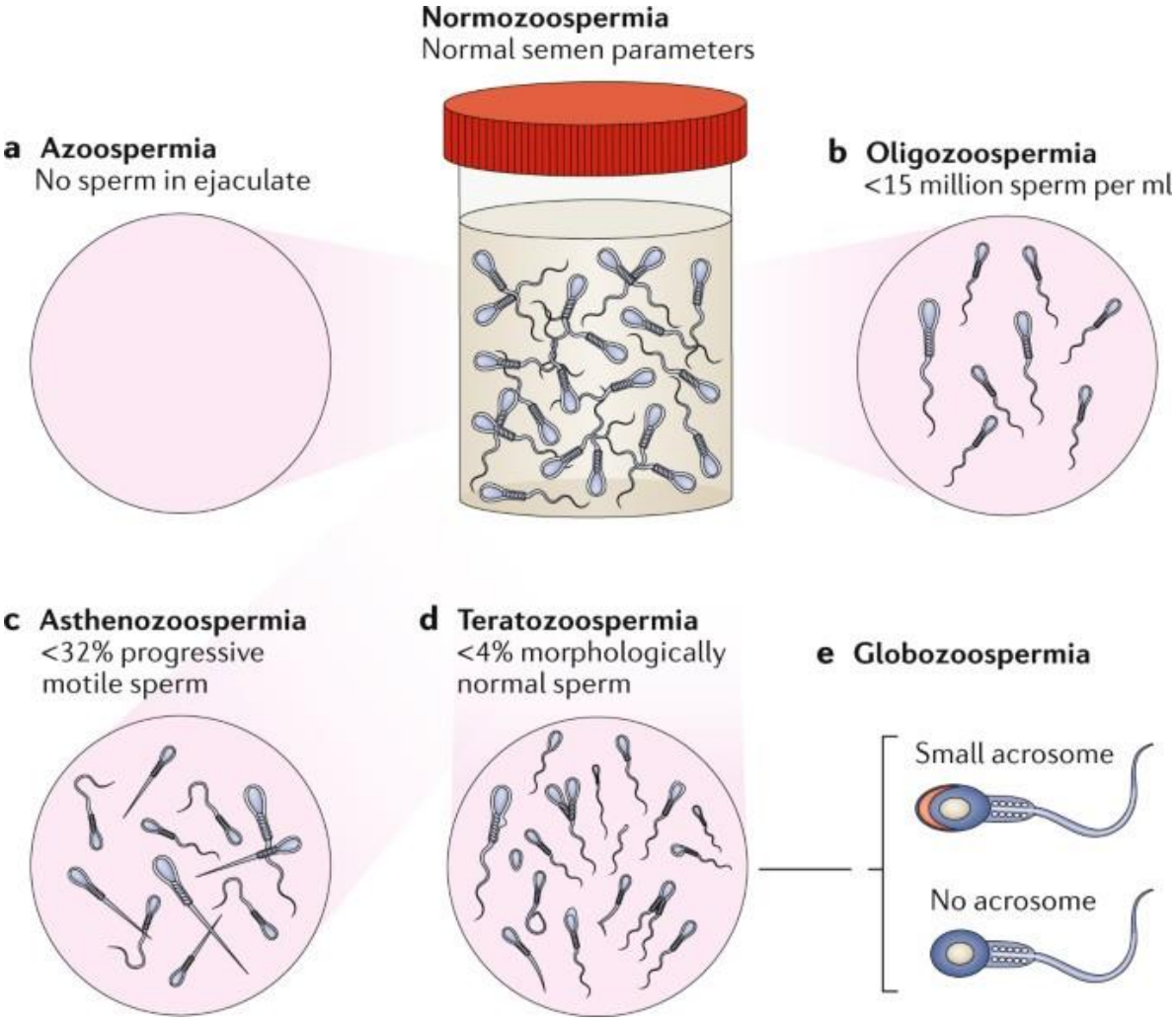


Figure 1.1: Seminal alterations associated with male infertility. Reproduced with permission from Springer Nature. (Esteves et al., 2018)

1.2 Epidemiology of Male Infertility

Whilst it is well reported that around 7 % of all males suffer from some form of infertility worldwide (Krausz, 2011), the epidemiology of the disease across different regions of the globe is difficult to report. Most often, the burden of infertility is placed upon the woman and in societies where male infertility is stigmatised or culturally not recognised, collecting accurate data is incredibly difficult. A study published in 2015 attempted to calculate the distribution of male fertility around the world (Agarwal et al., 2015). For many regions such as the Middle East, Asia, Latin America, and Africa however, data had to be extrapolated from couple infertility data. The study estimated that infertility rates in males were highest in Eastern Europe at around 8 % - 12 % of all males suffering from some form of infertility. A relatively substantial number of non-genetic factors are known to be risk factors for an increased likelihood of MI, with the most confidently linked factor known to be increased age. One study found that the time to pregnancy for men aged over 45 years was five times higher than that of men aged 25 years and under (Hassan & Killick, 2003). It has been well documented for a number of years that, as a male ages, his blood testosterone production eventually begins to decrease, beginning approximately at the age of 40 (Feldman et al., 1994). This, alongside a decreased libido, lower semen volume and increased abnormal sperm morphology, are all thought to be potential reasons for decreased fertility in older men (Kumar Mahat & Arora, 2016). Smoking has also been shown to have a detrimental effect on fertility in males with one study of 2542 health males showing a 19 % reduction in sperm concentration in those who smoked >20 cigarettes per day (Ramlau-Hansen et al., 2007). A further study of 1,786 males identified a 15.3 % reduction in sperm density, 17.5 % decrease in total sperm count and a 16.6 % decrease in total motile sperm in males who smoked (Künzle et al., 2003). Obesity is another lifestyle factor which has been shown to increase the risk of infertility in males. Decreased levels of serum testosterone and Luteinising Hormone (LH) can be associated with obesity as well as a higher chance of suffering from qualitative sperm defects (Sermondade et al., 2013; M. Al-Ali et al., 2014). It is thought that the essential hormones needed for the regulation of spermatogenesis are highly dependent on the correct levels of calorie and nutrition intake, and thus, any disruption to this is likely to have a knock-on effect (Michalakis et al., 2013).

1.3 Biology of Male Fertility

To understand and infer the potential mechanisms of male infertility, it is important firstly to understand the basics of male fertility and normal spermatogenesis. The human testes have two main functions, to produce hormones such as testosterone via the HPG axis, and to produce

male gametes via a process known as spermatogenesis – a single round of which takes approximately 74 days (Shalet, 2009; Cannarella et al., 2019). Spermatogenesis is a highly complex process regulated by a number of specific hormones and cells whereby undifferentiated diploid spermatogonial cells slowly evolve into the haploid cells (spermatozoa) (Sharma & Agarwal, 2011). Much of the volume of the testis consists of a series of convoluted tubules known as the seminiferous tubules enervated by blood vessels, nerves, lymphatic vessels and Leydig cells. The seminiferous tubules are split into three distinct layers based on the cells present at a given position, these are the basal compartment, the adluminal compartment and the lumen (Figure 1.2). The process of spermatogenesis is and driven through temporal regulation by the HPG axis (Singh et al., 2017; Clavijo & Hsiao, 2018; Kaprara & Huhtaniemi, 2018). In males, this axis consists of the hypothalamus, the pituitary gland and the testis (gonads). The hypothalamus secretes gonadotropin-releasing hormone (GnRH) in a pulsatile manner (Spaziani et al., 2021). GnRH is a key regulator in initiating the reproductive cascade by stimulating the release of gonadotropins from the pituitary gland. The GnRH binds to gonadotroph cells within the pituitary gland which then initiates the release of both follicle-stimulation hormone (FSH) and luteinizing hormone (LH) into the reproductive tract. LH is essential for the production of testosterone within the testis, with its function being bind to luteinizing hormone receptors (LHRs) and to stimulate the production of the hormone within the Leydig Cells in a pulsatile manner (Zirkin & Papadopoulos, 2018). The hormone testosterone plays many roles within the testis and other tissues including the acquisition of secondary sex characteristics, differentiation of the male gonads and the initiation and maintenance of spermatogenesis (Sharma et al., 2021). . Testosterone also plays an important part in the feedback regulation of the HPG axis itself, with elevated levels of testosterone exerting negative feedback on the GnRH secreting neurons, leading to decreased GnRH secretion (Bliss et al., 2010; Kaprara & Huhtaniemi, 2018). In response to LH, Leydig cells also produce a small amount of estrogens (Laan et al., 2012). Follicle-stimulation hormone acts within the seminiferous tubules themselves, stimulating Sertoli cells to induce the proliferation and differentiation of the Sertoli cells as well as promoting the secretion of the androgen binding protein (ABP) (Wang et al., 2022; Norman & Henry, 2015). Sertoli cells make up around 40 % of the seminiferous tubule and are essential for successful spermatogenesis; providing structural, functional, and metabolic support to the surrounding germ cells and are found in all three compartments of the seminiferous tubule (O'Donnell et al., 2022). Sertoli cells are often referred to as "nurse cells" due to their function in providing the necessary nutrients and support for germ cells as they progress from the less developed spermatogonia found adjacent to the basal membrane into the differentiated spermatozoa found in the lumen (França et al., 2016).

Additionally, these nurse cells regulate the levels of FSH secretion by producing the hormone inhibin B via a negative feedback mechanism acting on the pituitary gland (Pierik et al., 1998). The successful regulation of this axis during puberty is essential for the development of primary and secondary sexual characteristics in both males and females (Maggi et al., 2016; Singh et al., 2017; Spaziani et al., 2021).

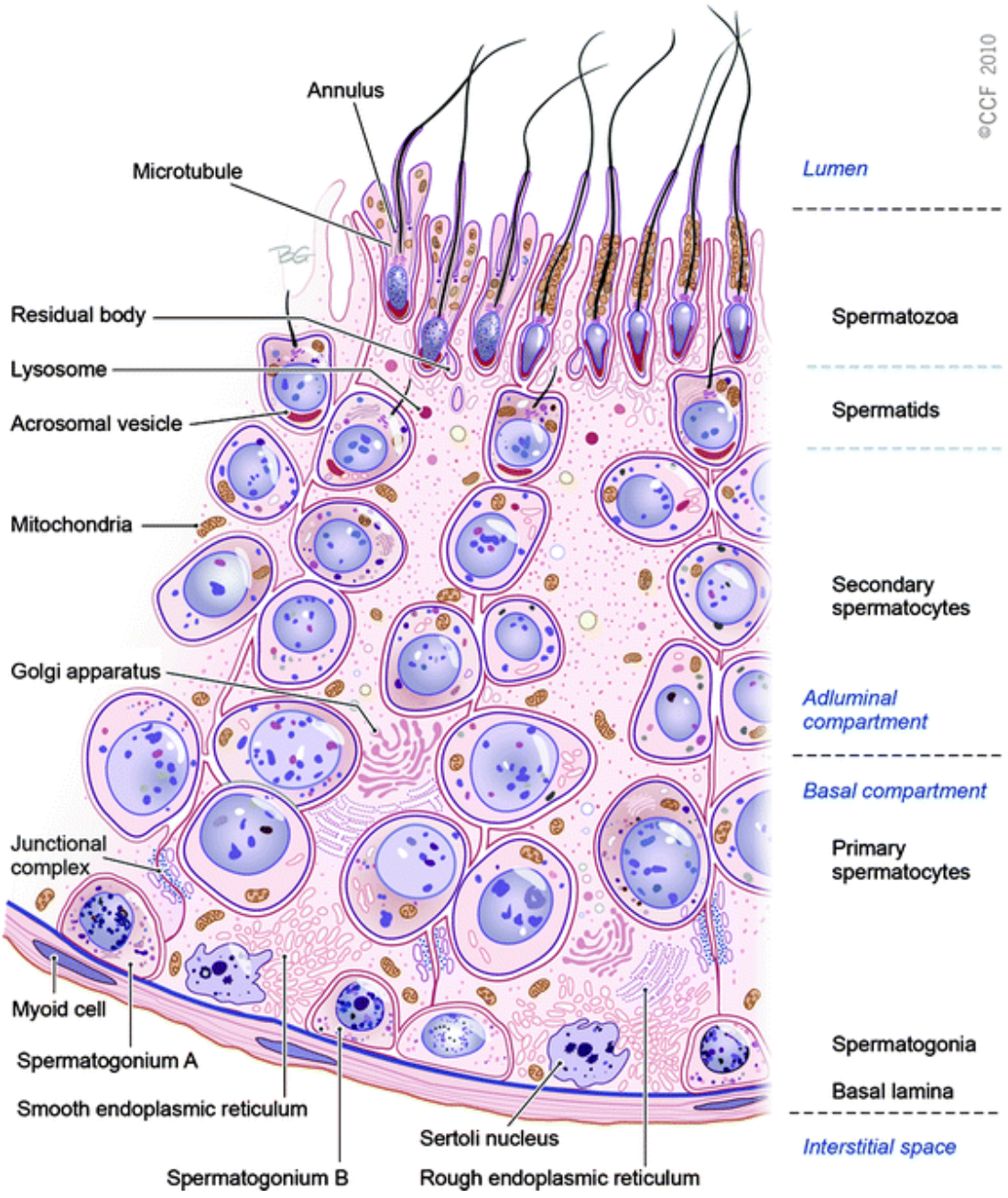


Figure 1.2: Section of the germinal epithelium in the seminiferous tubule. Sertoli cells divide the germinal epithelium into a basal and adluminal compartment, via the Sertoli cell. Spermatozoa are released into the lumen. Reproduced with permission from Springer Nature (Sharma & Agarwal, 2011)

The initial stage of spermatogenesis is defined as the proliferation and differentiation of the diploid spermatogonia. Spermatogonia can be split into two different subgroups, the stem cell type A spermatogonia, and the differentiated type B spermatogonia. Type A spermatogonia undergo mitosis in order to either produce more type A spermatogonia for stem cell population replenishment or differentiate into type B spermatogonia. The type B spermatogonia then undergo another round of mitosis to produce diploid primary spermatocytes. At this point, all proliferation has occurred within the basal compartment of the seminiferous tubule (Figure 1.3). During meiosis I, the primary spermatocytes move from the basal compartment of the tubule into the adluminal compartment and become haploid secondary spermatocytes. Meiosis II then occurs with the secondary spermatocytes transitioning into round spermatids. Spermatids then mature, undergoing structural and nuclear changes and forming the head and tail of the spermatozoa (Muñoz et al., 2015). The differentiated spermatids then progress towards the lumen, detaching from the Sertoli cells and become free spermatozoa (Shalet, 2009). A final stage of maturation occurs for these spermatozoa after release into the lumen, with the cells transitioning into the epididymis to acquire motility required for successful reproduction (Sharma & Agarwal, 2011).

Given the complexity of the process of spermatogenesis, with multiple stages of proliferation and differentiation, it is easy to see how MI presents with such heterogeneous phenotypes. Arrest or disruption at any of the distinct stages of the process can often lead to the same conclusion of MI however upon further investigation, the pathology can differ hugely. For this reason, the more severe NOA is often further classified using histological samples as caused by either Hypospermatogenesis (HS), Germ Cell Arrest (GCA) or Sertoli Cell Only syndrome (SCO) (Cioppi et al., 2021). Hypospermatogenesis is seen when all stages of spermatogenesis are present, but the numbers of the cells are decreased whereas GCA and SCO represent the disruption of spermatogenesis at various stages and are highlighted by the absence of either some or all germ cells in the seminiferous tubules (Figure 1.3). SCO is arguably the most severe of all male infertility phenotypes with no germ cells present (Leslie, Mejias, et al., 2023).

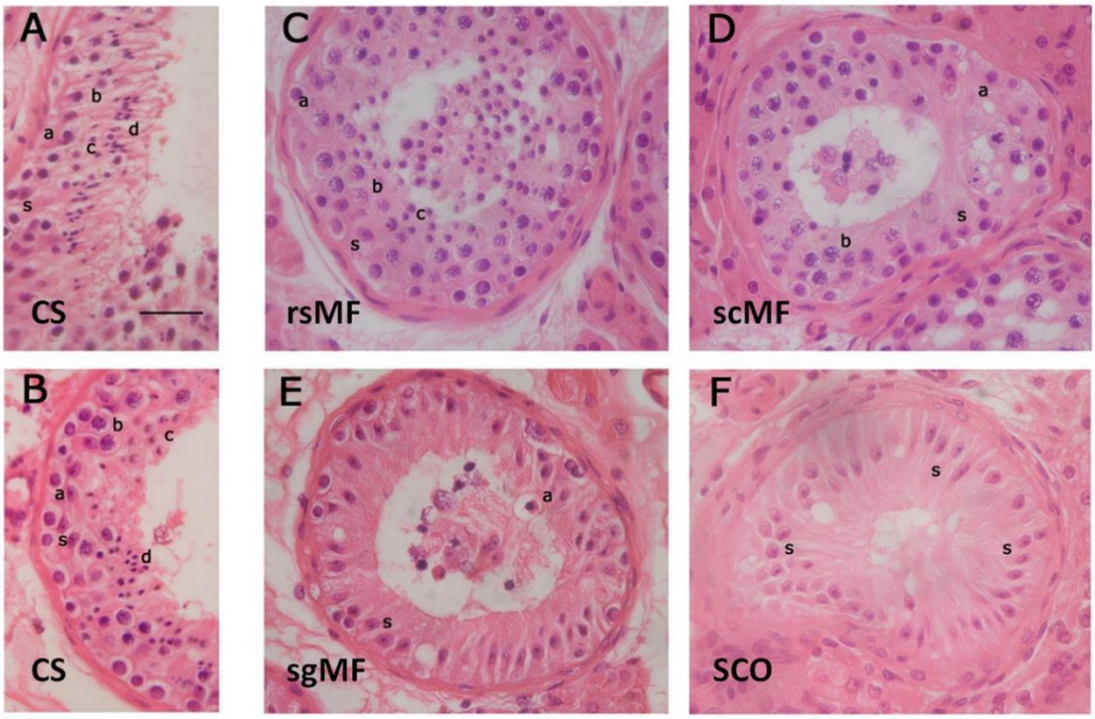


Figure 1.3: Testicular histology of representative sections of seminiferous tubules from infertile men showing different spermatogenic patterns, conserved spermatogenesis (CS) containing all germ-cell stages *A,B),* maturation failure at the round spermatid stage (rsMF) showing no elongated spermatids *C),* maturation failure at the spermatocyte stage (scMF) with spermatogonia and spermatocytes but few or none spermatids *D),* maturation failure at the spermatogonia stage (sgMF) displaying only these primordial cells *E)* and Sertoli cell-only phenotype or germ-cell aplasia (SCO) where seminiferous tubules contain exclusively Sertoli cells *F).* Examples of different cell types, including Sertoli cells and specific germ-cell stages, are identified with lower case letters: *a)* spermatogonia; *b)* primary spermatocytes; *c)* round spermatids; *d)* elongated spermatids; *s)* Sertoli cell nucleus. Haematoxylin and eosin staining. Bar = 50 μm. (Muñoz et al., 2015)

1.4 Non-Genetic Causes of Male Infertility

Upon initial presentation at a fertility clinic, males are screened for a number of non-genetic conditions which are known causes of MI. The leading correctable cause of MI is varicocele (K. Wang et al., 2022). Varicocele is the abnormal dilation and enlargement of the scrotal venous pampiniform plexus, tasked with draining the blood from each testis (Leslie, Sajjad, et al., 2023). The prevalence of varicocele in the general population is around 15 % - 20 % of men, but this rises to around 40 % of all men undergoing fertility investigations (de Sousa Filho et al., 2018; Leslie, Sajjad, et al., 2023). In men presenting with varicocele, low sperm count can be seen with poor sperm motility and an increase in the number of abnormal sperm (Xue et al., 2012). It is suggested that an increase in reactive oxygen species can occur in men with varicocele leading to oxidative stress and subsequently, sperm dysfunction (Wang et al., 2022).

Endocrine disorders are another known cause of MI, presenting in around 2 % - 5 % of all cases, with disturbance to the essential hormones driving spermatogenesis causing a diminishing effect on the fertility of males. Hypogonadotropic hypogonadism (HH) is a condition seen in infertile males which is caused by a dysfunction of the HPG axis (Morris et al., 2021; Sang et al., 2023). This form of secondary gonadotrophin deficiency results from a number of acquired factors which disrupt the normal function of either the hypothalamus or the pituitary gland. The dysfunction of the HPG axis at this stage leads to deficient secretion of GnRH and, therefore, a subsequent deficiency in both LH and FSH resulting in reduced fertility (Kumar Mahat & Arora, 2016). Whilst HH can be due to genetics in the case of congenital hypogonadotropic hypogonadism (CHH), there are many non-genetic causes resulting in acquired HH. For example, tumours affecting the hypothalamus and pituitary gland can disrupt the function of the HPG axis physically through compression or infiltration of the glands (Famini et al., 2011). In the case of tumours such as prolactinomas, hormonal secretions from the tumour can cause inhibition of the secretion of GnRH (Brue & Delemer, 2007). Other causes of HH include traumatic brain injury, inflammation or infections such as meningitis, damage caused by radiation therapy, certain hormonal and steroidal medications and some chronic illnesses such as chronic kidney disease (Constine et al., 1993; Hohl et al., 2009; Rahnema et al., 2014; Garibotto et al., 2021). Separately, infections such mumps (viral) and STDs (bacterial) can lead to orchitis, an inflammation in either one or both testicles. This inflammation has been shown to cause testicular atrophy, leading to a decrease in sperm count, mobility, and morphology (Masarani et al., 2006).

1.5 Genetic Causes of Male Infertility

Whilst the majority of diagnosed MI cases are due to a combination of environmental risk factors and acquired conditions, as previously mentioned, around 15 % of all cases are thought to be caused by a genetic factor (Krausz, 2011). A correlation has been identified between the severity of the disease and the percentage of cases currently attributed to a causative genetic factor, with a 25 % risk of genetic abnormalities amongst azoospermia sufferers (Riera-Escamilla et al., 2022) (Figure 1.4). As the severity of the infertility decreases, so does the contribution of genetics to the disease with a decrease in monogenic contribution as the sperm count of a patient increase.

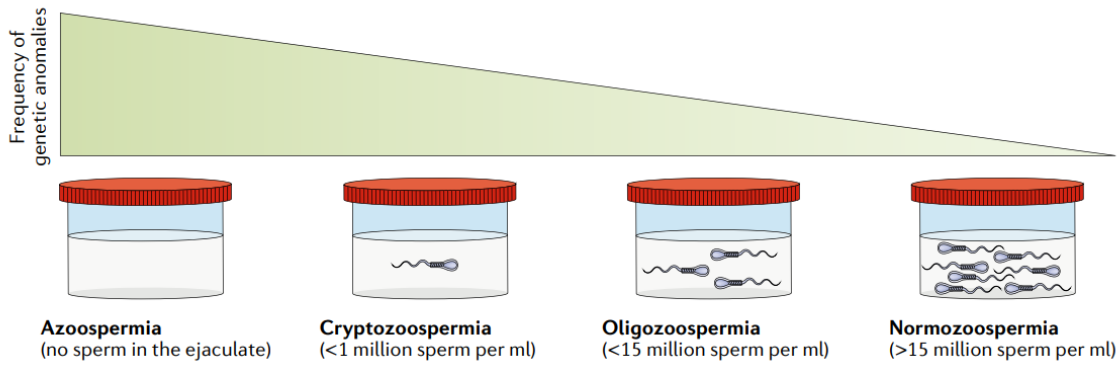


Figure 1.4: The different types of quantitative disturbances of spermatogenesis and the frequency of genetic factors in each category. An inverse correlation exists between sperm count and prevalence of genetic anomalies (Krausz & Antoni Riera-Escamilla, 2018)

1.5.1 Chromosomal Abnormalities in Male Infertility

Chromosomal abnormalities are one of the most common genetic causes of MI and are identified in around 5 % of all men with oligozoospermia, and up to 15 % of men presenting with azoospermia (Flannigan & Schlegel, 2017). For this reason, men with significantly decreased sperm count will undergo karyotyping and other chromosomal analyses to identify any potential abnormalities.

Klinefelter syndrome (KS) is an aneuploidy of the X chromosome resulting in a 47, XXY phenotype and is the most common chromosomal abnormality causing NOA (up to 15 % of NOA patients) (Vockel et al., 2021). This syndrome occurs due to nondisjunction of homologous chromosomes during meiosis I or chromatid pairs during mitosis or meiosis II, leading to the presence of an extra X chromosome (Flannigan & Schlegel, 2017). Klinefelter syndrome was first described in 1942 as a condition characterised by gynecomastia, small testes and severe spermatogenic disruption (Klinefelter et al., 1942), it was not until karyotyping was introduced in the late 1950s that the 47, XXY karyotype was determined as the cause (Jacobs & Strong, 1959). Males suffering from KS are found to have small firm testes with hyalinization of the seminiferous tubules leading to spermatogenic failure (Vockel et al., 2021).

A second chromosomal abnormality known to lead to MI a disorder of sex development (DSD), 46, XX syndrome, also known as 46, XX male syndrome. This syndrome was first reported in 1964 and occurs in only 1 in 20,000 newborn males (Vorona et al., 2007; Chapelle et al., 1964). Whilst presenting with an XX phenotype, around 90 % of all males with this disorder are *SRY* positive due to a translocation of the *SRY* gene onto the X chromosome during meiosis, causing the male presentation and infertility (Capron et al., 2022). Patient phenotypes commonly

include undescended testis, short stature, gynaecomastia and always present with azoospermia (Röpke & Tüttelmann, 2017).

Robertsonian translocations (RT's) occur in around 0.9 % of all infertile men and in around 4 % – 8 % of all oligozoospermic men (Krausz & Riera-Escamilla, 2018). The main characteristic of these translocations is a loss of genetic material due to the fusion of two long arms from acrocentric chromosomes and the loss of genetic material on the short arms. The most common RT is the 13q;14q translocation, resulting in impaired spermatogenesis and increased rates of sperm aneuploidy (Almesned et al., 2020). Large autosomal inversions can also cause male infertility with a chromosome 9 inversion thought to be responsible for between 3 % - 5 % of MI cases with the severity of the phenotype varying from oligozoospermia to asthenozoospermia (Flannigan & Schlegel, 2017).

After KS, Y chromosome microdeletions (YCMD) are the most common genetic cause of MI, leading most often to azoospermia or severe oligozoospermia and are found in around 2 % of all infertile males (de Sousa Filho et al., 2018). In 1991, it was discovered that the distal portion of the Y chromosome contained a region with suggested involvement in spermatogenesis (Bardoni et al., 1991). In 1996, the region was mapped, and three separate loci were identified as the AZFa, AZFb and AZFc regions (Vogt et al, 1996). It was found that microdeletions in these AZF regions were present in an increased rate in infertile patients at as. The region as a whole contains multiple genes which are expressed in the testis and are essential for successful spermatogenesis with at least 14 critical protein coding genes (Figure 1.5) (Neto et al., 2016).

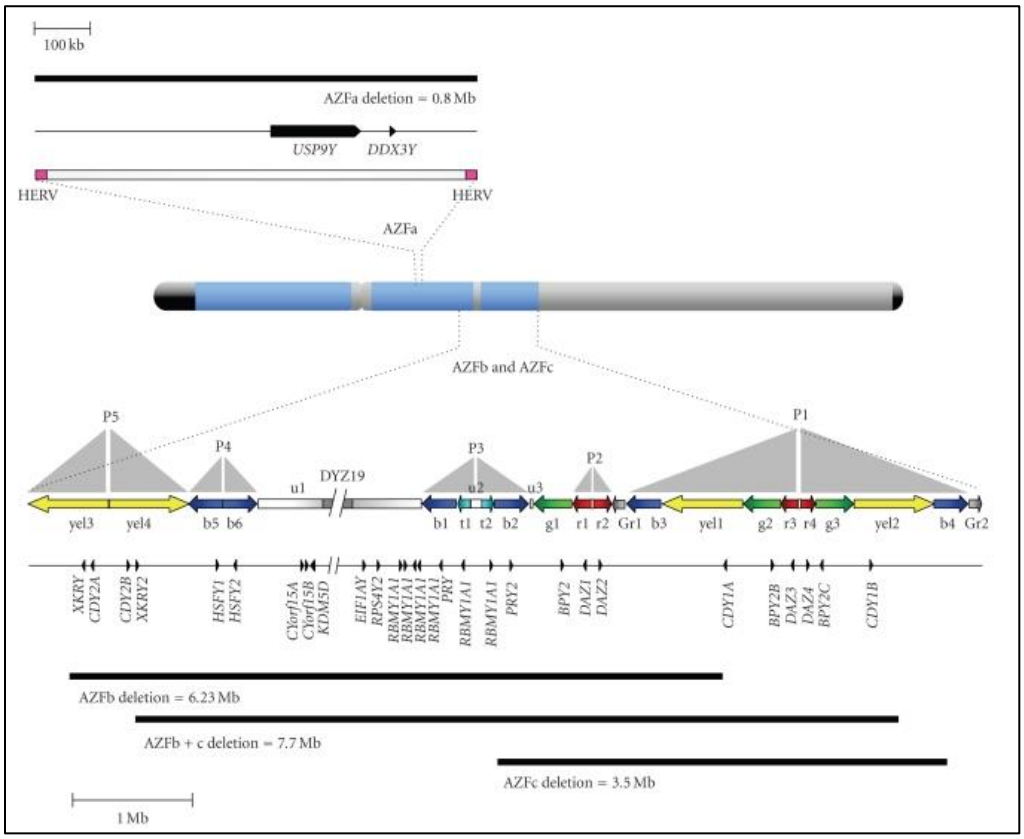


Figure 1.5: Schematics on the genomic architecture and gene content of the reference AZFa, AZFb and AZFc regions of the human Y chromosome. (Navarro-Costa et al., 2010)

Microdeletions can occur in one or multiple of the AZF regions resulting in a wide array of infertility phenotypes depending on the genes affected. Around 60 % of all Y chromosome microdeletions occur within the AZFc region at the distal aspect of the chromosome (Neto et al., 2016). This region contains the *DAZ* gene which has been implicated in MI. Microdeletions in the AZFc region are thought to be the least severe with phenotypes ranging from azoospermia to oligozoospermia (Nickkholgh et al., 2015). The AZFb region is implicated in around 15 % of all Y chromosome microdeletions with the phenotypic consequences of full deletions typically being azoospermia with testis histology showing SCO or GCA (Flannigan & Schlegel, 2017). The AZFb region contains the gene *RBMY1* which is expressed in the nucleus of germ cells from spermatogonia to round spermatids. *RBMY1* is a testis specific splicing factor and thus deletion of this gene provides severe consequences for the process of spermatogenesis (Yan et al., 2017). The most severe type of YCMD are complete deletions of the AZFa region. This only occurs in around 3 % of all YCMD cases but all patients present with azoospermia due to SCO syndrome with no viable sperm produced (Neto et al., 2016). The AZFa region is home to

the genes *UTY*, *USP9Y* and *DDX3Y,* the latter of which has been shown to play essential roles in spermatogenic regulation within the spermatogonia (Flannigan & Schlegel, 2017).

1.5.2 Monogenic Causes of Male Infertility

Nearly 10 years prior to the discovery of the AZF region on the Y chromosome, the first gene on the X chromosome was linked to male infertility; the *androgen receptor gene* (*AR*) (Brown et al., 1988). Mutations in the *AR* gene lead to a variety of issues collectively known as Androgen Insensitivity Syndrome (AIS) with over 1000 different *AR* mutations having been described (Gottlieb et al., 2012). AIS is, like 46, XX syndrome, a form of DSD (Gulía et al., 2018). The *AR* gene is critical for developmental processes and is a known DNA-binding transcription factor (Vockel et al., 2021). Androgen insensitivity syndrome can be further divided into complete androgen insensitivity (CAIS), partial androgen insensitivity (PAIS) and mild androgen insensitivity (MAIS) (Gottlieb & Trifiro, 1993). Genetic males (46, XY) who present with CAIS appear to have normal female external genitalia however lack internal female organs such as ovaries and a womb, often presenting with undescended testes. In contrast, patients who suffer from PAIS and MAIS have varying degrees of masculinization of the external genitalia and typically suffer from gynaecomastia, sparce facial and body hair and impaired spermatogenesis (Hughes et al., 2012). Around 1.5 % - 2.0 % of infertile males are identified as having *AR* mutations (Ferlin et al., 2006; Rocca et al., 2023).

Congenital hypogonadotropic hypogonadism (CHH) is a rare condition which effects as few as 1 in 86,000 males (Laan et al., 2021). Like HH, this is a DSD in which a failure of the secretion of GnRH leads to delayed puberty and, subsequently, infertility (Young et al., 2019; Grinspon et al., 2020; Laan et al., 2021). CHH patients often present with a lack of secondary sexual characteristics and, in certain sub forms of CHH such as Kallmann syndrome, an impaired sense of smell (Ichioka et al., 2024). In 1991, the X chromosome gene “anosmin-1” (*ANOS1*) was identified as deleted in patients suffering from Kallmann Syndrome (Legouis et al., 1991; Franco et al., 1991). The gene is an essential gene for development of the olfactory system and migration of GnRH neurons, and mutations within *ANOS1* have been shown to cause anosmia and the reduced secretion of GnRH (C.I. Gonçalves et al., 2017; de Castro et al., 2017). A decade later, mutations were identified in a second gene “fibroblast growth factor receptor 1” or *FGFR1* as autosomal dominant (AD) causes of Kallman syndrome (Dodé et al., 2003). Other genes known to be associated with Kallmann syndrome or CHH are *CHD7* (AD), *GNRHR* (AR) and *PROK2* (AR) (Sarfati et al., 2010; Catarina I Gonçalves et al., 2017; Balasubramanian & Crowley, 2017). Mutations in the gene *SOX9* have been strongly associated with DSD in males

as the gene plays an important role in male sexual development and mutations lead to testicular dysgenesis (Katoh-Fukui et al., 2015). The gene *NR5A1* is a known transcriptional factor which has a critical role in the development and function of the gonads. NR5A1 is mostly expressed within the Sertoli and Leydig cells of the developing testis and the Sertoli cells of adult testis (Bashamboo et al., 2010). *NR5A1* is also responsible for the transcriptional regulation of several genes involved in hormone production, sexual development and reproduction (Lin & Achermann, 2008). Variants in this gene have been identified as the second most common genetic variation responsible in cases of 46,XY DSD (Camats et al., 2012; Fabbri-Scallet et al., 2020). Despite most commonly being associated with DSD, recent studies have identified patients suffering from NOA and severe oligozoospermia with predicted pathogenic variants in *NR5A1*, mirroring the heterogeneity seen widely in male infertility (Rocca et al., 2020).

Shortly after the discovery of *AR* mutations in infertile males, the *Cystic Fibrosis Transmembrane Conductance Regulator gene* (*CFTR*) on chromosome 7 was discovered to house mutations underlying the disease Cystic Fibrosis (Riordan et al., 1989; Kerem et al., 1989). Over the course of the next few years, biallelic mutations in *CFTR* were identified in a number of cases of MI, specifically in patients who suffer from azoospermia as a result of Congenital Bilateral Absence of the *Vas Deferens* (CBAVD) (Dumur et al., 1990; Culard et al., 1994). The *vas deferens* are the ducts that transport sperm from the testis to the ejaculatory ducts during ejaculation. Absence of these ducts in cases of CBAVD is classified as a ductal obstruction and results in obstructive azoospermia (Persily et al., 2021). Mutations in the *CFTR* gene can produce highly variable phenotypes with over 2000 *CFTR* variants currently identified ranging from severe to mild based on their phenotypic presentation (De Boeck, 2020). Typically, males presenting with azoospermia due to CBAVD will have bi-allelic *CFTR* mutations that are not in themselves causative of Cystic Fibrosis but rather, *CFTR* dysregulation, leading to the milder, CBAVD phenotype (Bieth et al., 2021). Interestingly, in around 20 % of all cases of CBAVD, no *CFTR* mutations are found (Yu et al., 2012). Potentially pathogenic variants have, however, been reported in the X chromosome gene *ADGRG2* at an increased frequency in males with *CFTR* negative CBAVD (Table 1.1) (Patat et al., 2016; Yang et al., 2017).

One year prior to the discovery of the AZF regions, deletions in a region on the Y chromosome were found in a number of azoospermic males and absent in healthy controls. Within this region the *DAZ* (Deleted in Azoospermia) gene was identified (Reijo et al., 1995). The *DAZ* gene is expressed within pre-meiotic germ cells and plays a vital role in RNA-binding during

spermatogenesis (Reynolds & Cooke, 2005). The year 2015 saw the publication of two separate papers in which mutations in the testis-expressed 11 gene (*TEX11)* were identified as a potential X-linked causes of azoospermia in infertile males (Yang et al., 2015; Yatsenko et al., 2015). Both studies identified loss-of-function (LoF), splice site and missense variants in TEX11 and conclude that disruption of this gene leads to meiotic arrest and NOA.

In 2003, *SYCP3* was identified as a monogenic cause of male infertility with mutations in this gene leading to meiotic arrest during spermatogenesis. This was the first of only four currently confirmed AD male infertility genes with NOA or oligozoospermic phenotypes, the other three being *DMRT1*, *KLHl10* and *SYCP2* (Oud et al., 2019; Miyamoto et al., 2003; Schilit et al., 2020; Yatsenko et al., 2006; Lopes et al., 2013).

A 2014 paper investigating the rare form of qualitative male infertility, MMAF, identified five subjects who presented with homozygous mutations in the *DNAH1* (Dynein Axonemal Heavy Chain 1) (Ben Khelifa et al., 2014). DNAH1 is expressed in the testis and the gene encodes an axonemal inner dynein heavy chain which helps to drive proper flagellar movement (Jiang et al., 2021). Mutations in *DNAH1* are found to disrupt the function of the dynein complex and cause defects in sperm flagellar structure and motility (Ben Khelifa et al., 2014; Amiri-Yekta et al., 2016; Jiang et al., 2021).

	Isolated Infertility					Endocrine Disorder-related Infertility	Syndromic Infertility
Phenotype	**CBAVD**	**NOA**	**Severe Oligozoospermia**	**Globozoospermia**	**MMAF**	**CHH**	**PCD**
Phenotypic presentation	Bilateral absence of the *vas deferens*	Complete absence of sperm in the ejaculate	Severe depletion of sperm (0.1 - 5 million sperm per ml)	Sperm present with round heads and acrosomal malformation or absence	Sperm present with a variety of irregularities of the flagella	Incomplete development of secondary sexual characteristics and azoospermia	Sperm motility is severely reduced due to malfunction of motile cilia in sperm axoneme
Prevalence in infertile males	1 - 2 %	10 - 15 %	Up to 30 %*	< 0.1 %	< 0.1 %	< 1 %	< 0.1 %
Number of genes with moderate-definitive linkage	2	10	4	1	14	34	8

Table 1.1: Summary table representing the key forms of male infertility, their prevalence among infertile patients and the number of genes causatively linked to the phenotype. CBAVD: Congenital Bilateral Absence of the Vas Deferens, NOA: Non-Obstructive Azoospermia, MMAF: Multiple Morphological Abnormalities of the Flagella, CHH: Congenital Hypogonadotropic Hypogonadism, PCD: Primary Ciliary Dyskinesia. Data regarding number of genes linked to each phenotype at a moderate to severe level is adapted from a 2021 review on monogenic causes of male infertility (Houston et al., 2021). Statistics for phenotype prevalence in infertile males was collected from multiple sources: (Lin & Huang, 2020; Millar et al., 2021; Fakhro et al., 2018; Huang et al., 2022; Laan et al., 2021)
* This value includes cases of severe OligoAsthenoTeratozoospermia as often patients present with both qualitative and quantitative sperm defects upon diagnosis. (Tüttelmann et al., 2018)

1.5.3 Whole Exome Sequencing for Monogenic Disease Gene Discovery in Male Infertility

Since the commercial introduction of Next Generation Sequencing (NGS) technology in the mid-2000s, a large number of genes have been implicated in MI due to the huge expansion in sequencing capabilities this provided (Xavier et al., 2021). The introduction of whole exome sequencing (WES) meant that instead of genes or small regions of a chromosome being studied, the entire coding sequence of a patient's DNA could now be investigated, opening up a whole new method of disease gene research. In specific rare MI aetiologies such as MMAF, whole exome sequencing has been instrumental in identifying a number of causative recessive genes (Touré et al., 2021). The use of WES has led to identification of additional homozygous mutations in the known MMAF gene *DNAH1* as well as two novel genes involved in sperm axonemal assembly *CFAP43* and *CFAP44* (Coutton et al., 2018). It is now thought that up to 60 % of all MMAF cases can be explained by mutations in 18 genes which have been implicated within the last 10 years, largely through WES (Touré et al., 2021; Oud et al., 2021). For more "common" aetiologies, such as non-obstructive azoospermia, investigation into causative genes has been more difficult due to the enormous heterogeneity of the disorder and the wide array of phenotypes as a result. A small number of genes have been identified in azoospermic men utilising WES over the last 10 years, with *TEX15* and *MEIOB* being two such examples (Okutman et al., 2015; Gershoni et al., 2019). Both of these studies were performed, however, in singular families and therefore potentially only explain a very small proportion of azoospermia cases. Within the last few years, multiple studies with international collaboration between MI research groups have been published with the aim of progressing disease gene discovery via analysis of multiple large cohorts (Appendix A). In particular, the International Male Infertility Genomics Consortium (IMIGC) was set up with the aim of gathering and sharing genomic datasets from large cohorts of infertile males to help accelerate replication of novel genomic findings. In 2020, a cohort of 58 males with unexplained meiotic arrest underwent whole exome sequencing. A homozygous frameshift variant c.676dup was identified in three unrelated males in the gene *M1AP* (meiosis 1 associated protein) (Wyrwoll et al., 2020). Utilising the IMIGC collaboration, WES datasets from an additional 2000 infertile NOA males in four different European cohorts were then screened for *M1AP* variants. In total, six additional infertile males were found to possess likely causal bi-allelic variants in *M1AP* providing strong evidence for the role of M1AP as an autosomal recessive (AR) NOA-associated disease gene. The role of recessive variants causing NOA was further explored in a 2022 paper published by the GEMINI group (Nagirnaja et al., 2022). A cohort of 924 infertile males with a NOA

phenotype underwent WES to identify rare, likely causative recessive mutations. In 20 % of all these cases (n = 178) a potential molecular cause was identified. After assessing the WES of a further 1,148 infertile males from additional IMIGC cohorts, a further 21 genes were also implicated. Within this study, 11 patients were affected by rare recessive variation in six piRNA biogenesis genes (*PLD6, PNLDC1, RNF17, TDRD9, TDRD12, TDRKH*) of which two have previously been found disrupted in infertile males (Arafat et al., 2017; Nagirnaja et al., 2021, 2022). piRNAs are small non-coding RNA molecules which are abundant in male germ cells and are critical for the survival of the germ cell pool (Mann et al., 2022; Nagirnaja et al., 2021). The work in this paper indicates the importance of large-scale unbiased WES studies in infertile cohorts as well as pathway analyses of disease genes in order to identify the underlying biological mechanisms disrupted in different male infertility phenotypes. An additional paper in 2021 performed WES on 1305 azoospermic males, 90 of whom specifically presented with meiotic arrest (MeiA). In two MeiA males and four azoospermic males, likely pathogenic homozygous variants were identified in the gene *MSH5*. Additional biallelic mutations were identified in a further two men with MeiA in the closely related *MSH4* gene. MSH4 and MSH5 form a heterodimer which is known to be required for prophase of meiosis I during spermatogenesis. Functional studies performed in HEK293 cells revealed an arrest in early prophase of meiosis I in individuals with pathogenic *MSH4* or *MSH5* variants. This study reported the first known variants in the gene *MSH5* causing male infertility (Wyrwoll et al., 2021).

Almost three decades after the initial discovery of the three separate AZF loci, WES data for 1,600 NOA males has been used to identify the key gene located in the AZFa region which is indispensable for successful germ cell maturation (Dicke et al., 2023). Four different, likely pathogenic, LoF variants were identified in the AZFa gene *DDX3Y* in four individuals with three of the patients showing the typical AZFa SCO phenotype. Given that YCMDs of the AZFa region present with the most severe phenotype, and both full and partial deletions of the gene *USP9Y* has been identified in normozoospermic males (Alksere et al., 2019; Luddi et al., 2009), *DDX3Y* was an ideal candidate for the key spermatogenic factor in this region. The identification of these highly damaging mutations in NOA patients further corroborates the key role for *DDX3Y* in NOA.

A study published in 2019 aimed to consolidate the findings of all published research into monogenic male infertility disease gene discovery, giving a list of all definitive and potential causative genes identified based on the gene-disease relationships and their clinic validity (Oud et al., 2019). In total, 78 genes were identified with at either a moderate, strong or a definitive

link to 92 different MI phenotypes. Of the 24 genes identified in patients with isolated MI, only 4 (17 %) showed an AD inheritance pattern with the majority (n = 16, 66 %) showing recessive inheritance. An additional four variants were X-linked (17 %). An update to this review was published in 2021 with an expanded list of genes linked to MI (Figure 1.6) (Houston et al., 2021). An additional 30 genes were identified in novel literature with at least a moderate link to male infertility. Of these 30 novel genes, 12 were related to isolated infertility with 11 novel AR genes and one AD gene. This novel AD finding was associated with isolated MI with an MMAF phenotype.

This study, as well as others, highlights how despite 15 % of all cases of MI having a likely genetic cause, current diagnostic rates in these males is only around 4 % - 9 % (Tüttelmann et al., 2018; Olesen et al., 2017). This highlights the need for further research into the genetic causes of this disease, with larger cohort sizes and further NGS gene discovery studies needed to identify and validate novel causative genes in order to reach the genetic diagnostic levels of other heterogenous diseases such as Intellectual Disability (ID) (Vissers et al., 2016). A pilot study performed by the 100,000 genomes project highlighted the potential of using WES in idiopathic rare disease cases with a diagnostic yield of 40 % - 55 % for specific rare conditions such as ID and hearing or vision disorders. Likely causative variants were also identified in 30 % of all patients presenting with an idiopathic case of a rare disease with a typically monogenic cause (Smedley et al., 2021). Unfortunately, MI was not one of the rare diseases assessed in this study. The success rate of WES seen in similar conditions to MI such as ID strengthens the argument for NGS utilisation in idiopathic infertile males.

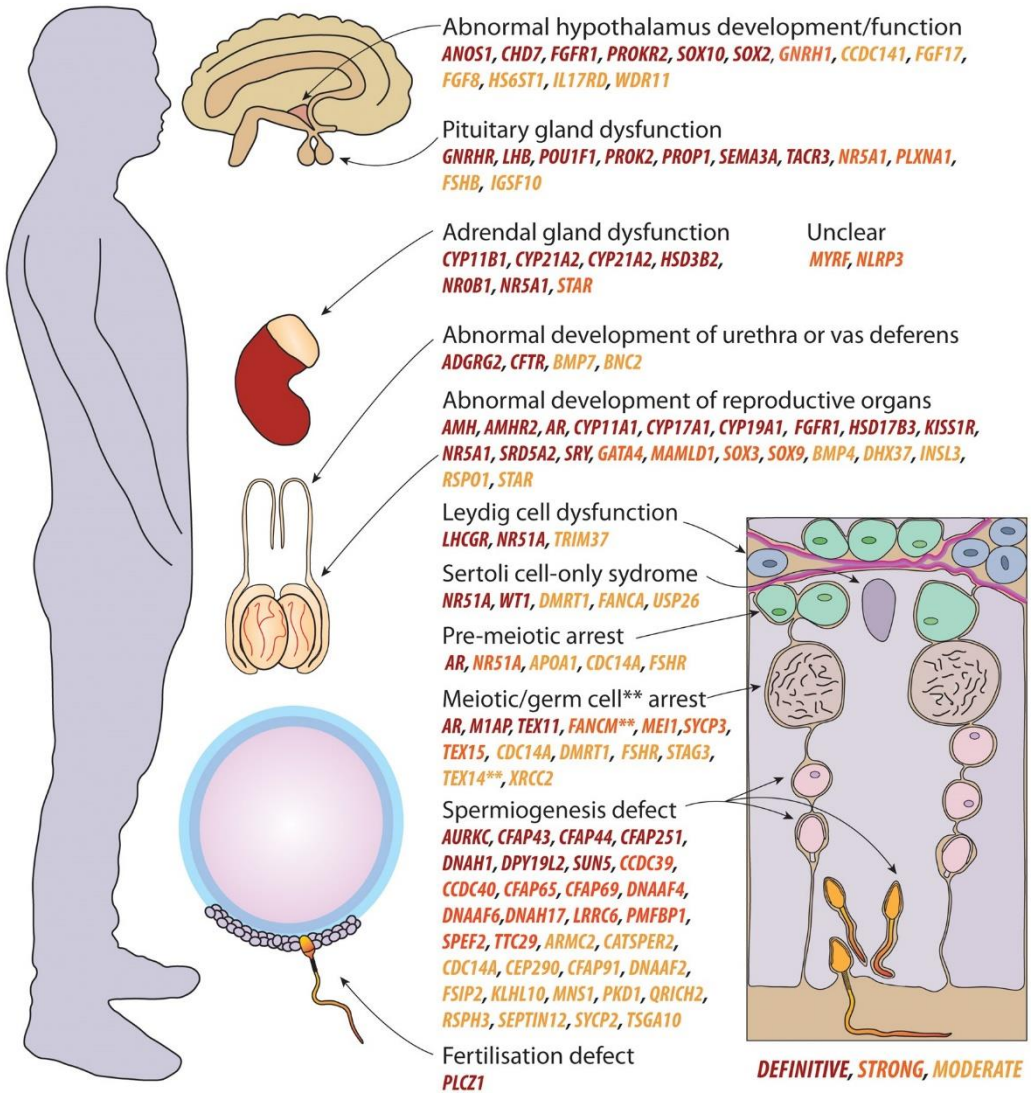


Figure 1.6: Overview of all genes associated with male infertility phenotypes at an organ/cell level. Definitive genes are labelled in red, strong in orange and moderate in yellow. Organs top to bottom: brain, adrenal gland (and kidney), testes and epididymis with *vas deferens*. Bottom left: sperm fertilising an oocyte surrounded by cumulus cells. Right: seminiferous tubule cross-section. Leydig cells (blue), Sertoli cells (purple), basement membrane (pink), spermatogonia (green), spermatocytes (brown) and spermatids (pink and orange). ** denotes germ cell arrest gene for *FANCM* and *TEX14*. Genes were classified as unclear when they were not clearly linked to a specific organ. (Houston et al., 2021)

1.6 *De novo* Mutations (DNMs) as a Model for Dominant Disease

1.6.1 De novo mutations in rare disease

DNMs are spontaneous mutations which can occur either during gametogenesis or post zygotically and are therefore not present in either parent (Veltman & Brunner, 2012). Multiple studies investigating DNMs within the human genome have concluded that in the average human genome there are around 44 - 89 DNMs with around 1 - 2 of these lying within the coding regions of the genome (Neale et al., 2012; Kong et al., 2012; Francioli et al., 2015; Acuna-Hidalgo et al., 2016). It is also known that the number of DNMs present per genome increase with paternal age with approximately 1 - 1.5 additional DNMs present per year (Goldmann et al., 2016; Jónsson et al., 2017; Kong et al., 2012; Smits et al., 2022; Michaelson et al., 2012). Over the course of the last 10-15 years, DNMs have been recognised as contributors to a wide range of human disorders. These include disorders such as Autism Spectrum Disorder (ASD), Intellectual Disability (ID), Schizophrenia, Retinitis Pigmentosa, Amyotrophic Lateral Sclerosis (ALS) and Congenital Heart Disease (CHD) (Georgieva et al., 2014; Müller et al., 2022; Martin-Merida et al., 2019; Satterstrom et al., 2020; Jin et al., 2017; Deciphering Developmental Disorders Study, 2017; Hamdan et al., 2017; De Vas et al., 2023). Most notably, as of 2023, it is thought that around 40 % of all cases of ID can be explained by DNMs in protein-coding genes (De Vas et al., 2023). Due to the spontaneous nature in which DNMs arise, the typical process of evolutionary selection that occurs with inherited variation has not occurred (Roach et al., 2010; Acuna-Hidalgo et al., 2016). This phenomenon gives DNMs the potential to be more deleterious and the ability to produce stronger phenotypic effects.

Across multiple DNM studies, a consistent finding is the rate at which damaging missense or LoF DNMs occur is statistically higher amongst patient cases than in healthy controls (O'Roak, Vives, Girirajan, et al., 2012; Iossifov et al., 2014; Homsy et al., 2015). A study investigating DNMs in Schizophrenia patients identified a 3.5-fold increase of proportion of nonsense DNMs in sporadic trios compared to healthy trios (McCarthy et al., 2014). This same study concluded that the DNMs identified in these trios were significantly more likely to occur in genes with a higher probability of haploinsufficiency. Separately, in a cohort of 2270 ASD trios, a statistically significant 13.8 % carried a *de novo* LoF mutation (De Rubeis et al., 2014). In some cases of ASD, it has been noted that patients who present with a causative DNM typically show a more severe phenotype than those with inherited variation (O'Roak et al., 2014; Krumm et al., 2015). In summary, the results of all these studies into the role of DNMs in varying diseases

indicate a predisposition for damaging DNMs to selectively distribute within genes and regions in which they are not tolerated and cause potentially stronger phenotypic responses.

1.6.2 De novo mutations in male infertility

Despite a clear role for genetic causes in male infertility, there is a significant lack of diagnostically relevant genes with idiopathic male infertility accounting for around 40 % cases (Tüttelmann et al., 2018; Oud et al., 2019; Kasak & Laan, 2021; Riera-Escamilla et al., 2022). To date, the majority of disease gene discovery relating to MI has been focussed on AR inheritance models or X/Y linked male infertility (Oud et al., 2019; Houston et al., 2021). AD causes of MI can be either maternally inherited or occur *de novo* in the patient. Whilst both inheritance models are important to investigate in this disorder, the ability of DNMs to avoid evolutionary selection makes these mutations potentially more deleterious and therefore of great interest to investigate in MI (Acuna-Hidalgo et al., 2016).

In conditions such as ID and other developmental disorders, where there is a high degree of reproductive lethality similar to that seen in MI, the role of DNMs has been proven to be of high aetiological importance (Veltman & Brunner, 2012; Vissers et al., 2016). Additionally, *de novo* factors are already known to play a significant role in male infertility in the case of AZF deletions on the Y chromosome and Klinefelter syndrome. As MI inherently cannot be passed on from father to child via spontaneous pregnancy, YCMD can be expected to occur mostly *de novo* in the affected men. Together, 25 % of all cases of NOA are explained by AZF deletions and Klinefelter syndrome (Krausz & Riera-Escamilla, 2018; Kasak & Laan, 2021). In 1999 a study was performed in which a DNM was identified on the Y-chromosomal gene *USP9Y* in a man with azoospermia that was absent in his fertile brother (Sun et al., 1999). This gene is located within the AZFa region on the Y chromosome and is thought to act as deubiquitinase, playing a role in the development of male germ cells (Colaco & Modi, 2018). Since this first study in 1999, however, there have been conflicting reports as to *USP9Y* and its role as an essential gene in male fertility, with examples of normozoospermic males presenting with deletions within the gene (Luddi et al., 2009; Krausz et al., 2006). It has recently been shown that the gene *DDX3Y* is most likely the causative gene for NOA within the AZFa region of the Y chromosome (Dicke et al., 2023).

DNM discovery requires the sequencing of both parents as well as the affected patient in order differentiate between inherited heterozygous mutations and DNMs. In conditions such as early onset developmental delay, parents are typically available and eager to participate in such research. In the case of male infertility, one of the great limiting factors to *de novo* studies are the acquisition of parental samples. The age of infertile males at diagnosis, along with the

stigma that often comes with the disorder, means that parents are sometimes unavailable to provide a sample, or the patient does not want to disclose this information to family members. Currently there is a significant lack of large-scale systematic analyses into the role of DNMs in male infertility. In 2021 a pilot study was published in which 13 infertile males and their parents underwent WES in order to identify potential genes of interest containing pathogenic DNMs (Hodžić et al., 2021). The study identified five *de novo* mutations in genes which could potentially be associated with the azoospermic phenotype seen in the patients. Whilst this work provides a starting point for patient-parent trio analyses in cases of severe idiopathic male infertility, the authors acknowledge a lack of functional information connecting these genes to the NOA phenotype seen in the patients as well as a limitation on gene discovery due to the small sample size. With over 2,000 genes expressed during the process of spermatogenesis alone (Gao et al., 2022), it is highly likely that there are a large number of MI disease genes which are currently unidentified in patients. The use of NGS to detect DNMs in large-scale cohorts of infertile patients and their parents is an important line of research to address this.

1.7 Current Treatments for Male Infertility

For males who present with disorders of the HPG axis or other endocrine issues as the cause of their infertility, treatment will often involve hormonal therapy such as pulsatile gonadotropin-releasing hormone (GnRH) infusion to help restore balance to the HPG axis and allow normal spermatogenesis to occur (Mao et al., 2017). Varicoceles can often be treated with surgically with around 50 % of patients showing significant sperm count improvements after surgical correction (Pathak et al., 2016). For those, however, who suffer from more permanent and incurable forms of MI, assisted reproductive techniques (ART) can be used to bypass the cause of the infertility and allow conception to occur. Most commonly, in vitro fertilisation (IVF) or intracytoplasmic sperm injection (ICSI) are utilized (Steptoe & Edwards, 1978). IVF is the method in which oocytes are harvested and fertilized *in-vitro* before implantation by mixing the oocyte with spermatozoa directly from the ejaculate. ICSI is utilised in cases where sperm count, or quality, is severely diminished by injecting an individual spermatozoon directly into the cytoplasm of an oocyte (Aggarwal et al., 2021; Esteves et al., 2018). In cases such as NOA and severe oligozoospermia a procedure called testicular sperm extraction (TESE) or micro-TESE can be performed where the sperm count is so low that the ejaculate of the patient is unusable, and biopsies of the testis must be performed in order to isolate useable spermatozoa for ICSI (Achermann et al., 2021).

A study published in 2018 concluded that approximately 1.7 % of all births in the USA are conceived using IVF or other ART (Sunderam et al., 2018). This has appeared to be a huge leap in the treatment of male factor infertility with many couples now able to conceive but, in reality, only around 20 % - 30 % of all ART cycles are successful (Sunderam et al., 2018; Human Fertilisation and Embryology Authority (HFEA), 2021). Not only do these technologies not offer a solution for the large majority of patients undergoing treatment, but the potential to pass on problematic genetic mutations cannot be ignored. This is particularly important in cases of idiopathic male infertility with potential AD inheritance.

The potentially negative effects of utilising sperm from males suffering from NOA can be seen as aneuploidy screening on embryos obtained after Intracytoplasmic Sperm Injection (ICSI) in NOA men showed increased levels of aneuploidy and mosaicism than expected with the mean sex chromosome aneuploidy rate being 0.8 % in ICSI children as opposed to 0.2 % in the general population (Bonduelle et al., 1998). Karyotyping of miscarriages occurring after Testicular Sperm Extraction ICSI (TESE-ICSI) procedures also highlighted an increase in aneuploidy levels (Bettio et al., 2008). It is logical therefore to assume that if increases in known genetic defects are seen in children who are conceived in this manner, the potential for passing on idiopathic male infertility to offspring is high. As many of the first children who were conceived from NOA males using ICSI have only reached reproductive age within the last 5 - 10 years, very few studies have been performed on the overall fertility rates in ICSI conceived children. In 2016, the first comparison was performed between ICSI conceived children and those conceived via spontaneous pregnancy (Belva et al., 2016). The 54 young men conceived via ICSI were found to have significantly lower median sperm count and total motile sperm count than the spontaneously conceived males. It should be noted that, whilst the ICSI offspring were seen to have decreased sperm count, the median was still 31.9 million per ml of ejaculate which is far above the guidelines for quantitative sperm defects (Mild oligozoospermia = less than 15 million sperm per ml of ejaculate). As with many studies involving follow up investigations in ART conceived children, the number of patients assessed in this group is too small to provide any definite arguments regarding the use of ART and its effects on the children born from these procedures. An additional study from 2022 performed WES on 35 children conceived via IVF or TESE-ICSI and their parents to compare the number of DNMs found in these ART conceived children compared to 17 spontaneously conceived children and their parents who also underwent WES (Smits et al., 2022). It was found that there was no significant increase in the number of DNMs in the ART children compared to those who were naturally conceived. Whilst this appears to be a positive result, the size of this study is again is relatively

small to the number of children who are conceived via ART and further large-scale WES investigations are needed to truly determine the risk. A greater understanding of the genetic causes of MI, as focussed on within this thesis, will likely be invaluable for genetic counselling of infertile couples when assessing their treatment options.

1.8 Project Aims and Chapters Outlines.

1.8.1 Project Hypothesis and Aims

Despite recent advancements in next generation sequencing technologies and decades of research into human MI, a staggering number of patients still lack insight into the cause of their infertility. With around 40 % of all MI cases remaining idiopathic, there is an overwhelming need for more in-depth investigations into novel disease genes related to MI. In this thesis, I aim to progress research into idiopathic MI by utilising NGS techniques to sequence large cohorts of infertile males and their parents to identify potential causative *de novo* mutations (DNMs). My overall hypothesis is that, like in diseases such as developmental delay, DNMs play an important role in idiopathic MI.

Aims:

1. To use NGS techniques to identify potentially causative DNMs in patients suffering from idiopathic MI.
2. To investigate any genes containing DNMs for functional enrichment in male infertility associated pathways.
3. To perform functional studies to investigate the potential disease mechanisms of any novel male infertility disease genes of interest.

1.8.2 Chapter Outlines

In the Introduction of this thesis, I have aimed to describe MI in terms of both its biological background and currently known causes of the disease as well as bringing attention to the need for further investigation into the large number of idiopathic cases of male infertility worldwide.

In Chapter 2 of this thesis, I provide a detailed description of the Materials and Methods I have used throughout the entirety of the work described in this thesis.

In Chapter 3, I discuss my primary investigation into the role of DNMs in idiopathic cases of MI by designing a targeted NGS sequencing assay to investigate 54 potential genes of interest for DNMs in an infertile patient parent trio cohort.

The work in Chapter 4 is my next study, where 185 idiopathic males and their patients undergo whole exome sequencing (WES) to identify novel disease genes containing potentially causative DNMs. The work in this chapter has been published in the journal "Nature Communications" for which I am co-first author.

In Chapter 5, I utilise the results from Chapter 4 to investigate the potentially damaging effects of a number of mutations in the pre-mRNA splicing gene *RBM5* by attempting to determine any damage to the genes splicing abilities.

In Chapter 6, I discuss the results of all the work presented in the thesis in relation to each other and what this means for the field of MI as well as further improvements.

Chapter 2: Materials and Methods

2.1 Patient Sample Information

2.1.1 Newcastle and Nijmegen Patient and Parental Samples

As the focus of this PhD is to investigate *de novo* mutations (DNMs) in cases of severe male infertility (MI), all patients who had their DNA collected and sequenced suffered from severe idiopathic MI. This means that all the males suffered from non-obstructive quantitative sperm defects, either Azoospermia (total absence of any sperm within the ejaculate), Extreme Oligozoospermia (<0.1 million sperm per ml of ejaculate) or Severe Oligozoospermia (<0.1 - 5 million sperm cells per ml of ejaculate) (Krausz & A Riera-Escamilla, 2018; Lopes et al., 2013).

The samples included in this study were collected either in Newcastle, UK, or Nijmegen, The Netherlands. Patients, as well as their fertile parents, were enrolled at the Radboud University Medical Centre (Radboudumc) outpatient clinic between July 2007 and October 2017 (n = 175) and at the Newcastle upon Tyne Hospitals NHS Foundation Trust between January 2018 to January 2020 (n = 17). The study protocol was approved by the respective Ethics Committees / Institutional Review Boards (Nijmegen: NL50495.091.14 version 5.0, Newcastle: REC ref. 18/NE/0089) and written informed consent from all participants was obtained prior to enrolment in the study. For the patients, residual genomic DNA was extracted from a blood sample taken at the time of evaluation and treatment at the respective fertility centre. DNA from all parents was obtained from saliva by using the Oragene OG-500 kit (DNA Genotek, Ottawa, Canada). The reference values and semen nomenclature were used according to the WHO guidelines (World Health Organization, 2010) with all patients providing two consecutive semen samples indicative of either azoospermia or oligozoospermia. A semen analysis performed at the respective clinics diagnosed azoospermia in 114 of 192 men (59 %) and 78 men with oligozoospermia (40 %). The oligozoospermic men were then classified as extreme oligozoospermia in 41 cases (53 %) and severe oligozoospermia in 37 (47 %). Clinical evaluation did not lead to an etiologic diagnosis and all patients were negative for *AZF* deletions and chromosomal anomalies. Most men were of Dutch or British descent (n = 167 and n = 16 respectively). One male was from the United Arab Emirates, one male was from Saint Martin (Caribbean Islands) and the remaining seven men were of unknown descent.

In this thesis, patient-parent trios were investigated using two different sequencing methods in order to identify DNMs which could help to explain their currently idiopathic infertility. Of these 192 patient-parent trio samples, 75 were subjected to targeted sequencing using single molecule Molecular Inversion Probes (smMIPs, see Chapter 3) and 185 underwent Whole Exome Sequencing (WES, see Chapter 4). Between the two studies, there was an overlap of 68

patients. In addition, a single father and son duo was sequenced using the smMIP assay as the mother sample from the trio failed at a later stage of the smMIP protocol.

Alongside these patient-parent trios, an additional 145 patients were included in the WES study without availability of parental samples, following the above procedures and protocols. A total of 69 of these 'singleton' patients were sequenced using smMIPs, 13 from Newcastle and 56 from Nijmegen. These samples were not analysed in this thesis but were prepared alongside the other 227 samples for sequencing using the smMIP assay for future investigation.

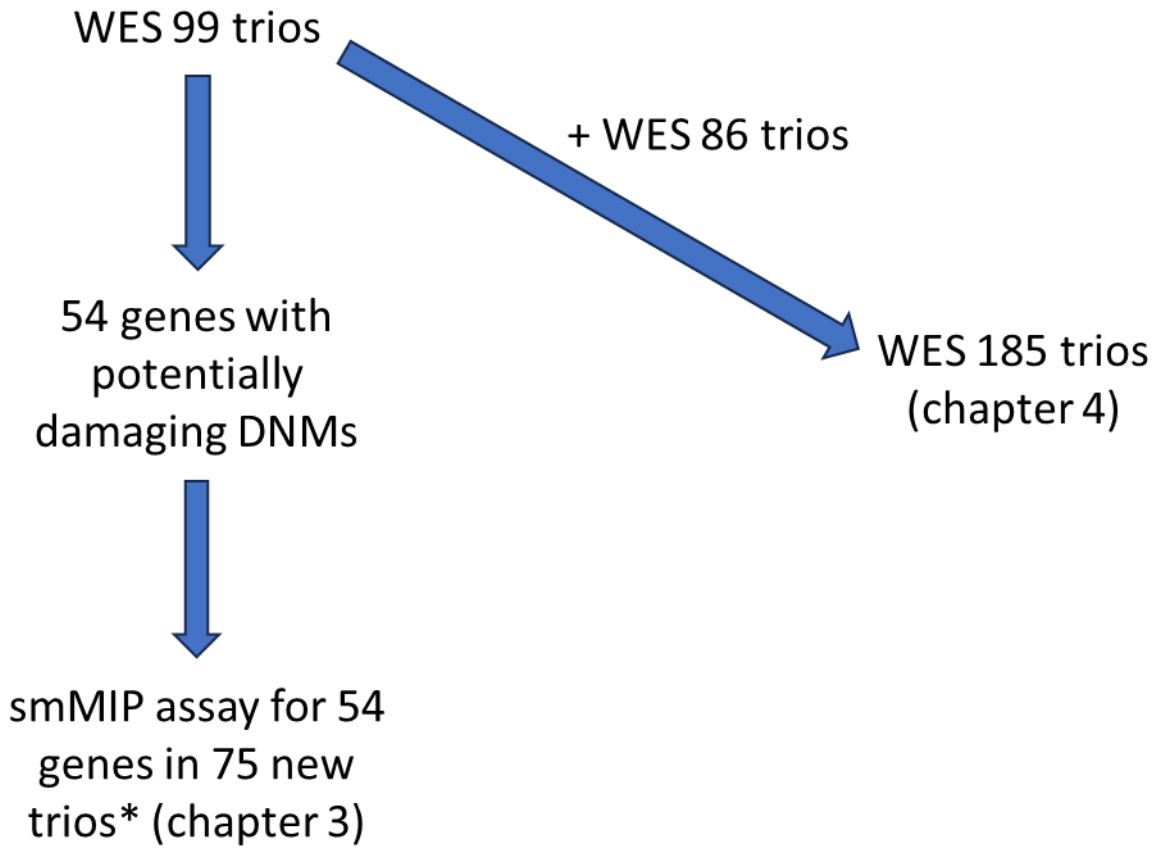


Figure 2.1: A chart to show the breakdown of all infertile patient-parent trios sequenced using various techniques throughout this thesis. In total 185 patients and their parents underwent whole exome sequencing using either an Illumina's Nextera DNA Exome Capture kit (n = 99) or the Twist Bioscience's Human Comprehensive Exome Kit (n = 86). The initial 99 trios were first analysed by Dr Manon Oud at Radboud University, Netherlands before being re-analysed and prioritised by myself for the work in Chapter 4 of this thesis. From Dr Oud's initial investigation, 54 genes of potential interest were highlighted and smMIP panel was created to identify any further mutations in these genes in a cohort of 75 patient parent trios, this is discussed in Chapter 3 of this thesis. Both the initial 99 and secondary 86 trios were analysed by myself in Chapter 4 of this thesis.
*Of these 75 trios analysed using the targeted smMIP panel, 68 were then included in the secondary cohort of 86 infertile patient-parent trios who underwent whole exome sequencing.

2.1.2 Fertile Control Cohorts

An anonymised exome dataset derived from 5,784 Dutch men and 5,803 Dutch women who had conceived at least one child was used as a control cohort for the frequency of rare variants in fertile men and fertile women. These men and women received routine exome sequencing at the Radboud University diagnostics centre as the healthy parent of a child with a severe developmental disorder. Although these men fathered a child with such a disorder, their fertility was expected to be comparable to that of an unselected sample from the human population. These samples were all sequenced either using Illumina's Nextera DNA Exome Capture kit which is consistent with the processing of infertile patient samples sequencing in both Newcastle and Nijmegen. The pipeline used to process this sequencing data was the foundation on which the processing pipeline in section 2.3.2 was created making the data highly compatible. All of the sequencing data from these control samples was re-annotated using our own custom annotation system as mentioned in section 2.3.2 to bring the control data set in line with all the NCL and NIJ samples. All variant information from these control samples was then filtered and assessed using the same methods as described in sections 2.3.3 - 2.3.4.

2.1.3 Additional Patient Cohorts

Exome data from four additional cohorts of infertile men were investigated in Chapter 4. The four external cohorts consisted of an Italian cohort (n = 48), a GEMINI cohort (n = 1,011), a MERGE cohort (n = 1,114) and a Geisinger-Regeneron DiscovEHR cohort (n = 88).

For the Italian cohort of 48 patients with non-obstructive azoospermia (NOA), exome sequencing was carried out as a service by Macrogen Inc. (Republic of Korea) utilizing the SureSelect_V6 enrichment (Agilent) and a NovaSeq 6000 (Illumina).

The Genetics of Male Infertility INitiative (GEMINI) is a multicentre study funded by the United States NIH. The GEMINI project performed WES on 1,011 unrelated men diagnosed with spermatogenic failure, the vast majority with unexplained NOA. Sequencing of genomic DNA was performed at the McDonnell Genome Institute of Washington University in St. Louis, MO, USA, using an in-house exome targeting reagent capturing 39.1 Mb of exome and 2 × 150 bp paired-end sequencing on Illumina HiSeq 4000.

The German Male Reproductive Genomics (MERGE) study comprised exome data of 1,114 men with azoo-, crypto-, or severe oligozoospermia. Known causes for male infertility such as chromosomal aberrations and microdeletions of the AZF region were excluded in advance. WES was performed as previously described (Wyrwoll et al., 2020).

The 88 patients diagnosed with MI participating in the Geisinger-Regeneron DiscovEHR collaboration were selected from deidentified EHR information using the ICD-10CM code N46 which refers to "Male Infertility" including oligospermia, azoospermia, other male infertility, and male infertility unspecified. All patients were sequenced at the Regeneron Genetics Center (RGC) as previously described (Dewey et al., 2016). Captured libraries were sequenced on the Illumina HiSeq 2500 platform with v4 chemistry using paired-end 75 bp reads. Exome sequencing was performed such that >85 % of the bases were covered at 20× or greater. Raw sequence reads were mapped and aligned to the GRCh38/hg38 human genome reference assembly using BWA-mem (Li & Durbin, 2010), single nucleotide and indel variants were called using GATK's HaplotypeCaller (Auwera et al., 2013).

For comparison purposes, all data retrieved from these distinct cohorts was converted to use the GRCh37 reference genome using LiftOver where necessary and then all variants were re-annotated using Ensembl's VEP v99 (McLaren et al., 2016) to minimize differences between datasets.

2.2 Targeted Sequencing Work using smMIPs

The initial experiment described in this thesis is based off work performed by a previous PhD student associated with our group, Dr Manon Oud. She had performed WES in 99 trios to identify any potentially pathogenic and disease causing DNMs in these infertile males. From this study, 54 genes were highlighted containing variants of interest. The experiment within this thesis involved a targeted sequencing method using single-molecule molecular inversion probes (smMIPs) to sequence these 54 genes in 75 trios and one father and son duo. The aim was to identify any additional potentially damaging DNMs in these genes of interest within these patients.

The method of using smMIPs has been well documented as a cost effective and highly efficient technique where specific genes can be targeted and sequenced in multiple patients in a single run and has been used by several groups to selectively screen genes for potentially damaging mutations (Pogoda et al., 2019; Gallon et al., 2020; Hiatt et al., 2013; O'Roak, Vives, Fu, et al., 2012). smMIPs consist of a backbone containing a unique molecular barcode for differentiation of smMIPs during multiplex sequencing, a ligation arm, and an extension arm (Figure 2.2A). These two arms are designed to target specific regions within the gene of interest before then capturing the target sequence (Figure 2.2B).

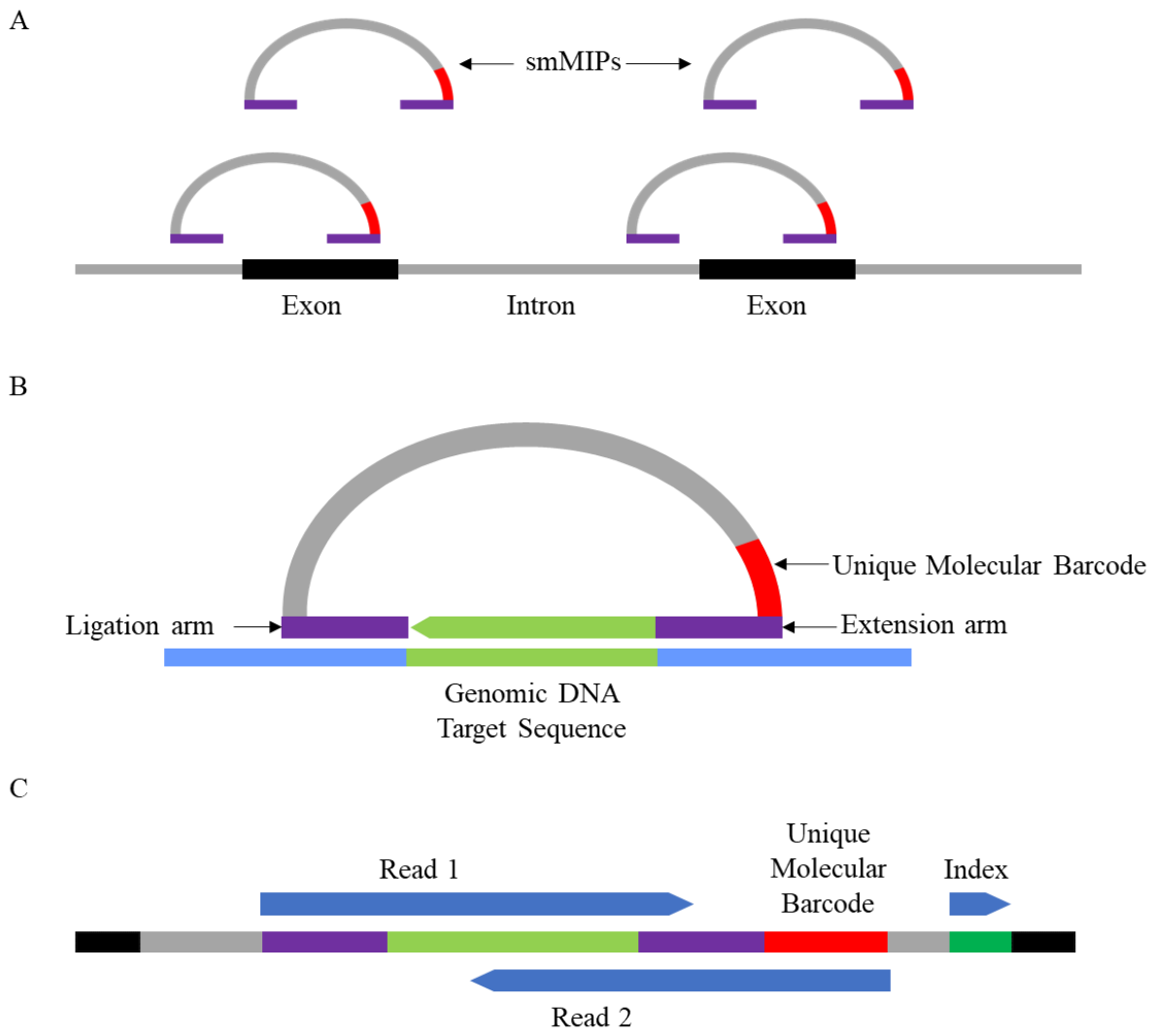


Figure 2.2: Single molecule Molecular Inversion Probe (smMIP) protocol from initial design to final sequencing. *A)* smMIPs are designed to cover specific regions of DNA using the program MIPgen, this aligns smMIPs with a target region of 109 bp to the target sequence, allowing for a minimum overlap of 10 bp and a maximum of 40 bp to ensure complete coverage of all exons. *B)* All smMIPs were designed with a ligation and extension arm totalling 40 bp in length which hybridise to the regions flanking the specific target to be amplified and an 8 bp unique molecular barcode which allows the differentiation between the reads assigned to each individual smMIP after sequencing. During the capture reaction, the smMIPs hybridise to the target region and the polymerase begins a 'gap-fill' reaction to copy the target sequence starting from the extension arm and forming a circularised smMIP after filling to the phosphorylated ligation arm. *C)* After exonuclease clean up, the smMIP is left as a linear single strand, through the PCR stage the index primer is added, which allows de-multiplexing of samples from the sequencing run, and then the sample can be sequenced. During the sequencing run, the smMIP is sequenced via Read 1, Read 2 and then the index is read separately.

2.2.1 smMIP Design

A total of 2,904 smMIPs were designed to fully cover and sequence 54 candidate genes for MI. All the work involving smMIPs and their design used the GRCh37/hg19 genome build. The University of California, Santa Cruz (UCSC) genome browser tool (https://genome.ucsc.edu/) was used to produce a Browser Extensible File (BED file) containing the start and end genomic

coordinates for every exon in the canonical transcripts of all 54 genes of interest (Table 2.1). The files also contained information regarding whether the gene lies on the positive or negative strand. The BED file was then edited manually to add an additional 20 bp buffer onto both the start and end coordinates to ensure full coverage when smMIPs are designed. All exons were then named within the BED file in the following format: [*gene*]_cds[*exon number*].

The previously produced BED file was run through a script which made use of the MIPgen program (Boyle et al., 2014), 1,000 genome index database (Hg19) and a list of all common Single Nucleotide Polymorphism (SNP) locations within these regions from the NCBI Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 1999). The list of common SNPs was essential to account for any common SNPs that could be present at the targeting arm position of a given smMIP. In these cases, two alternate smMIPs were created at this position, the second accounting for the potential variation. The script then output a list of 2,904 smMIPs with information on their targeting arm sequences, region of interest sequences, arm sizes and coordinates, and their own individual name based on the gene and the position of the smMIP within that gene. The format of smMIP names can be seen here: INF1_smMIP_[*gene name*]_[*given number*], with the given number allowing a record to be of the smMIP and its position.

Each smMIP has its own unique molecular barcode (UMB) to allow smMIP differentiation between reads for each individual smMIP after sequencing, and the size chosen was 8 bp which gives 65,536 potential base combinations. The region of interest for the smMIPs was set at 109 bp and the ligation and extension targeting arms had a total of 40 bp in length. The minimum overlap of adjacent smMIPs was set to 10 bp and the maximum 40 bp. These parameters for the smMIP creation were chosen to give the best coverage of the targets possible. Finally, the logistic optimal score, which determines at what point to stop optimising and accept a smMIP, was set to a minimum of 0.7 as recommended by the MIPgen program specifications (Figure 2.2A).

The smMIPs were ordered from Integrated DNA Technologies (IDT) as 25 nmol DNA Oligos in 96 well plates which had been normalized to 100 µM in IDTE Buffer pH 8.0. When not in use, the 96 well plates containing the oligos were stored at -80 °C.

Genes Investigated Using smMIPs		
ABCF3	*ERG*	*NUP210*
ABLIM1	*EVC*	*ODF1*
APC2	*EXOSC9*	*OSBPL3*
APOBEC3G	*FBXO5*	*PLCL1*
ATP1A1	*FNDC8*	*RASEF*
ATP8A1	*FUS*	*RPA1*
ATP8B4	*FZD3*	*SIX2*
C9orf50	*HELZ2*	*SMC2*
CAPN10	*HNRNPL*	*SOGA1*
CD81	*HR*	*STX7*
CDC5L	*ILVBL*	*STXBP2*
CDK5RAP2	*IQSEC1*	*TACC2*
CRACR2A	*LEO1*	*TLN2*
CWC27	*LZTS2*	*TMEM62*
DGKZ	*MCM6*	*TOPAZ1*
DNAJC2	*MPRIP*	*U2AF2*
DNMT1	*MYRIP*	*WDR17*
EMILIN1	*NEO1*	*ZBTB39*

Table 2.1: All 54 genes for which smMIPs were designed.

2.2.2 smMIP Pooling and Preparation

All 2,904 smMIPs were pooled together in a single library for sequencing. Due to the large number of smMIPs, they were initially pooled by gene, making 54 individual smMIP gene pools. Into each gene pool, 5 µl of the original smMIP was pooled giving a varied final volume depending on the number of smMIPs designed per gene. To achieve an equal concentration of each smMIP when pooling together from the individual gene pools, a 0.1x aliquot of each gene was added to the final smMIP pool, resulting in an equimolarly balanced pool at a volume of 290.4 µl.

In order for the smMIP region of interest to be properly captured, they needed to first undergo phosphorylation, which allows binding of the polymerase to the extension arm during the capture reaction. This was performed with 29 µl of the smMIPs, 1.2 µl of T4 Polynucleotide Kinase (PNK) (New England Biolabs), 23.8 µl of H_2O and 6 µl of 10X T4 DNA ligase buffer with 10 mM ATP (New England Biolabs) for a final reaction volume of 60 µl. This reaction ran in the thermocycler at 37 °C for 45 minutes and then followed by a kinase inactivation at 65 °C for 20 minutes. The final concentration of smMIPs in this reaction mixture was 0.000167 µM.

2.2.3 Capture Reaction

For the capture reaction, a similar approach was used to those previously described in literature, with some minor modifications (Oud et al., 2017; Hiatt et al., 2013). A molecular ratio of 1:800 between genomic DNA and smMIPs was used during the capture reaction. Using the value of 330 haploid genome molecules per 1 ng of DNA, it was determined that 26,400,000 smMIP molecules were required to capture the 100 ng DNA fully. Using Avogadro's constant, it was possible to determine the picomoles required of the 1x smMIP pool in the capture reaction at 4.38×10^{-5} pM. The formula used to calculate this was:

$$\left(\frac{smMIP\ molecules\ needed\ per\ 100\ ng\ DNA}{6.02214076 \times 10^{2}}\right) \times (1 \times 10^{12})$$

From this it was determined that 0.000262854 µl of the 1x smMIP pool would need to be added to the capture reaction to give the final concentration of 4.38×10^{-5} pM. This was determined using the following formula:

$$\frac{required\ picomoles\ in\ capture\ reaction}{concentration\ of\ smMIPs\ in\ 1x\ pool\ in\ picomoles\ per\ microliter}$$

With this being such a minute volume, two serial dilutions of 1 in 100, with a final volume of 200 µl in each, were performed starting with the initial smMIP pool to give a final concentration of 1.668×10^{-5} µM, or 1.668×10^{-5} pM/µl. The required volume of the diluted smMIP pool required in the capture reaction was 2.63 µl.

Per individual sample, the capture reaction consisted of 2.63 µl of the diluted smMIP pool, 2.5 µl of 10x Ampligase buffer (Epicentre), 0.08 µl dNTP 0.1 mM (Promega), 0.32 µl of Hemo Klentaq 10 U/µl (New England Biolabs), 0.2 µl of Ampligase 5 U/µl (Epicentre), 9.27 µl of H_2O and 10 µl of sample DNA at a concentration of 10 ng/µl. The capture protocol was then performed with an initial step of 95 °C for 10 minutes and then 65 °C for 20 - 22 hours to allow for smMIP-DNA hybridisation and gap filling to occur (Figure 2.2B).

The captured samples then underwent an exonuclease clean up reaction to digest any residual linear DNA present. To each sample, 0.5 µl of Exonuclease I (New England Biolabs), 0.5 µl of Exonuclease III (New England Biolabs), 0.2 µl pf Ampligase 10X Reaction Buffer (Epicentre) and 0.8 µl of H_2O was added. The exonuclease reaction contained an incubation at 37 °C for 45 minutes then 95 °C for 2 minutes to deactivate the exonuclease.

2.2.4 Addition of Specific Primers

Post clean-up, all samples were processed through a PCR reaction, involving the addition of a unique reverse primer to each individual sample. This 'index' primer allows the identification and differentiation of reads assigned to each sample during the bioinformatic processing of the next generation sequencing data. The index primers were used in previous smMIP studies (Hiatt et al., 2013; O'Roak, Vives, Fu, et al., 2012) and ordered for use from both Metabion and Integrated DNA Technologies (IDT). Each reaction tube contained 12.5 µl of 2X Iproof (BioRad), 0.42 µl of 10 µM Universal forward primer (Illumina), 5 µl of exonuclease-treated smMIP-captured sample, 5.83 µl of H_2O and then 1.25 µl of a unique reverse index primer with a total reaction volume of 25 µl. The PCR protocol used was as follows:

Temperature	Time	Cycles
98 °C	30 Seconds	x1
98 °C	10 Seconds	
60 °C	30 Seconds	x18
72 °C	30 Seconds	
72 °C	2 Minutes	x1
4 °C	Indefinitely	x1

Table 2.2: PCR protocol used for the addition of specific primers to the smMIP captured samples.

Gel electrophoresis was used to confirm that both the capture of the target sequence and the PCR reaction itself had been successful. The 1.5 % agarose gel was made by first dissolving 2.25 g of UltraPure Agarose (Invitrogen) in 150 ml of 1x TBE buffer (ThermoFisher). This was then microwaved for approximately 3 - 5 minutes to obtain a homogenised solution. Once it had cooled slightly, 9 µl (6 % of total volume) of GelRed (Biotium) was added and mixed in, after which, the gel mixture was then poured into the electrophoresis tray and left until set. The GelRed was added to allow visualisation of the DNA bands after running the gel, this was chosen as opposed to ethidium bromide due to the increased sensitivity and safety in comparison. Once set, the gel was placed into the Bio Rad Sub-Cell GT Cell tank which contained 1X TBE buffer (ThermoFisher). Samples were then individually pipetted into to each well in the gel at a total volume of 6 µl, with 3 µl of sample and 3 µl of 6X TriTrack DNA Loading Dye (Thermo Scientific) to allow visualisation of the samples whilst running. A ladder

was used in the first (and occasionally last) wells of each row of samples in order to be able to compare the size of product. In this case the ladder used was the BenchTop 100 bp DNA Ladder (Promega) diluted with 6X TriTrack DNA Loading Dye (Thermo Scientific) dye. The gels were then run at 80 volts for 60 minutes. All gels were then visualised under UV light using the GelDoc-It (UVP) system to confirm the presence of the correct DNA band sizes.

2.2.5 Sample Pooling and Clean Up

After confirming the correct product was amplified, the concentrations of all the samples were measured using the Qubit 3 Fluorometer (Invitrogen) using the Qubit dsDNA HS assay kit. A working solution was created containing 1 µl of Qubit dsDNA HS Reagent and 199 µl of Qubit dsDNA HS Buffer for each individual sample to be checked, and for the two standards needed to calibrate the Qubit machine. From this working solution, 190 µl was added to two individual tubes, one of these containing 10 µl of Qubit dsDNA HS Standard #1, the other containing 10 µl of Qubit dsDNA HS Standard #2. For every DNA sample, 199 µl of the working solution was added to an individual tube and 1 µl of that sample was then added. The tubes containing the two standards and the samples were then all vigorously vortexed for 3 - 5 seconds before resting at room temperature for two minutes. Standards #1 and #2 were then read to calibrate the fluorometer before reading each sample. All samples were measured in triplicate to ensure accuracy, and an average of the concentrations was used as the final value. All samples were pooled equimolarly in a single pool, allowing for an even allocation of reads during sequencing to each individual sample. For the two initial test runs, the samples were all pooled directly into one tube however due to the large number of samples in the processed in the final NovaSeq (Illumina) run, samples were initially pooled with others of a similar concentration before then creating the one final pool.

The pool was screened for non-specific PCR products using the Agilent High Sensitivity D1000 ScreenTape System on the Agilent 4200 TapeStation according to the manufacturer's standard protocol (Agilent). Any impurities were removed using the Agencourt AMPure XP bead PCR clean-up kit (Beckman Coulter), following the manufacturers protocol. Post clean-up, a repeat screening was performed on the TapeStation to confirm that the process had been successful. The concentration of the final pool was measured using the previously mentioned Qubit protocol, ready to then undergo library preparation.

2.2.6 Library Preparation and Sequencing Setup

The final sample pool was diluted and prepared for sequencing using the manufacturer's protocols for Illumina's MiSeq and NovaSeq platforms. The initial two sequencing runs of four samples each, were performed on the Illumina MiSeq platform. The final pool containing the indexed, smMIP captured samples, was diluted down initially to 4 nM as required to proceed with the MiSeq Denature and Dilute library preparation (Illumina). This was prepared using the following formula:

$$\frac{concentration\ of\ sample\ pool\ (ng/\mu l)}{660\ g/mol\ \times library\ size\ in\ bp} \times 10^6$$

The final sequencing run of all 300 patient and parental samples was performed using the Illumina NovaSeq platform using an Illumina S2 Flow Cell, using paired end reads, with an output of 2 x 150 bp and providing an output of 1000 – 1250 Gb of data. The library preparation was performed by staff within the Genomics Core Facility (GCF) at Newcastle University.

Both the MiSeq v2 reagent kit (Illumina) and the NovaSeq 6000 S2 Reagent Kit v1.5 (Illumina) allowed for 325 cycles of sequencing to be performed for each run, 158 of these cycles were allocated to Read 1 sequencing and 158 to Read 2 sequencing to allow for paired end reads (Figure 2.2C). An additional eight cycles were allocated to the sequencing of the index sequence which allows the differentiation of reads assigned to individual samples. The MiSeq platform allowed for these parameters, as well as the sample names and index primer sequences to be input to the Illumina BaseSpace program which could be used to track the runs as they progressed as well as providing data quality metrics. The parameters were uploaded via a .csv file to the NovaSeq platform.

2.2.7 Re-pooling with 1x and 10x smMIP Pool

Re-pooling of the unphosphorylated smMIPS was performed after the first MiSeq sequencing run to assess whether any errors had occurred during the initial pooling. A new 1x unphosphorylated pool was made from the initial gene smMIP pools previously created in section 2.2.2 and the same protocol was followed as with the previous 1X pool. This new pool was then processed and captured using two of the same samples 'Arg1' and 'BA' that had been used in the initial run for a fair comparison.

Information on the efficiency of smMIPs provided by the first MiSeq run was then used to rebalance the smMIP pool and improve the performance of the assay. The efficiency of each

individual smMIP within each individual sample was able to be determined using the UMB sequence to assign reads to separate smMIPs and the index sequence to allocate them into samples.

The read fraction (RF) of all 2,904 smMIPs was determined from this initial run using the number of reads assigned to that smMIP as a fraction of the total number of reads produced during the sequencing run. The ideal RF for each smMIP was defined to be 0.0003, determined by the total number of smMIPs and their fraction of one individual read when working at equal efficiency. Using the assigned RF values, the smMIPs were split into four different groups: Overperformers with an RF >0.003, Within-Range with an RF of >0.00003 and <0.003, Underperformers with an RF of <0.00003 and Zero with smMIPs showing no assigned reads. All smMIPs that underperformed or were allocated zero reads across multiple samples were then pooled at 5 ul per smMIP, into a new 10x Pool which was then processed using the same methods starting from sections 2.2.2 to 2.2.5.

2.2.8 Bioinformatic Processing of Sequencing Data

All preliminary quality control checks of the sequencing data produced by the MiSeq runs were retrieved from BaseSpace (Illumina). These initial quality checks provided information such as percentages of clusters passing filter, percentages of clusters with a Q score of 30 or above, the number of reads achieved from the sequencing and the percentage of these reads assigned to each sample. Optimally, 100 % of all clusters would have a Q score of 30 as this is equivalent to the probability of an incorrect base being called once in 1000 times, providing a base calling accuracy of 99.9 %.

After quality control checks, FastQ files produced by the sequencing run were fed into a smMIP python pipeline custom made by Dr Simon Cockell at the Bioinformatic Support Unit at Newcastle University. The first step in processing the FastQ files involved retrieving the UMBs to ensure only unique reads continue processing and are aligned. Any reads without valid UMBs or lacking the matching smMIP arm sequences were discarded before mapping. The pipeline then aligns sequence reads to the UCSC reference genome hg19 with Burrows-Wheeler Aligner (bwa-mem 0.7.17). SAMtool version 1.3.1, was used to convert alignment files into BAM files by sorting and indexing them. The Picard toolkit v2.13.2 was used to mark duplicates and SAMtools was used to de-duplicate and index the resulting modified BAM files. Genome Analysis ToolKit v3.8 (GATK3) (Auwera et al., 2013), using IndelRealigner, BaseRecalibrator and Haplotypecaller, was used to perform indel realignment, base-quality score recalibration

and variant calling. The generated variant call files (VCF) were then annotated using the Ensemble's Variant Effect Predictor (VEP)v92 (McLaren et al., 2016).

2.2.9 Filtration of Identified DNMs

The information contained within annotated VCF files provided detailed data on all variants detected per patient including the wild type (WT) and detected variant alleles, amino acid changes, consequences of the variant, allele frequency among different populations and multiple different pathogenicity scores predicting how potentially damaging any given mutation would be to the structure and function of the protein or gene product.

Each proband sample file underwent manual filtering to remove any variants with a gnomAD database allele frequency of >0.001 or 0.1 % to ensure only rare variants were highlighted. All variants which were non-synonymous and produced either missense, frameshift, stop-gained or start-lost consequences within exonic regions mutations were kept. It was after this step that the proband filtered variants were then cross referenced to the parental VCF files to rule out any inherited variants, keeping only those with *de novo* inheritance.

As a further confirmation step, all remaining variants were visualised in the initial BAM files produced using the Integrated Genome Viewer (IGV).The variants which appeared to be true DNMs or were classified as unclear underwent Sanger sequencing for a final confirmation.

2.2.10 Validation of Identified DNMs using Sanger Sequencing

Validation of DNMs was performed using standard Sanger Sequencing approach on an Applied Biosystems SeqStudio 3200 Genetic Analyzer (ThermoFisher) to confirm the presence of the mutation in probands and its absence in the parents.

Primers were designed for each variant using the program Primer3 (Untergasser et al., 2012). The design specifications included an optimum melting temperature of 60 °C with a range of 2 °C either side, an optimum length of between 18 bp and 25 bp and a buffer region around the variant of interest of a size of 100 bp in both directions, to allow for any sequencing errors at the beginning and end of the Sanger output. Primers were checked for self-complementarity using the program Oligocalc (Kibbe, 2007) and for region specificity using the UCSC In Silico PCR tool (Kent et al., 2002). All Oligos were then ordered from Integrated DNA Technologies (IDT).

All variants were amplified using a similar PCR protocol to that in Table 2.1 with a 61.5 °C annealing temperature and 30 cycles as opposed to the previously mentioned 18. Each reaction tube contained 10 µl of 2X Iproof (BioRad), 0.3 µl of the custom forward primer, 0.3 µl of the custom reverse primer, 2 µl of patient DNA and 7.4 µl of H_2O with a total reaction volume of 20 µl.

Once the presence of the correct size band was seen on an agarose gel, the variant was then taken forward for clean-up by method of EXO-SAP protocol. This reaction involved the addition of 0.025 µl of Exonuclease I, 0.25 µl of Shrimp Alkaline Phosphatase (SAP) (New England Biolabs) and 9.725 µl of H_2O to the 20 µl post-PCR reaction mix. The reaction then ran at 37 °C for 30 minutes before inactivation of the reaction at 95 °C for 5 minutes.

To prepare the amplified target regions for Sanger sequencing, the BigDye Terminator v3.1 cycle sequencing kit (Applied Biosystems) was used. During this reaction, 0.5 µl Of BigDye Terminator 3.1, 4 µl of BigDye Terminator 5X sequencing buffer, 12 µl of H_2O, and 0.5 µl of either the forward or reverse primer was added to 3 µl of the EXO-SAP treated PCR mix. The cycle times for this reaction can be seen in Table 2.3. All BigDye reactions were performed in 96 well plates.

Temperature	Time	Cycles
96 °C	1 Minute	x1
96 °C	15 Seconds	
50 °C	15 Seconds	x25
60 °C	2 Minutes	
4 °C	Indefinitely	x1

Table 2.3: BigDye Terminator v3.1 protocol.

To purify the samples, Ethanol Precipitation was carried out immediately after completion of the BigDye terminator protocol. 2 µl of 125 mM EDTA (Invitrogen), 2 µl of 3M Sodium Acetate (Invitrogen) and 70 µl of 100 % Ethanol (Sigma Aldrich) were added to each well. The plates were then sealed, mixed by inversion, and incubated at room temperature for 15 minutes before spinning at 2000g for 30 minutes. After spinning, the plate lid was removed and inverted onto tissue paper and gently tapped, before being placed in the centrifuge (still inverted on tissue paper) and spun up to 100g to remove the waste ethanol. 70 µl of 70 % ethanol was then added

to each well and the plate was spun at 1650g for 15 minutes. The waste ethanol was then removed again as previously described and the plate was left to air dry in the dark for 10 minutes. After this stage, the pellets were then resuspended in 10 µl of Hi-Di Formamide (Applied Biosystems) before running on the Applied Biosystems SeqStudio 3200 Genetic Analyzer (ThermoFisher). ThermoFisher's own variant analysis tool was then used to determine the inheritance model to which each variant adhered.

2.3 WES Protocols

2.3.1 WES Enrichment Kit Information

Samples were initially prepared and enriched by the GCF at Newcastle University following two different exome enrichment kits. The first 99 trios were enriched using Illumina's Nextera DNA Exome Capture kit, whereas the second set of 86 trios were enriched using the Twist Bioscience's Human Comprehensive Exome Kit. This was done after an in-house comparison of both kits showed the Nextera kit was not very specific in also targeting a lot of non-coding regions, resulting in a lower average coverage in the coding region. These results will be further discussed in Chapter 4 of this thesis. All sequencing was performed on the NovaSeq 6000 Sequencing System (Illumina) on an S2 flow cell using pair-end reading (2 x 100bp). More than 99 % of all exonic regions were covered at least 1x using both enrichment kits, but the average coverage varied considerable for the coding region with 72× for the Illumina Nextera Kit and 99x for the Twist Bioscience's Kit.

2.3.2 WES Post-Sequencing Processing

Post sequencing processing of WES data was performed by the BSU and Dr Miguel Xavier. Sequenced reads were aligned to Human Reference Genome (GRCh37.p5) using BWA-Mem v0.7.17 (Li & Durbin, 2010), Picard (https://broadinstitute.github.io/picard/), and GATK v4.1.4.1(Auwera et al., 2013). The sex, ancestry and relatedness of each sample was calculated using Peddy (Pedersen & Quinlan, 2017), samples found to have the incorrect sex or were unrelated to the correct samples were excluded from this study. Following best practice recommendations, single nucleotide variations and small indels were identified and quality-filtered using GATK's HaplotypeCaller obtaining comparable results independently of the kit or the origin of the DNA. Afterwards, all variants were further analysed using a custom GATK4-based algorithm to identify and separate high- and low-confidence *de novo* variants from inherited variants. Briefly, posterior genotype probabilities (GQ) were recalculated for each

sample at each variant site using Bayes' rule to consider family and population priors (Auwera et al., 2013; Poplin et al., 2017). Variants were classified as low-confidence DNMs when they were absent in parental samples but detected in the proband with GQ ≥ 10 and either an allele count (AC) below four or allele frequency (AF) < 0.1 % in all samples, whichever is more stringent. Variants with GQs ≥ 20 with either ran AC < 4 or AF < 0.1 % were classified as high confidence DNMs. Ensembl's Variant Effect Predictor v99 (VEP) (McLaren et al., 2016) was used to fully annotate all *de novo* variants.

2.3.3 Variant QC Filtering

After sequencing of the samples, a large number of variants were present and many of these did not fit the parameters we would expect of a variant causing a rare disease. This is why a stringent set of filtering criteria were used to remove (1) variants which were too common in the general population, (2) variant positions that were poorly covered during sequencing or (3) variant calls that did not fit the *de novo* heterozygous frequency expected. The primary stages in filtering of variants included removing all variants with an allele frequency of >0.1 % in the gnomAD database to only include rare variants in our analysis. All variants then with <10 reads in the exome data and/or less than 15 % of these reads containing the mutation were then removed. Finally, any remaining variants located outside the exonic regions were removed. This provided the initial list of rare *de novo* variants.

2.3.4 Variant Pathogenicity Prediction and Prioritisation

A number of tools were used to assign a level of potential pathogenicity to all the remaining rare, protein altering DNMs in order to highlight any mutations which may have a detrimental effect on the function of the gene in which they lie in, thus being a mutation level investigation. Pathogenicity prediction was performed using the tools Sorting Intolerant From Tolerant (SIFT) (Vaser et al., 2016), MutationTaster (Schwarz et al., 2010), and PolyPhen2 (Adzhubei et al., 2010) and all variants were classified according to the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) 2015 guidelines (Richards et al., 2015). Both SIFT and Polyphen work by assessing the impact a mutation will have on the amino acid structure of protein and whether this is likely to alter the protein's function (Vaser et al., 2016; Adzhubei et al., 2010). Mutation Taster on the other hand works by performing a series of *in silico* tests to determine the pathogenicity of a given SNP by investigating amino acid substitution, protein domain loss, potential splicing effects,

conservation of DNA at the point of the mutation, protein truncation and disruption of regulatory elements (Schwarz et al., 2010). All protein altering variants predicted to be pathogenic by at least two out of three prediction models, absent from the fertile male cohort, present in <5 males in the gnomAD database were considered for further functional analysis. Maternally inherited mutations present in genes identified as having a protein altering DNM were also independently identified in all patients and submitted to the exact same method of filtration and interpretation as described above.

2.3.5 Interpretation of DNM-Containing Genes

To further assess the potential pathogenicity of any DNMs highlighted, the function of each gene affected by a DNM was investigated. This highlighted whether any of the DNMs identified were found in genes likely to be involved in spermatogenesis or similar processes and whether they were likely to cause disruption to the normal function of these genes in these processes. The functional analysis was split into six different categories; each category provided a score of either 1 or 0 depending on whether a DNM gene met the threshold for that category. These categories included: RNA expression of the gene in the testis, RNA enrichment in the testis or presence in spermatogenesis, protein expression in the testis, whether an infertile mouse model already exists for the given gene, the protein function in relation to spermatogenesis and finally whether the given gene interacts with any known fertility genes. For expression levels retrieved for each gene of interest from the GTEx database (https://www.gtexportal.org), an expression of medium (≥10 < 100 TPM) or high (>100 TPM) gave a score of 1 with low (>2 < 10 TPM) and no expression (≥10 < 100 TPM) or high (>100 TPM) gave a score of 1 with low (>2 < 10 TPM) and no expression (<2 TPM) giving a score of 0. RNA enrichment was based on elevated expression (tissue enriched, group enriched, or tissue enhanced) in the Human Protein Atlas (Uhlén et al., 2015) or being among the genes up or down regulated during spermatogenesis as found in a recent single cell RNA sequencing study (Guo et al., 2018). Protein expression was retrieved from the Human Protein Atlas (Uhlén et al., 2015)and interactions with known infertility genes was calculated using STRING version 11 (Szklarczyk et al., 2017). The final classification of the genes was then split into Not Causative, Unlikely Causative, Unclear and Possibly Causative. These classifications were given based on the variant scores out of 6 with: [0 points + not expressed/not detected/not present on several occasions = Unlikely Causative], [0 points + “Unknown” on several occasions = Unclear], [1–2 points = Unclear] and [3–6 points = Possibly Causative].

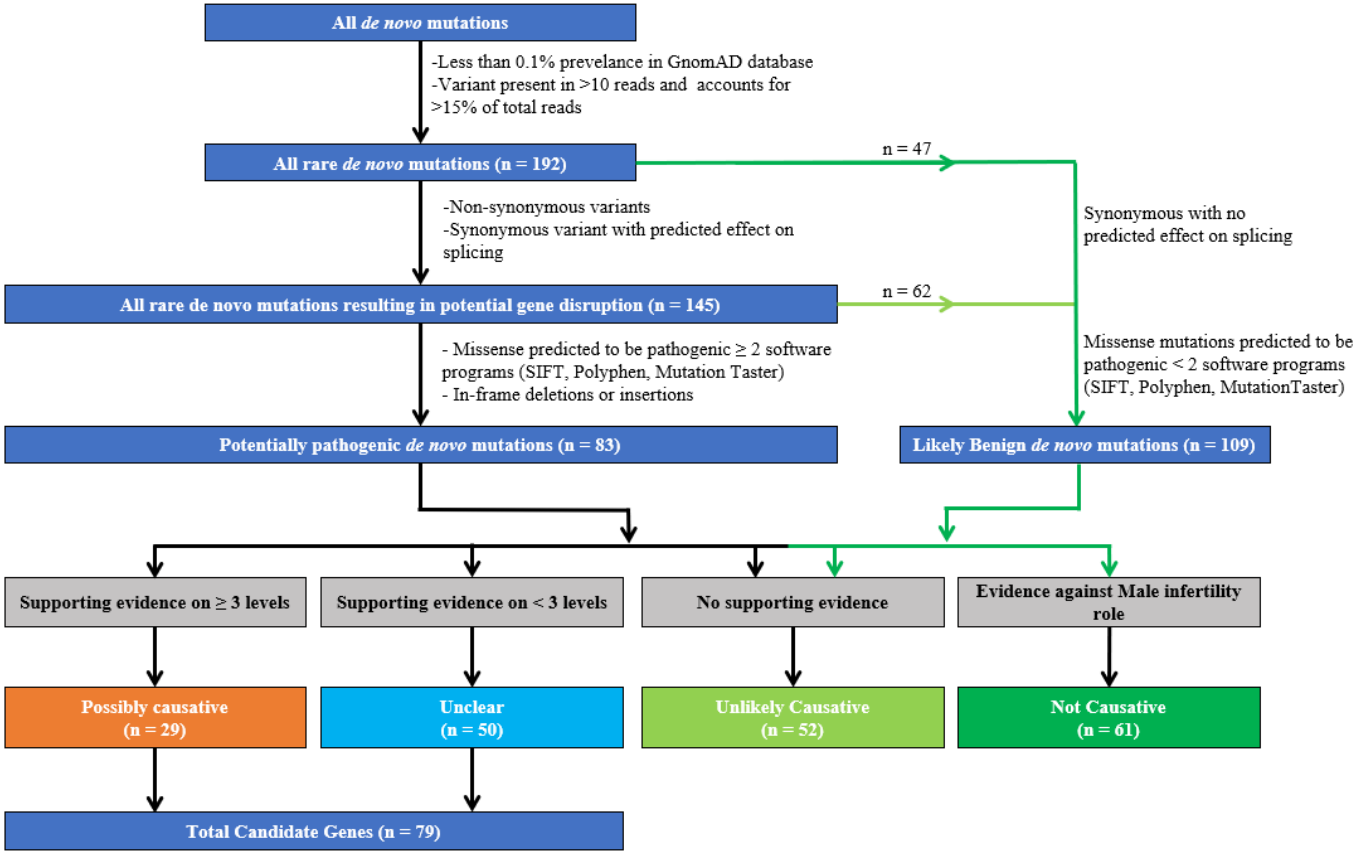


Figure 2.3: A flow chart representing the variant pathogenicity assessment pipeline used to prioritise and assess all DNMs highlighted after sequencing. All the filtration steps shown in this flow chart were performed by hand in excel on data provided by VEP (McLaren et al., 2016) annotated VCF files as a product of a custom pipeline as described in Section 2.3.2. All the information regarding used to filter "All *de novo* mutations" down to the "Potentially pathogenic *de novo* mutations" was present in these files. The supporting evidence used to categorise variants into either "Possibly Causative", "Unclear", "Unlikely Causative" and "Not Causative" was also performed by hand by assessing 6 points of information. These were; RNA expression of the gene in the testis, RNA enrichment in the testis or presence in spermatogenesis, protein expression in the testis, whether an infertile mouse model already exists for the given gene, the protein function in relation to spermatogenesis and finally whether the given gene interacts with any known fertility genes. The databases used to gather this information are discussed and referenced in section 2.3.5.

2.3.6 Copy Number Variation (CNV) Analysis

CNV calling was performed on the trio-based exome data with a custom GATK4-based pipeline, CNVRobot (https://github.com/AnetaMikulasova/CNVRobot). This workflow exploited the GATK4 sequence read depth normalization (McKenna et al., 2010) and a custom R-based segmentation and visualization (https://www.R-project.org). Parental samples from the trios under examination were used as controls for the normalization step. The CNVs detected were annotated using AnnotSV (https://lbgi.fr/AnnotSV) (Geoffroy et al., 2018). CNVs present in more than 1 % of the samples of the Database of Genomic Variants present in more than 10 % of the patients were excluded from the analysis. The remaining rare deletions and duplications were individually inspected through the genomic profiles and detailed Log2Ratio plots generated by the workflow. Only CNVs involving more than two exons were further considered to minimize the inclusion of false positives, and two CNVs present in the probands but absent in their parents were selected for further validation. The CNV calling and processing was performed by Dr Kumara Mastrorosa, a previous PhD student at Newcastle University, however the genes affected by these *de novo* CNVs were independently investigated by me and are discussed within this thesis.

2.3.7 Mutational Intolerance Statistical Tests

To investigate whether there was an enrichment of *de novo* loss-of-function (LoF) mutations within genes that were intolerant to LoF mutations, the probability of LoF intolerance (pLI) score was used. The pLI score indicates how intolerant a gene is to a LOF mutation with a pLI score of 1 indicating a significant intolerance to LOF mutations with a likelihood for functional disruption to the gene if this mutation type occurs. This was based on data from the Genome Aggregation Database (gnomAD) (Lek et al., 2016) containing genetic data from 141,456 individuals. This score was not used for filtering but rather for highlighting those variants which were likely to be more damaging.

The likelihood was computed of the observed median pLI score of each gene (LoF in controls) set compared to the expected median pLI based on a previously described method (Lelieveld et al., 2016). The expected number of recurrently mutated genes was computed by redistributing the observed number of mutations at random over a determined set of genes based on their specific LoF and functional mutation rates. In contrast to previous studies (Lelieveld et al., 2016; Samocha et al., 2014) however, a set of 2,766 coding DNMs in 1,941 control individuals

was used instead of using the complete set of 18,226 pLI annotated genes to obtain expected median pLI scores. This control data set was downloaded from the *de novo*-db version 1.6.1 (http://de novo-db.gs.washington.edu/de novo-db) and was used to correct for the gene-specific mutation rate. The empirical P value was calculated by comparing the observed median pLI to the expected pLI following 100,000 random sampling simulations. Case and fertile controls were processed using the exact same filtration and annotation parameters as described above so that each variant detected was evaluated in a comparable manner. The same method was then repeated using the Loss-of-function observed/expected upper bound fraction, or LOEUF score (Oved et al., 2020), which also is an indicator of LoF intolerance. In contrast to the pLI score, the lower the LOEUF score, the more intolerant the gene is to LOF mutations.

To evaluate the impact of the *de novo* missense mutations to each gene, the missense Z-scores for all genes containing *de novo* missense mutations, as calculated by gnomAD (Lek et al., 2016; Samocha et al., 2014), were used to predict the tolerance of each gene to variation in place of the pLI scores when applying the methodology described above (Lelieveld et al., 2016) following 100,000 simulations. The missense Z-score, like both the pLI and LOEUF scores, indicates an intolerance of a gene to a certain kind of mutation which in this case is a missense mutation. The higher the missense Z-score, the more likely the gene is to be intolerant to these missense mutations. The presence of *de novo* missense mutations in missense intolerant genes was compared between predicted pathogenic and benign using a two-tailed Mann–Whitney U test in infertile samples and in controls independently. These statistical tests were all performed in R (R Core Team, 2020)by Dr Miguel Xavier of Newcastle University, however, the results were then analysed and investigated by me.

2.3.8 Protein-Protein Interactions

In order to predict the affected protein function and the potential role in disease, the protein-protein interactions between the genes containing a DNM were evaluated using STRING version 11 (Szklarczyk et al., 2017). All genes containing DNMs were input into the STRING analysis tool and processed to highlight any functional links between genes affected by DNM. This was performed with for genes containing predicted pathogenic protein altering DNMs and then compared to all those containing likely benign DNMs, and separately, for synonymous DNMs.

2.3.9 Burden Tests

After identifying likely pathogenic rare loss-of-function and missense DNMs, a gene-based burden test was performed to determine whether any of the 152 DNM genes contained more predicted pathogenic missense mutations in other cohorts of infertile men (n = 2,506) compared to fertile men (n = 5,784).The proportion of individuals with pathogenic variants in each of the genes carrying a DNM in the infertile cohort was statistically evaluated using two-tailed Fisher's Exact tests, individual p values were corrected using the Bonferroni method corrections to adjust for performing consecutive statistical tests and reduced the risk of Type I errors. Similarly, a gene-based burden test was performed to compare fertile fathers (n = 5784) with fertile mothers (n = 5,803) to investigate whether any of the sexes predominantly carried a greater number of rare pathogenic mutations. These statistical tests were all performed in R (R Core Team, 2020) by my colleague Dr Miguel Xavier of Newcastle University and the results were then analysed and investigated by me. The results of this analysis are discussed in Chapter 4 of this thesis and form the foundation of Chapter 5.

2.4 Functional Studies

The following methods relate to the data presented in Chapter 5 of this thesis. After the identification of a likely pathogenic *de novo* missense variant in *RBM5* and identification of a further six patients with heterozygous, predicted pathogenic, missense mutations within the same gene in other infertile men, functional studies were performed to investigate the potential damage these mutations may have on the function of this gene. The gene *RBM5* is a known splicing factor and has been well investigated regarding MI in mice and is also linked to lung cancer (Yu et al., 2020; O'Bryan et al., 2013). This study makes use of a known splicing target for the *RBM5* gene, named FAS in order to investigate any potential disruption to RBM5's ability to function as a splice factor caused by the specific point mutations found in these infertile men.

2.4.1 gBlock Gene Fragment Design and Preparation

To assess the effects of five of our *RBM5* mutations on the genes splicing capabilities, gBlock gene fragments (Integrated DNA Technologies) were used. Six different gBlocks gene fragments were ordered from Integrated DNA Technologies (IDT), five containing the RBM5 cDNA sequence with single base mutations detected in five different infertile patients (Chapter 4), the final containing the WT sequence.

To prepare the gene fragment for transformation into a p3XFLAG-CMV-14 expression vector (Sigma-Aldrich), a PCR reaction was used. The primers here introduced a flag epitope into the C-terminus of the RBM5 and left out the stop codon to ensure RBM5 reading frame reads immediately into the flag epitope sequence. The gene fragments were then purified into a volume of 50 µl using the QIAquick PCR Purification Kit and its associated protocol (Qiagen).

After the PCR amplification, 4 µl of each gBlock insert underwent a double restriction digest reaction using 2 µl of both the HindIII and EcoRI restriction enzymes and 5 µl of Promega Buffer E (Promega). The same digest was performed on the empty p3XFLAG-CMV-14 expression vector (Sigma-Aldrich) with a reaction mix of 10 µl vector, 2 µl HindIII, 2 µl EcoRI, 5 µl Promega buffer E (Promega) and 26 µl H_2O. The digests ran simultaneously at 37 °C for 3 hours and were then purified using the previously mentioned QIAquick PCR purification kit (Figure 2.4a). 15 µl of each digested insert was then added to 1 µl of digested plasmid, 2 µl of 10X Ligase buffer, and 2µl of T4 DNA Ligase to undergo a ligation reaction to form the plasmids for transformation of bacterial cells. This ligation reaction was left at 16 °C overnight.

The next morning, the ligation mix was then re-cleaved using the NotI restriction enzyme to remove any empty plasmids which did not ligate with the insert and still have an intact NotI restriction site. 20 µl of ligation mix was added to 2 µl of NotI, 23 µl of H_2O, and 5 µl of Promega buffer D. The reaction was incubated for 1 hour at 37 °C before the enzyme was deactivated at 65 °C for 10 minutes.

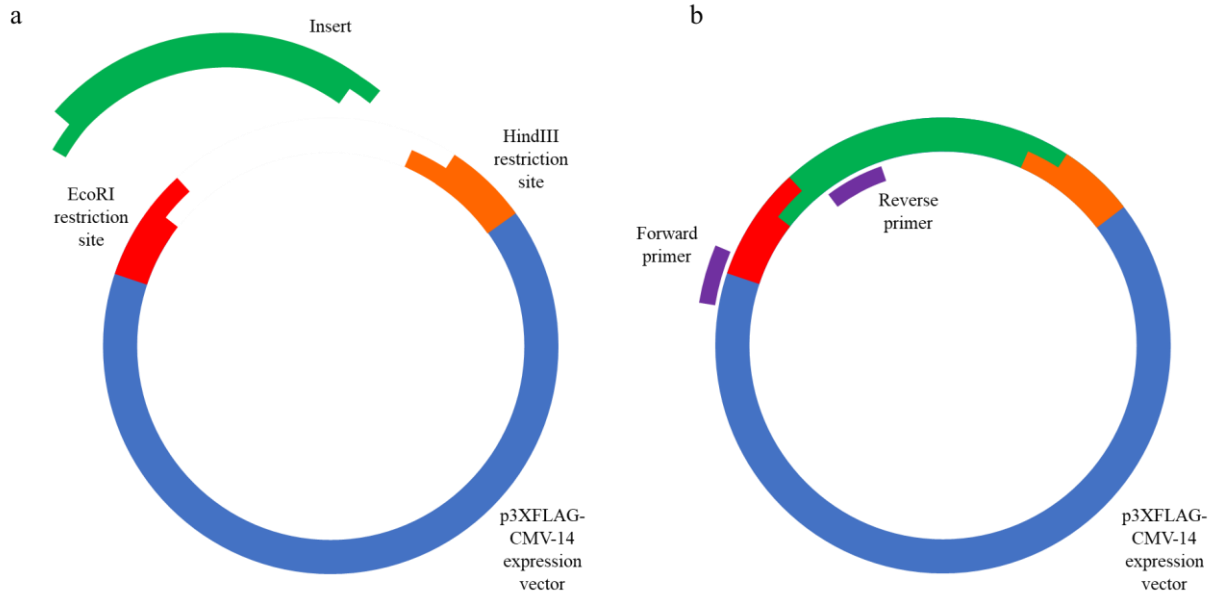


Figure 2.4: Plasmid containing insert. *a)* Digested EcoRI and HindIII sites within the p3XFLAG-CMV14 expression vector and the digested *RBM5* insert after restriction digest using HindIII and EcoRI restriction enzymes. *b)* Ligated plasmid containing the *RBM5* sequence insert and the primer positions for Colony PCR to confirm the insertion of the *RBM5*

sequence and the correct orientation of the sequence. The forward primer lies within the backbone of the plasmid to ensure the insert has been correctly orientated.

2.4.2 Transformation of Cells and Plasmid

In order to amplify the plasmid DNA for further use, 5 µl of recombinant plasmid DNA was added to 50 µl of host strain competent cells. As well as the 6 plasmids containing the RBM5 sequences, a Fas minigene, provided by Professor Juan Valcarcel's group at the Center for Genomic Regulation in Barcelona, was also transformed into the competent cells (Bonnal et al., 2008). The cell-plasmid mix was then left on ice for 10 minutes before a 1-minute heat shock at 42 °C and a further rest on ice for 5 minutes. The cell-plasmid mix was then added to 1 ml of Luria-Bertani (LB) Media (5g Tryptone 1 % w/v, 5g NaCl 1 % w/v, 2.5g Yeast extract 0.5 % w/v, 500ml ddH_2O) for a 1-hour cell recovery phase. The tubes were then spun down at 3000 rpm for 1 minute and the remaining pellet was then resuspended in 200 µl of LB broth before spreading onto an 1.5 % LB agar plate containing ampicillin. These plates were then inverted and incubated overnight at 37 °C in a stationary incubator before colony picking. To confirm the presence of the correct DNA insert within the plasmids, a colony PCR was performed. PCR primers were designed to confirm both the presence of the insert and the correct orientation with the forward primer lying in the backbone of the plasmid and the reverse lying within the insert (Figure 2.4b). Colonies were picked from the plates using a 10 µl pipette tip and dipped into the PCR mixture before being transferred to a tube of LB media and grown overnight again. The colonies containing the correctly formed plasmids were then processed using the QIAprep Spin Miniprep Kit (Qiagen).

2.4.3 Sanger Sequencing

To confirm that the plasmid contained the correct unaltered insert sequence, Sanger Sequencing was used. Four different primer pairs were designed to amplify the entire region of the insert with a 150 bp overlap to ensure a full sequence coverage. The Sanger sequencing protocol was performed as described in section 2.2.10.

2.4.4 Transfection

Confluent HEK293 cells grown in Gibco Dulbecco's Modified Eagle Medium (DMEM) containing 10 % Gibco Fetal Bovine Serum (FBS) were used for transfection. Media was removed from the t75 flask containing HEK293 cells and a wash with 5 µl of Gibco phosphate

buffered saline 7.4 pH (PBS) was applied to remove leftover media. 2 µl of Gibco Trypsin-EDTA (0.05 %) was then added, and the cells were left to incubate for roughly 5 minutes at 37 °C until all cells had lifted from the flask. Cells were then seeded onto 6 well plates with a final concentration of around $1x10^6$ cells/ml. The cells were then grown overnight at 37 °C to 70 % -80 % confluency.

To transfect the minigene and mutant *RBM5* into the HEK293 cells, 3.75 µl of Lipofectamine 3000 Reagent (Invitrogen) was diluted in 125 µl of Opti-MEM Reduced Serum Medium (Gibco) , all volumes given are for an individual well. 1 µl of *FAS* minigene was added to the diluted Lipofectamine 3000 reagent. DNA (G-blocks) was then diluted in a separate Eppendorf tube to a concentration of 200 ng/µl in 250 µl of Opti-MEM. 10 µl of P3000 Reagent (Invitrogen) was added then added to the diluted DNA. 125 µl of diluted DNA was then added to 125 µl of diluted lipofectamine 3000 reagent (Invitrogen) at a ratio of 1:1. The DNA-Lipofectamine complex was then incubated for 10 - 15 minutes at room temperature. 250 µl of the DNA-lipid complex was then added to each well and the cells were then incubated for 1 day at 37 °C .

To harvest the cells, media was removed from each well and 1 µl of PBS was used to wash the cells. 250 µl of Gibco Trypsin-EDTA (0.05 %) was added per well and the cells were left to incubate for 5 - 10 minutes at 37 °C. 1 µl of DMEM media was added to the cells before being transferred to an Eppendorf tube. The media-cell mix was spun for 5 minutes at 7500g in order to pellet the cells. After spinning, the media was removed and 1 µl of PBS was added to resuspend the pellet. The 1 µl of PBS-cell mix was then split into 500 µl in two individual Eppendorf tubes. The tubes were then spun down and the PBS removed leaving the cell pellet.

2.4.5 RNA Extraction and Processing

One of the two Eppendorf tubes containing cell pellets was used to extract RNA. 100 µl of TRIzol (Invitrogen) was added to each cell pellet and this was vortexed for 5 minutes at room temperature before adding 20 µl of chloroform (Sigma Aldrich). The tubes were then shaken vigorously before a 3-minute incubation at room temperature. After centrifugation for 15 minutes at 12000g and 4 °C , the clear liquid phase was moved to a sterile Eppendorf tube. The clear phase was then precipitated with 50 µl of chilled isopropanol, incubated at room temperature for 10 minutes and then centrifuged for 15 minutes at 12000g and 4 °C. The supernatant was then discarded, and the RNA pellet was washed with 250 µl of 75 % EtOH before vortexing, and a final centrifugation for 5 minutes at 7500g and 4 °C. The RNA pellet

was then dried at room temperature for 5 minutes before re-suspending in 25 µl of DepC H_2O. The concentration of all extracted RNA was measured using the Nanodrop machine, with a 260/280 ratio value of 2 and 260/230 ratio value between 2 - 2.2 indicating 'pure' RNA. All RNA samples were then diluted and normalised to a concentration of 100 ng/µl.

2.4.6 Reverse Transcription PCR

Primers were designed to analyse the levels of each isoform of the *FAS* minigene present in each sample and a reverse transcription PCR (RT-PCR) was performed. The QIAGEN OneStep RT-PCR Kit (Qiagen) was used to perform the RT-PCR following the manufacturers protocol.

The samples were then run on the QIAxcel Advanced (Qiagen) capillary electrophoresis system to visualise and quantify the levels of the two different isoforms produced. The concentration was determined for each isoform per sample and a percentage splice inclusion (PSI) was calculated for each sample using the formula:

$$PSI = \frac{concentration\ of\ upper\ band}{concentration\ of\ upper\ band + concentration\ of\ lower\ band} \times 100$$

The PSIs of the five variants were then compared to those of the wild type using paired t-tests to determine any statistically significant difference in the means.

2.4.7 Western Blot Analysis

SDS polyacrylamide gel electrophoresis (SDS-PAGE) was used to identify the RBM5 protein according to the molecular weight, described below.

To the second 500 µl Eppendorf containing the other half of the cell pellet, 25 µl of 2x loading buffer was added and the cells were resuspended. To denature the proteins, the cell suspension underwent sonication before boiling the samples at 95 °C for 5 minutes on a heat block. 20 µl of each sample was then loaded onto an NuPAGE 4 % – 12 % Bis-Tris Gel (Invitrogen) alongside the PageRuler Plus Prestained Protein Ladder (Thermo Scientific) and electrophoresed in 1x NuPAGE MOPS SDS Running Buffer at 200 V for 40 minutes. The transfer of proteins from the polyacrylamide gel to the nitrocellulose membrane was performed using a Mini Blot Module (Invitrogen) and run in 1x Transfer buffer containing 50 ml of NuPAGE Transfer Buffer (20X) (Invitrogen), 100 ml of Methanol and 850 ml of deionized H_2O. The transfer protocol used on the Invitrogen PowerEase Touch 350W (Invitrogen) ran for 60 minutes at 10 volts.

After protein transfer, the nitrocellulose membrane was blocked in SuperBlock Blocking Buffer (Thermo Scientific) for 40 minutes at room temperature under light agitation. The membrane was then incubated over night at 4 °C on a rotating mixer in the primary Anti-Flag antibody (Sigma-Aldrich) diluted to a concentration of 1 in 1000 in SuperBlock Blocking Buffer (Thermo Scientific). The membrane was then washed 3 times for 5 minutes each in PBS-T to remove any unbound primary antibody before being incubated in a secondary Anti-Mouse antibody (Invitrogen) diluted to 1 in 2000 concentration in SuperBlock Blocking Buffer (Thermo Scientific) for 1 hour at room temperature. The membrane was then washed twice for 5 minutes each time in PBS-T to remove any unbound antibody, and a final time for 5 minutes in PBS. This process was repeated with an α-tubulin primary antibody (Invitrogen) as a loading control and the Anti-Mouse secondary. This allowed for visualisation of both the protein of interest and the loading control on the same membrane.

Visualisation of the membrane was performed on the Amersham Imager 600 after incubating the probed membrane in SuperSignal West Pico PLUS Chemiluminescent Substrate consisting of equal parts SuperSignal West Pico PLUS Luminol/Enhancer Solution and SuperSignal West Pico PLUS Stable Peroxide Solution.

Chapter 3: Investigating *de novo* Mutations in a Panel of Candidate Infertility Genes Using Targeted Next Generation Sequencing

3.1 Introduction

Next Generation Sequencing (NGS) technologies have improved greatly since their inception in the early 2000s, with the data output and range increasing dramatically and the cost per sample decreasing over time. In 2014 it was announced that the $1000 genome had been achieved (Check Hayden, 2014) and Illumina now claim that the $100 genome is within reach (Illumina, 2022). The advancement of these technologies has allowed the investigation and discovery of a large number of disease genes in rare diseases where this was previously much more difficult (Prokop et al., 2018). A paper published in 2013 lists over 180 novel genes associated with rare diseases through the use of NGS techniques discovered up to that point (Boycott et al., 2013). Since their widespread introduction in 2010, NGS techniques have been rapidly replacing conventional methods for Mendelian disease gene discovery, from around 7 % of genes reported in OMIM in 2010 to 87 % in 2017 (Bamshad et al., 2019). This same study highlights the growing proportion of disease genes being identified as containing pathogenic *de novo* variants.

Whilst NGS has provided a huge breakthrough in the discovery of many disease genes, Male Infertility (MI) presents with a large amount of genetic heterogeneity, with over 2,000 genes known to be involved in spermatogenesis alone (Krausz & Antoni Riera-Escamilla, 2018; Kasak & Laan, 2021). This, alongside the difficulty in acquiring large cohorts of patient-parent trios, has proven to be an issue when performing whole exome sequencing (WES) and whole genome sequencing (WGS) to identify monogenic causes of MI. NGS often results in the detection of a multitude of possible candidate variants, with mutations found in only one individual which are potentially causative of a patient's infertility. This means that whilst a library of potential candidate genes is available, until there are more replicates, these genes cannot be clearly identified as MI associated genes - and a higher level of evidence is required (Houston et al., 2021; Oud et al., 2019).

A cost-effective and efficient technique to identify recurrently mutated genes is to use a multiplex targeted gene panel assay. This approach allows multiple targets to be sequenced in many samples at one time, with high precision. Single-molecule Molecular Inversion Probes (smMIPs) provide a simple and adaptable means of investigating these smaller gene panels (Oud et al., 2017).

3.1.1 Single-molecule Molecular Inversion Probes as a Research Tool

The smMIP assay was first described in a 2013 paper where researchers combined the already existing technique of Molecular Inversion Probes (MIPs) with the method of single molecule tagging (Hiatt et al., 2013). In the past, MIPs alone did not allow for the identification of the single genomic equivalents which were present in the input material. Resulting in a technique that, whilst boasting a low cost per sample and low sample input requirements, was not as sensitive or specific as often required (Turner et al., 2009; Shen et al., 2011; O'Roak, Vives, Fu, et al., 2012). With the addition of single molecule tagging, in the form of Unique Molecular Barcodes (UMBs), each individual DNA molecule sequenced can then be quantified to allow for consensus calling of single genomic equivalents and deduplication of PCR replicates (Kinde et al., 2011; Casbon et al., 2011; Hiatt et al., 2010). The structure of these smMIPs can be seen in Figure 2.2 in Chapter 2 of this thesis.

Since then, there have been multiple studies using smMIPs to identify causative mutations within genes known to be linked with specific diseases. A recent study compared a smMIP assay, which was designed to target two specific exons within the *GNAS* gene, to a TaqMan genotyping assay to detect two specific point mutations in the *GNAS* gene (Bekers et al., 2019). The authors concluded that although both techniques showed high concordance levels, the smMIP assay allowed for the sequencing of whole exons rather than just the single point mutations, and thus providing additional information and the opportunity for a broader range of mutations on the *GNAS* gene to be discovered.

In another study, a smMIP assay was used to further investigate what appeared to be germline *de novo* mutations in the gene *SCN1A* (de Lange et al., 2019). The smMIP assay was utilised to identify low-level mosaicism of pathogenic *SCN1A* variants in parents of children with suspected *de novo SCN1A*-related epilepsy (Dravet Syndrome). It was found that the targeted smMIP panel perform to a higher standard than Sanger Sequencing alone by identifying low-level mosaicisms in an additional 5 % of all families involved in the study. This was particularly relevant for patients suffering from Dravet Syndrome as low-level mosaicism in the parents could have important implications for any future offspring, and thus genetic counselling would be required for the potential parents.

Furthermore, smMIPs have been shown by our group and collaborators to potentially increase the diagnostic yield of detecting genetic variations in males affected by severe infertility (Oud et al., 2017), particularly when focusing on genes previously identified as a known cause of MI, or as likely candidate genes to cause MI when containing damaging mutations. In this study, smMIPs were used to target and screen 107 of these candidate genes in 1,112 infertile males all

presenting with idiopathic infertility. Following sequencing and variant analysis, 1 % - 1.5 % of these previously idiopathic patients received a probable diagnosis due to detection of chromosomal anomalies or AZF deletions. The authors of this study predict that this diagnostic yield value could be even further increased with the confident linkage of more genes to male infertility, something which can also be achieved by utilising the smMIP assay.

For this Chapter of my thesis, a smMIP assay was used to sequence 54 genes which were initially highlighted as potential novel candidate MI genes in a study by Dr Manon Oud, in collaboration with our group which provided the basis for the WES work described in Chapter 4 of this thesis. These 54 genes contained rare *de novo* mutations in individual males suffering from severe idiopathic MI and were identified as likely to affect male fertility when their regular function is disrupted. A list of all 54 genes can be found in Table 2.1 of this thesis. Each of these genes were identified as containing a DNM in a singular infertile male meaning despite the determined relevance of the genes to the disease of interest, there was little evidence to solidify them as disease genes. The focus of this chapter was to validate the relevance of these 54 genes as novel candidate infertility genes by identifying additional patients presenting with potentially damaging mutations in these genes of interest. In total 2904 smMIPs were designed to target all 54 of these genes, aiding in the identification of any potential replications of *de novo* mutations in these genes within a new cohort of infertile males. The smMIP assay was performed on 299 samples, including 75 new patient-parent trios in order to identify possibly causative *de novo* mutations.

3.2 Aims

In this Chapter the aims of this work are:

1. To design, assemble and rebalance a smMIP pool targeting 54 potential novel candidate male infertility genes to an optimal efficiency.
2. To utilise this rebalanced smMIP pool to identify and assess the pathogenicity of any rare coding *de novo* mutations in the DNA of 75 infertile males.
3. To highlight and further investigate any of the 54 genes containing rare coding de novo mutations in the new cohort as potential candidate male infertility genes.

3.3 Methods

In this study, four control samples were used for initial control runs followed by a cohort of 299 infertile patient and parental samples. Of these 299 samples, 225 were made up of 75 patient-parent trios with an additional 74 singletons. These 75 patient-parent trios are the focus of this chapter and are also included in Chapter 4 where they undergo WES to investigate DNMs in all genes. These are fully described in Section 2.1.1 of Chapter 2: Materials and Methods. The methodology for designing and preparing the smMIP pool is found in Sections 2.2.1 and 2.2.2. Additionally, Sections 2.2.3 - 2.2.7 of the Materials and Methods describe the sample processing and sequencing preparation including the rebalancing stages. The bioinformatic processing of sequencing data and subsequent variant filtration, analysis and validation are described in Sections 2.3.8 - 2.3.10. The variant filtration and prioritisation workflow used is identical to that shown in Figure 2.3 of Chapter 2 of this thesis.

All smMIP design, pool preparation, sample preparation, variant analysis and Sanger Sequencing was performed by me. The initial control runs (Run 1 and Run 2) were also entirely performed by me and Run 3 using the NovaSeq platform was partially assisted by the Genomics Core Facility (GCF) at Newcastle University. Processing of the Fastq files from Run 3 to produce Variant Call Format (VCF) files was performed using a custom pipeline designed by Dr Simon Cockell at the Bioinformatics Support Unit at Newcastle University.

The 54 genes sequenced using this smMIP targeted panel were all previously identified in a preliminary WES study performed by Dr Manon Oud as potential male infertility candidate genes. These genes were identified to contain a single de novo mutation of interest in individual patients from an initial 99 infertile males sequenced in Nijmegen, and these WES results are included and expanded upon in Chapter 4 of this thesis. The breakdown of samples used in different experiments throughout this thesis is represented in Figure 2.1 of the Materials and Methods chapter. The initial DNMs in these 54 genes were highlighted as either “possibly causative” or “unclear” based on both variant level evidence and gene level evidence using an earlier less refined version of the methodology described in Section 2.3.8-2.3.10 of the Materials and Methods chapter of this thesis.

3.4 Results

3.4.1 Initial Control Run using a 1x smMIP pool

To determine the success and efficiency of the initial smMIP pool, quality scores were analysed which were generated from performing this assay on a small number of control samples. The first metrics used to determine the success of the initial smMIP sequencing run were the QC scores given through the BaseSpace platform (Illumina). The initial sequencing run performed very well with 94.62 % of all clusters passing the built-in chastity filters. Chastity is defined as the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities. Clusters pass this filter if no more than one base call has a chastity value below 0.6 in the first 25 cycles . The capacity of the flow cell was for 30,000,000 reads and this run provided 28,261,210 reads (94.20 %) after any low-quality reads were removed. Additionally, in this run 95.26 % of all clusters achieved a Q score greater or equal than 30, where a Q score of 30 represents a 99.9 % base calling accuracy (Liao et al., 2017). The smMIP pool was tested on four samples which have been named Sample 1, Sample 2, Sample 3, and Sample 4 - for both Runs 1 and 2.

As with all experiments performed within this Chapter, the samples used in this run were all pooled equimolarly and thus, after sequencing, it would be expected that each sample should be represented by an equal number of reads. Figure 3.1 shows the distribution of the reads passing filter between the four samples and it can clearly be noted that Sample 1 took up a higher percentage of the reads at around 37 % and Sample 4 underperformed with only around 12 % of the reads being assigned to this sample.

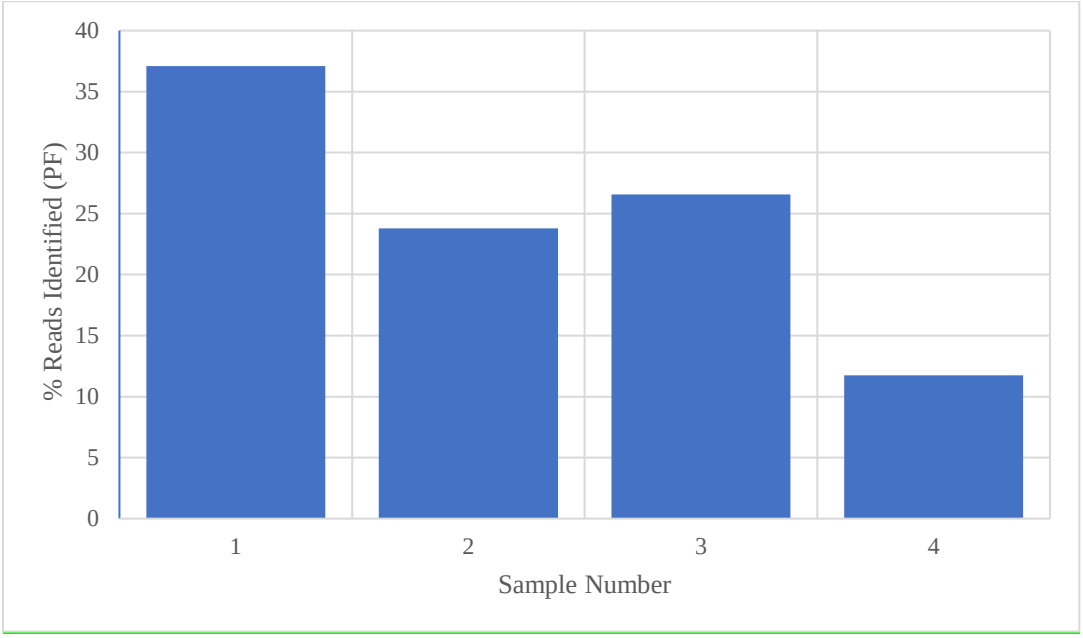


Figure 3.1: Percentage of passing filter reads identified per sample in the initial control run using a 1x smMIP pool.

After the run data had been processed via the in-house bioinformatics pipeline, the coverage data was available to assess the performance of each individual smMIP within each sample. To perform an evaluation of the overall quality of sequencing of each sample, the average coverage of each smMIP was identified, along with additional values including the maximum number of reads across the smMIPs and the total reads assigned to each sample. All coverage data presented was collected after the deduplication of the reads using the UMB of each smMIP. The coverage data for Run 1 can be seen represented in Table 3.1. As seen within the initial metrics, there is an obvious imbalance across the four samples, with Sample 4 significantly underperforming in comparison to the other three samples with only 9.5 % of all the deduplicated smMIP reads being assigned to this sample. Because of this underperformance, this sample was removed from further analysis and rebalancing runs.

The median number of reads per smMIP along with the mean, gives a solid indication of how the smMIP pool is performing on average, with both Samples 1 and 3 performing similarly with an average of 161 and 153 reads per smMIP respectively, as well as medians of 112 and 110 reads per smMIP. Across the three successful samples, an average of 81 % of all smMIPs were covered at a read depth of at least 10 reads.

Sample ID	S1	S2	S3	S4
Mean	161	126	153	46
Median	112	75	110	18
Min	0	0	0	0
Max	2006	1778	1400	593
StDev	178	151	158	66
Total	**468,919**	**366,720**	**443,042**	**134,428**

Table 3.1: Coverage data for Control Samples 1 – 4 from Run 1 using a 1x smMIP pool. The sample names given in the table are shortened whereby the "S" represents Sample. The metrics given here represent the number of reads per smMIP within each sample. From this table, it is clear that Sample 4 (S4) has greatly underperformed in comparison to the previous 3 samples with less than half of the total assigned reads compared to the next best performer.

In order to assess which smMIPs were underperforming and required rebalancing in the second control run, the Read Fraction (RF) value was determined for each smMIP. As mentioned in Chapter 2, the ideal RF for each smMIP was defined to be 0.0003, determined by the total number of smMIPs and their fraction of one individual read when working at equal efficiency (Oud et al., 2017). Using the assigned RF values, the smMIPs were split into four distinct groups: Overperformers with an RF >0.003, Within Range with an RF of >0.00003 and <0.003, Underperformers with an RF of <0.00003, and Zero with smMIPs showing no assigned reads. This data can all be found in Table 3.2.

On average, 2,281 out of 2,904 smMIPs (78.5 % of total smMIPs) performed within the acceptable range, with Sample 3 appearing to perform the best with 2,343 on-target smMIPs (81 % of total smMIPs). Across the three samples, approximately 12 % of all smMIPs underperformed and a further 10 % had zero reads assigned to them. Only 7 % (n = 202) of smMIPs gave zero reads in Sample 3 compared to 12 % (n = 354) in Sample 2. Out of these 202 smMIPs which had given zero reads in Sample 3, 160 were found to also have zero reads across Samples 1 and 2. Of the remaining 42 smMIPs with zero reads in Sample 3, 38 were underperformers across the two other samples and four were on target in either one, or both, of the other samples.

It was decided that Sample 3 had shown the best overall performance within the smMIP pool, due to its higher number of on target smMIPs (81 %) and significantly lower number of smMIPs giving zero reads (7 %). Thus, the data from this sample would be used to rebalance the smMIP pool for a second control run. To action this, all 560 smMIPs classified as Underperformers or

Zeros were then added into the new smMIP pool at a 10x concentration in the hopes of improving the performance of these smMIPs.

It was noted that many of these smMIPs which needed rebalancing belonged to specific genes which contain large numbers of smMIPs including: 83 % of all smMIPs in *APC2* (n = 85), 74 % in *HELZ2* (n = 90) and 53 % in *DGKZ* (n = 89). There were, however, several smaller genes which showed a high percentage of smMIPs requiring rebalancing including 59 % of *C9orf50* (n = 13) and 56 % of *CD81* (n = 10). Over 50 % of smMIPs needed rebalancing in six of the original 54 genes with only three genes (*APOBEC3G*, *ODF1* and *ZBTB39*) not containing any underperforming smMIPs. All this data is represented in Table 3.6.

smMIP quality	RF	Run 1 S1	Run 1 S2	Run 1 S3	Average Run 1
Zero	**0**	279	354	202	**278**
Underperformer	**<0.00003**	332	334	358	**341**
Within Range	**>0.00003 and <0.003**	2289	2212	2343	**2281**
Overperformer	**>0.003**	4	4	1	**3**
Total MIPs	**-**	2904	2904	2904	**2904**

Table 3.2: Read Fraction (RF) dataset for Samples 1 - 3 in Run 1 using a 1x smMIP pool. This table highlights the relative consistency in the number of underperformers across all three samples within the first run, whilst also showing an increase in the number of smMIPs performing within range in Sample 3

3.4.2 Control Run 2 including rebalanced smMIPs in a new 10x pool

The second control run was performed with the purpose of testing a newly rebalanced smMIP pool containing all 2904 smMIPs with the 560 Underperformers and Zeros highlighted from Run 1 at a 10x concentration within the pool. The 10x pool was initially processed using the three successful control Samples (1 - 3) from Run 1, however at a late stage in the experiment, Sample 3 failed and thus only Samples 1 and 2 were fully processed with the rebalanced 10x pool. The samples processed using the newly rebalanced 10x pool are referred to as Sample 1_10x and Sample 2_10x in all tables and figures within this chapter. An additional two samples were run on this flow cell; however, these samples were not part of this experiment and will not be further discussed in this chapter.

As with the first MiSeq control run, the metric scores gathered from BaseSpace were consistent with a high-quality sequencing run. The capacity of this flow cell again was 30,000,000 million reads and, after removing reads which did not pass filter, 29,020,418 reads were achieved meaning 96.7 % of the flow cell capacity was used. The number of clusters passing the inbuilt quality filters was 94.4 % and the average number of clusters with a Q score ≥ 30 was 95.0 %. Looking at the distribution of reads before any bioinformatic processing, it was observed that of the 15,195,091 reads passing filter assigned to the two samples captured with the rebalanced smMIP pool, 54 % were assigned to Sample 1_10x and 46 % to Sample 2_10x (Figure 3.2).

Post-deduplication coverage data showed that the difference in number of reads assigned to each of these samples was further reduced with 49 % assigned to Sample 1_10x and 51 % to Sample 2_10x. The two 10x pool samples in Run 2 showed lower mean and median number of reads when compared to the samples using the 1x pool in Run 1 with a reduction from a median of 112 and 75 reads per smMIP in Run 1 to 64 and 63 reads per smMIP in Run 2 (Table 3.3). Similarly, to Run 1, Samples 1 and 2 showed an average of 82 % of all smMIPs at a read depth of 10 or more reads.

Sample ID	S1_10X	S2_10X
Mean	105	107
Median	64	63
Min	0	0
Max	1,893	2,015
StDev	141	153
Total	303,858	311,895

Table 3.3: Coverage data for Samples 1 and 2 from Run 2 using a 10x smMIP pool. The sample names given in the table are shortened whereby the "S" represents Sample. The metrics given here represent the number of reads per smMIP within each sample. From this table it was seen that the smMIPs from the individual 10x pools performed similarly to each other despite a slight increase in the number of reads assigned to Sample 2_10x compared to Sample 1_10x.

The most important scores to compare the two runs after sample QC checks were the RFs, these values can be seen in Table 3.4 for the rebalanced 10x pool compared to the initial Run 1 values using a 1x smMIP pool. The difference between the two pools and their performance could be seen clearly. Samples 1 and 2 had an average of 317 smMIPs which were giving zero reads in the initial 1x smMIP pool run, however after utilising the 10x pool, an average of only 186

smMIPs gave zero reads in the same two samples (Figure 3.2). Samples 1 and 2 showed only 17 % and 19 % of smMIPs underperforming or giving zero reads in the 10x pool, compared to 21 % and 24 % in the initial run with the number of zero reads decreasing by almost 50 % in Sample 2. It was noted that in the second run, there were 81 smMIPs which gave zero reads in both Sample 1 and Sample 2 within the 10x pool that were also 'Zero' read smMIPs in the initial 1x Run 1 samples.

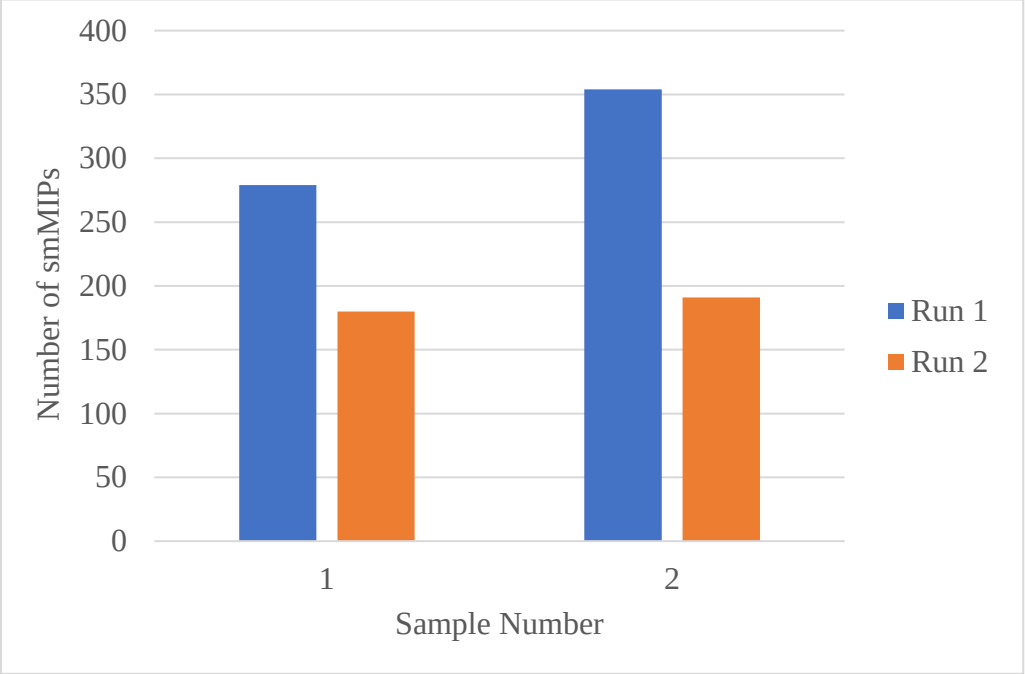


Figure 3.2: Number of smMIPs giving zero reads in Samples 1 and 2 across both initial runs. It was seen that after rebalancing in Run 2 with 560 smMIPs at a 10x concentration within the smMIP pool, both Samples 1 and 2 had large reduction in the number of smMIPs giving Zero reads. In Sample 1, a reduction of 35 % of smMIPs giving Zero reads is seen and in Sample 2, 46 %.

Of the three genes mentioned in Section 3.3.1 which had a large number of Underperformers or Zeros across all samples in the initial run, two of the genes (*APC2* and *HELZ2*) saw a decrease in the number of smMIPs underperforming or giving zero reads - with a decrease from 85 (83 %) in Run 1 to 76 (74 %) in Run 2 for *APC2*, and from 90 (74 %) to 86 (71 %) in *HELZ2* across the same run sets. This data can be visualised in Table 3.5 and was calculated by determining the average RF of each smMIP across the two 10x pool samples (Sample 1_10x and Sample 2_10x) giving a total of 517 smMIPs which underperformed or gave zero reads on average across the two samples. Of these, 53 were classified as within range in Sample 1_10x and 28 within range in Sample 2_10x. After the rebalancing, an increase was seen in the number of genes showing no underperforming smMIPs from three to eight. As with this rebalancing

round, more than 80 % of all the smMIPs from the 10x pool were within an acceptable performance range, the 10x pool was used in the final infertile sample cohort sequencing run.

smMIP quality	RF	Run 1 S1	Run 1 S2	Run 2 S1_10X	Run 2 S2_10X	Average Run 1	Average Run 2
Zero	**0**	279	354	180	191	**317**	**186**
Underperformer	**<0.00003**	332	334	317	373	**333**	**345**
Within Range	**>0.00003 and <0.003**	2289	2212	2393	2325	**2251**	**2359**
Overperformer	**>0.003**	4	4	14	15	**4**	**15**
Total MIPs	**-**	2904	2904	2904	2904	**2904**	**2904**

Table 3.4: Read Fraction (RF) dataset comparing Samples 1 and 2 with the 1x smMIP pool to their performance with the rebalanced 10x smMIP pool. This table highlights the increased number of smMIPs performing within range after application of the new 10x smMIP pool to Samples 1 and 2 when compared to the initial 1x pool utilised in Run 1.

3.4.3 Targeted Sequencing using a multiplex smMIP panel on a cohort of 299 infertile males and parental samples

The initial QC checks for the final 299 sample run were performed by the Genomics Core Facility (GCF) at Newcastle University (NCL) throughout the sequencing process. After processing through the BSU pipeline, the following analyses were performed.

The initial metrics for the 299 samples are represented in Figure 3.3 as the mean number of reads per smMIP per sample. Whilst there are a handful of samples originating from Nijmegen (NIJ) that had many sequencing reads assigned on average to each smMIP, the general range between all four sample groups is fairly consistent with an average read per smMIP of 203 in NCL blood group, 213 in the NCL Saliva group, 240 in the NIJ Blood group and 243 in the NIJ Saliva group. Across these four cohorts, it was seen that the lowest mean number of smMIPs per sample ranges from 66 reads in the NCL Saliva category and 99 in the NCL Blood group. Across all 299 samples, the average and median read depth was 236x and 140x, respectively. It should also be noted that none of the samples in the sequencing run failed, with all having a satisfactory coverage and that all data is post-deduplication of the smMIPs.

Looking at the overall coverage, the lowest number of reads assigned to a single sample was 189,227 reads and the highest contained 4,251,930 reads. The sample with the highest reads can be visualised on the graph in Figure 3.3, as one of the few outliers within the Nijmegen samples with a disproportionately large number of average reads per smMIP (n = 1464).

When assessing the success of the run, it was possible to analyse the proportion of smMIPs captured at a depth of at least 10 reads with an average of 92 % of all smMIPs having a coverage of 10 or more reads across all 299 samples. Figure 3.4 represents the percentage of smMIPs with ≥10 reads in each sample with the two lowest performing samples having 78 % of all smMIPs at 10 reads or more. A further eight samples have between 80 % - 84 % of all smMIPs with ≥10 reads, and 62 additional samples were seen with 85 % - 89 %. The remaining 227 samples saw over 90 % of all smMIPs achieving a read depth of at least 10 reads.

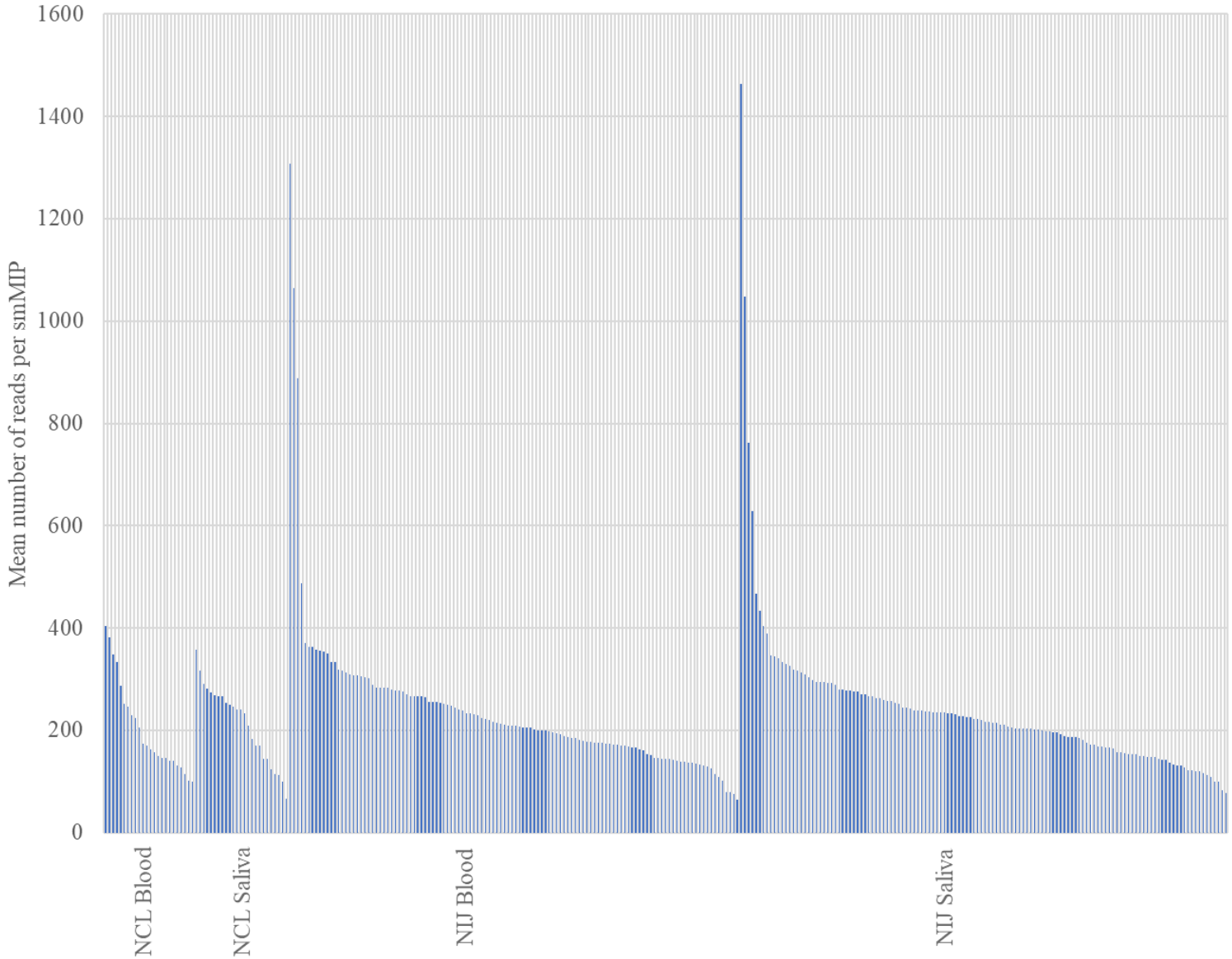


Figure 3.3: Comparison of Newcastle (NCL) and Nijmegen (NIJ) Blood and Saliva samples using the average number of reads per smMIP as a comparison tool.

The RF analysis of the 299 samples highlighted the improvement of the smMIP pool within this final run with an average of 87 % of all smMIPs performing within range across all 299 samples. This data is shown in Figure 3.5. The highest performing sample shows over 92 % of all the 2904 smMIPs performing Within Range, only 5 % Underperforming and 2 % giving Zero reads. This contrasts with the worst sample with 81 % Within Range, 13 % Underperforming and 5 % giving Zero reads. The fact that all the samples have been successfully sequenced, have a high percentage of their smMIPs performing within range, and good coverages across all the smMIPs gives a strong indication that the smMIP pool has been well rebalanced.

Upon further investigation, only seven smMIPs consistently gave Zero reads across all 299 samples and these smMIPs had also been Zero read smMIPs in Runs 1 and 2. Of these seven smMIPs, five lay within individual genes with the last two relating to the gene *EVC*.

To compare the Underperformers in the genes mentioned in Sections 3.3.1 and 3.3.2 to those in the final run, only 49 smMIPS underperformed in the gene *APC2* (48 %), 25 in *DGKZ* (28 %) and 35 in *HELZ2* (29 %), showing a vast improvement in the overall quality of the smMIP pool across the three runs. The gene *CD81* was seen to have the worst performance of smMIPs with 50 % (n = 9) underperforming or giving zero reads in comparison to the genes *RASEF, CDC5L, STX7, FZD3, ODF1* and *CWC27* which all had no smMIPs underperforming.

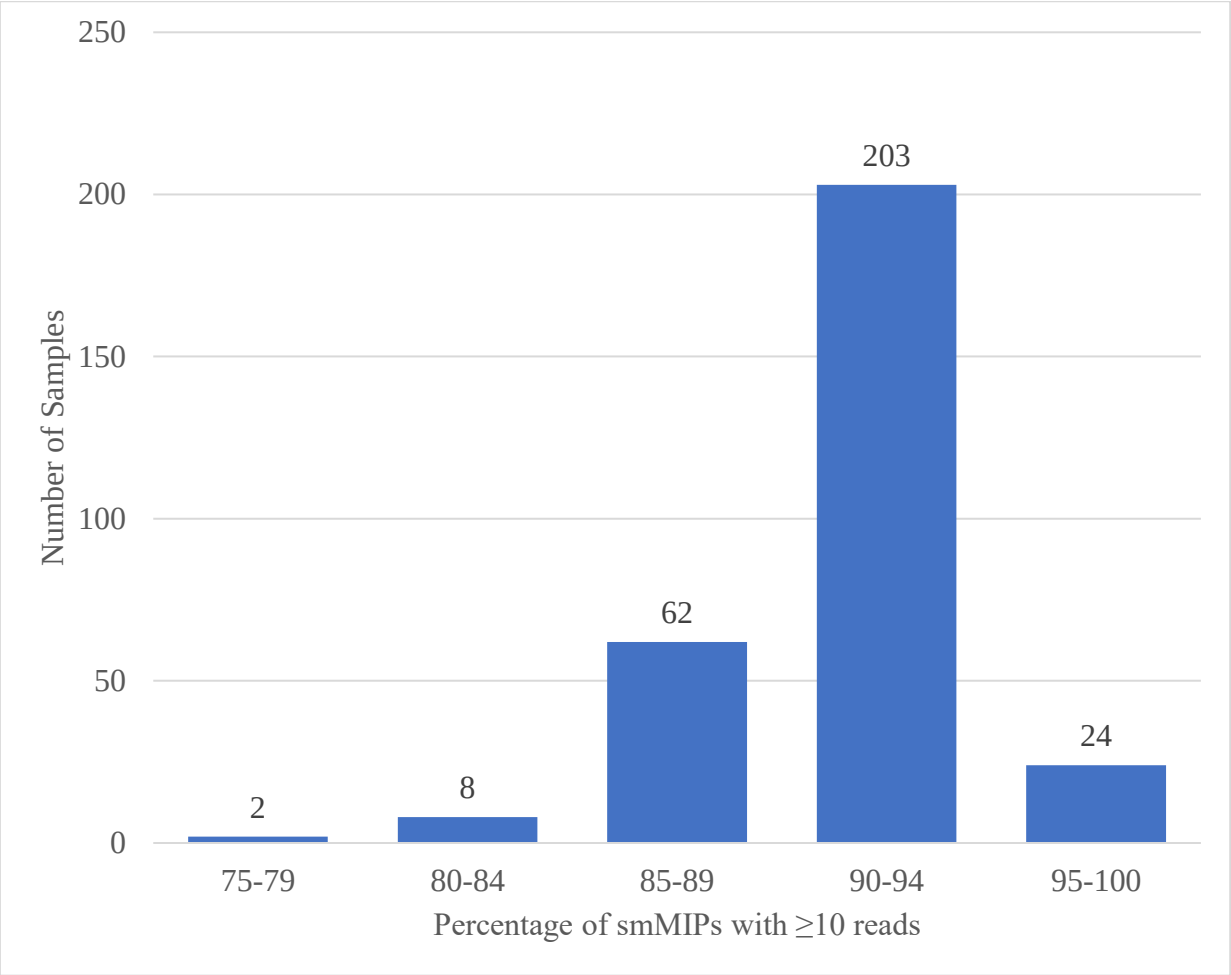


Figure 3.4: A graph to show the percentage of all smMIPs with at least 10 reads within 299 samples. It was seen that in 76 % of all samples (n = 227) sequenced in the third and final run using the 10x smMIP pool, 90 % or more of the 2904 smMIPs were sequenced with at least 10 unique reads. With 24 samples achieving more than 95 % of all smMIPs at ≥10 reads. Only two samples (1 %) of all samples had less than 80 % of all smMIPs covered with ≥10 reads with actual values of 78 % in both samples.

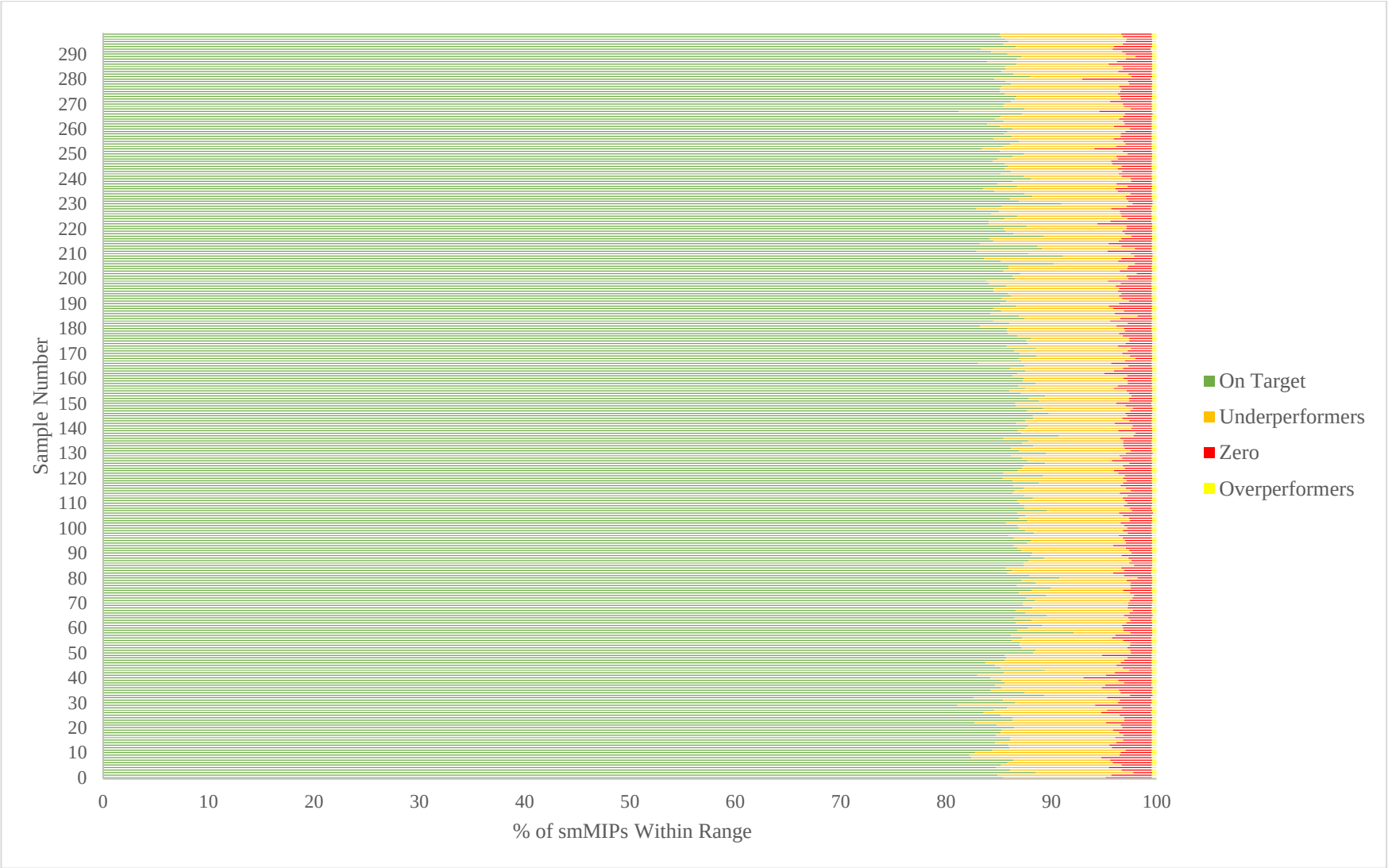


Figure 3.5: Visual representation of the performance of all 2904 smMIPs across 299 infertile patient and parental sample

Gene	Gene Length (bp)	Total smMIPs	smMIPs for rebalancing after Run 1	% of all smMIPs	Under-performers and Zeros Run 2	% of all smMIPs	Under-performers and Zeros Run 3	% of all smMIPs
ABCF3	2246	48	2	4 %	2	4 %	6	13 %
ABLIM1	7560	55	4	7 %	3	5 %	3	5 %
APC2	10179	103	85	83 %	76	74 %	49	48 %
APOBEC3G	1564	22	0	0 %	1	5 %	2	9 %
ATP1A1	3670	57	4	7 %	2	4 %	3	5 %
ATP8A1	8270	81	8	10 %	4	5 %	4	5 %
ATP8B4	5688	74	5	7 %	3	4 %	3	4 %
C9orf50	1945	22	13	59 %	10	45 %	8	36 %
CAPN10	2621	45	20	44 %	18	40 %	8	18 %
CD81	1482	18	10	56 %	13	72 %	9	50 %
CDC5L	6241	43	3	7 %	0	0 %	0	0 %
CDK5RAP2	6232	98	4	4 %	2	2 %	2	2 %
CRACR2A	2704	45	2	4 %	4	9 %	3	7 %
CWC27	2065	34	4	12 %	0	0 %	0	0 %
DGKZ	3588	89	47	53 %	55	62 %	25	28 %
DNAJC2	2291	37	4	11 %	0	0 %	2	5 %
DNMT1	5274	99	8	8 %	4	4 %	6	6 %
EMILIN1	3890	45	22	49 %	29	64 %	18	40 %
ERG	4904	32	4	13 %	4	13 %	3	9 %
EVC	6427	64	6	9 %	5	8 %	7	11 %
EXOSC9	1587	27	3	11 %	1	4 %	1	4 %
FBXO5	2054	20	4	20 %	2	10 %	2	10 %
FNDC8	1352	15	5	33 %	5	33 %	5	33 %
FUS	1824	35	4	11 %	1	3 %	3	9 %
FZD3	13770	26	1	4 %	0	0 %	0	0 %
HELZ2	10002	121	90	74 %	86	71 %	35	29 %
HNRNPL	2142	34	7	21 %	9	26 %	10	29 %
HR	5474	63	30	48 %	33	52 %	18	29 %
ILVBL	2305	44	10	23 %	5	11 %	7	16 %
IQSEC1	7700	61	10	16 %	16	26 %	9	15 %
LEO1	2165	34	1	3 %	2	6 %	1	3 %
LZTS2	2862	33	15	45 %	11	33 %	8	24 %
MCM6	3753	44	1	2 %	2	5 %	2	5 %
MPRIP	15419	61	5	8 %	7	11 %	9	15 %
MYRIP	5073	46	2	4 %	5	11 %	5	11 %
NEO1	7108	76	5	7 %	4	5 %	4	5 %
NUP210	7206	111	8	7 %	7	6 %	8	7 %
ODF1	1026	12	0	0 %	0	0 %	0	0 %
OSBPL3	6749	47	1	2 %	1	2 %	1	2 %
PLCL1	6696	42	6	14 %	8	19 %	3	7 %
RASEF	5585	43	7	16 %	0	0 %	0	0 %
RPA1	4847	36	2	6 %	0	0 %	3	8 %
SIX2	2206	14	6	43 %	3	21 %	2	14 %
SMC2	5928	72	1	1 %	2	3 %	4	6 %
SOGA1	14218	80	15	19 %	15	19 %	13	16 %
STX7	15845	20	1	5 %	0	0 %	0	0 %
STXBP2	1885	45	27	60 %	24	53 %	18	40 %
TACC2	9704	135	11	8 %	4	3 %	4	3 %
TLN2	12023	143	3	2 %	6	4 %	6	4 %
TMEM62	2707	34	4	12 %	4	12 %	1	3 %
TOPAZ1	5334	83	2	2 %	3	4 %	5	6 %
U2AF2	2146	29	14	48 %	13	45 %	11	38 %
WDR17	7333	80	4	5 %	1	1 %	2	3 %
ZBTB39	6268	27	0	0 %	2	7 %	2	7 %

Table 3.5: Underperformers and Zero smMIPs across three separate sequencing runs in all 54 genes. The smMIPs represented in column smMIPs for rebalancing after Run 1 are those which were found to be underperforming or giving zero reads within Sample 3 in Run 1 and were then rebalanced in the 10x pool in Run 2. Underperformers and Zeros Run 2 were determined by taking an average of the RF of all smMIPs in Sample 1_10x and Sample 2_10x. Underperformers and Zeros Run 3 were identified after the average RF was calculated for each smMIP across all 299 samples within the third Run.

3.4.4 Filtration of de novo Variants Identified in Patient-Parent Trios

After the coverage files were assessed, annotated VCF files were analysed to highlight any potentially causative rare variants found in one of the 54 genes sequenced using the smMIP panel. This, and all further analysis steps were only performed on the 75 patients who had undergone sequencing as a trio.

Initially, on average 250 unique variants were identified per sample before any filtration took place. The first filtering stage used to prioritize these variants included removing all variants which have a gnomAD allele frequency ≥0.1 % or those for which no global frequency is available. After removal of all common variants and a final filtration for only non-synonymous coding variants, a total of 13 rare, non-synonymous, coding *de novo* variants were identified. These 13 variants were present in 12 different samples, and all were located within a different gene each and can be seen in Table 3.6. It was noted that all 13 variants had a Minor Allele Frequency (MAF) between 10.34 % in the lowest sample and 33.33 % in the highest, with 10 % being the cut off for *de novo* variants to be called within the bioinformatics pipeline.

The variants then underwent a manual Binary Alignment Map (BAM) file check and from these, it was determined that five samples contained a MAF greater than 5 % in one of the parents (See Table 3.6), casting doubts on whether the variants were truly *de novo.* So, following this, Sanger Sequencing was performed in order to visualise the variant locations in the initial DNA samples to give an accurate determination of whether the variants were indeed true *de novo* variants.

Sample	Original Location	Codons	Coverage	Minor Allele Frequency	Gene	Consequence	BAM Frequency Proband	BAM Frequency Mother	BAM Frequency Father	Sanger Sequencing Result
MI_0010_P	chr10:116196083	cAg/cGg	489	29 %	*ABLIM1*	missense_variant	29 %	1 %	0 %	Inherited from Father
MI_0010_P	chr21:39947608	cCg/cTg	480	14 %	*ERG*	missense_variant	15 %	1 %	0 %	Inherited from Mother
MI_0019_P	chr5:64181325	cAa/ca	116	10 %	*CWC27*	frameshift_variant	10 %	0 %	2 %	Variant not present
MI_0031_P	chr2:136602151	agA/agC	346	16 %	*MCM6*	missense_variant	16 %	0 %	1 %	Inherited from Mother
MI_Proband00660	chr3:13417817	Gcg/Acg	150	22 %	*NUP210*	missense_variant	22 %	0 %	5 %	Inherited from Father
MI_Proband01024	chr20:62191548	Gtg/Atg	10	20 %	*HELZ2*	missense_variant	20 %	0 %	0 %	Variant not present
MI_Proband01181	chr12:3728463	Cgg/Tgg	302	15 %	*CRACR2A*	missense_variant	15 %	0 %	7 %	Inherited from Father
MI_Proband01320	chr6:153296225	aAt/aGt	48	33 %	*FBXO5*	missense_variant	33 %	7 %	0 %	Inherited from Mother
MI_Proband01586	chr10:123844023	Ctg/Gtg	122	16 %	*TACC2*	missense_variant	16 %	0 %	8 %	Inherited from Father
MI_Proband01697	chr2:27306220	cGc/cAc	76	14 %	*EMILIN1*	missense_variant	15 %	0 %	0 %	Inherited from Father
MI_Proband01968	chr8:21982863	Cgg/Tgg	44	14 %	*HR*	missense_variant	13 %	2 %	0 %	Variant not present
MI_Proband01982	chr19:7707105	cGc/cAc	20	30 %	*STXBP2*	missense_variant	30 %	0 %	0 %	Variant not present
MI_Proband02057	chr15:63125775	Gcc/Acc	465	17 %	*TLN2*	missense_variant	18 %	0 %	9 %	Inherited from Father

Table 3.6: All rare *de novo* variants highlighted after filtering. After filtration, 13 rare *de novo* variants were identified. The BAM frequencies can be seen within the table for both the proband and the parental samples and a final classification of the variant after Sanger Sequencing is present.

3.4.5 Validation of Rare de novo Variants Identified in Infertile Patients

The Sanger Sequencing determined that all 13 variants were either false positives (n = 4) or were inherited from a parent (n = 9). Of the four variants which were proven to be false positives, three had an alternative allele coverage of less than 10 reads and the fourth showed the variant in 12 reads however this only provided an MAF of 14 %. This can be seen represented in Figure 3.5, which shows data for the NM_005144:c.1711C>T variant in sample MI_Proband01968 that had a coverage of 44 in the initial sequencing files and an MAF of 14 %. The BAM files produced from the sequencing run further confirmed these values and showed only a 2 % MAF in the mother and 0 % in the father. However, Sanger Sequencing clearly highlighted that this variant was not present within the patient or either parent.

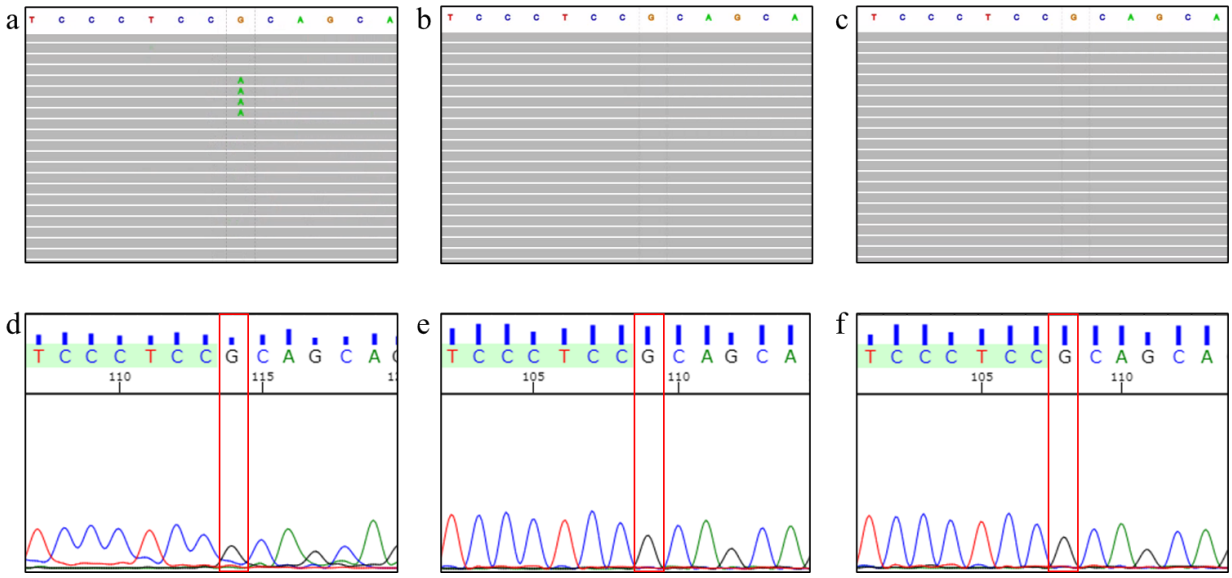


Figure 3.6: Validation of NM_005144:c.1711C>T variant in in sample MI-Proband01968. *a-c)* show BAM file results for proband (a), mother (b) and father (c) from the initial sequencing result. Within the BAM file, it was seen that in 14 % of all reads in the proband, the variant was present. It was also present in 4 % of the maternal reads and absent from the paternal sample. *d-f)* show Sanger Sequencing coverage of the variant region. As the gene is transcribed on the reverse strand, the wild-type allele is G. It can be clearly seen that the variant (G>A) is not present in either the proband (d), the mother (e) or the father (f).

3.5 Discussion and Conclusions

In this Chapter, I have designed and optimised a 2904 smMIP panel for targeted sequencing of 54 genes which are potentially of interest in MI. These genes were highlighted as part of a study by Dr Manon Oud where WES was used to identify potentially causative *de novo* mutations in genes which were likely to influence the patient's fertility when mutated. This preliminary study led to the discovery of 54 potential candidate male infertility genes and founded the work in Chapter 4 of this thesis. The aim was to look for any replications of mutations within these 54 genes in a new cohort of infertile males by using a more targeted approach with the intention to prioritise these genes for further investigation. This was achieved by creating a smMIP assay targeting these 54 genes of interest and using it to investigate *de novo* mutations in a cohort of 75 patient-parent trios after the assay was properly optimised. After multiple rounds of optimisation and then sequencing the infertile trios, it was found that there were no DNMs present in any of the patients within the 54 genes of interest. There were multiple potential reasons for this including sub optimal performance or coverage of the smMIPs as well as too small a gene panel utilised at such an early stage of gene discovery.

The initial run worked well despite one of the samples clearly being of too low quality to successfully utilise the smMIP pool. In theory, with all samples performing optimally, it would be expected that 25 % of the total reads would be assigned to each of the four samples, however we can see that this was not the case in this run with Sample 4 being assigned only 12 % of all the reads. It was decided that the performance issue was likely due to one of two issues. Either an error when pooling all four samples together resulting in a lower sample concentration, or an issue with the quality of the sample DNA itself, resulting in a less successful capturing of the smMIPs. This is potentially represented by significantly decreased number of reads assigned to each smMIP on average to this sample seen in Table 3.1.

Samples 1 - 3 showed a minimum of 76 % and a maximum of 81 % of all smMIPs performing within range without any form of rebalancing or optimisation of the pool. Whilst Sample 3 did show an increased number of smMIPs underperforming in comparison to Samples 1 and 2 in total, the number of smMIPs underperforming and giving zero reads was considerably lower overall, highlighting that many of those giving zero reads in Samples 1 and 2 were likely underperforming in Sample 3. One issue with this run was the relatively large number of smMIPs providing zero reads with 7 % - 12 % fitting this category. With this being the first run, it was highly likely that a number of smMIPs would need rebalancing due to the varying complexity of the genomic regions being targeted. The improvement of these Underperformers

and Zero read smMIPs was the aim of the second rebalancing run, with the intention of creating a shift of smMIPs into the Within Range category.

After rebalancing, the number of underperformers remained largely the same when comparing Samples 1 and 2 in both the initial 1x pool and the 10x pool (Table 3.2), but this could be explained by the fact that the average number of smMIPs performing within range across the two samples has increased by around 100 in each case. This indicated that the 10x pool had caused a shift in the Zero smMIPs to Underperformers, and the Underperformers to Within Range group. This was particularly noticeable when looking at the Zero category for smMIPs in both samples, with Samples 1 and 2 showing only 17 % and 19 % of smMIPs underperforming or giving zero reads in the 10x pool compared to 21 % and 24 % in the initial run. The number of zero reads in Sample 2 were also seen to decrease by almost 50 %. These results were determined to be satisfactory enough to proceed to sequencing of the full 299 patient and parental samples.

This final sequencing run showed yet another improvement with all samples showing greater than 80 % of all smMIPs performing within range, with one sample reaching 92 %. When an analysis was performed to investigate any differences between the DNA origin in regard to the sequencing success, it was found that Saliva and Blood samples performed at similar levels with the only real discrepancy lying between a handful of Nijmegen Blood and Saliva samples which had considerably higher mean reads per smMIP than the majority of the other samples, both Blood and Saliva from both cohorts. It is highly likely that either dilution errors occurred in these samples during the initial sample preparation or pooling stages, or an overamplification has occurred during the PCR stages of the smMIP protocol, leading to a significantly increased number of reads being assigned to these smMIPs. Whilst this is not optimal, it did not have any lasting effect on the variant calling in these samples.

When this study was compared to others using this smMIP assay, it was noted that the assay performed at a highly comparable standard. The work laid out in a previous study utilised a pool of 4,525 smMIPs to sequence a total of 107 causal and candidate genes in 1,119 infertile male patients (Oud et al., 2017). Within this study, 93 % of all smMIPs across the 1,119 samples were seen to be covered by at least 10 unique reads after sequencing. This is only slightly higher than the 92 % achieved within this thesis. Importantly, however, multiple rounds of rebalancing were performed and smMIPs of 1x, 5x, 10x, 20x and unphosphorylated smMIPs were added to the smMIP pool in the work by Oud et al., when compared to one round and either 1x or 10x concentration of smMIPs in this work. The further rebalancing and addition of smMIPs at 5x, 20x would likely have helped to reduce the number of smMIPs achieving zero reads during

sequencing as well as those found to underperform, however, as the percentage of underperformers and the number of smMIPs providing less than 10 reads was relatively small, it was unnecessary to perform these additional rebalancing.

Another study utilising a smMIP pool to investigate mutations in seven candidate genes for Developmental and Epileptic Encephalopathies assessed the performance of their smMIP pool based on the percentage of target bases covered at ≥10 reads within each sample (Hamdan et al., 2017). They noted that in 476 (80 %) of their samples, 90 % of the target bases were covered by at least 10 reads with the other 119 (20 %) of their samples obtaining a reduced coverage with only 70 % of their smMIPs achieving at least 10 reads. In the 299 samples described in this chapter, a similar statistic was seen with 227 (76 %) samples showing 90 % of their smMIPs with ≥10 reads; however, a significant improvement is then seen when compared to Hamdan et al. with 289 of the 299 samples having reached a coverage of 10 or more reads in 85 % of all smMIPs. The two worst performing samples within the cohort still achieved 78 % of all smMIPs covered at a read depth of 10x or greater showing the overall success of the smMIP assay for sequencing regions of interest whilst also highlighting how, with further rebalancing steps, the panel could be even further improved upon.

After analysis of the sequencing data 13 rare (AF <0.1 % in the global population) *de novo* variants were prioritized in 12 different patients. Unfortunately, after performing further Sanger Sequencing validations of these variants, all 13 were found to be false positives. Sanger Sequencing showed that either the variant was not real, or it was in fact inherited from one of the parents. Of the 13 variants found to be false positive, four were proven to be false positives after Sanger Sequencing showed an absence of the variant in both the probands and the parents. Typically, for a germline *de novo* mutation, a MAF of 50 % would be expected. However, the minimum filters were set to 10 % in the processing stages of the variants in order to not exclude any potential mosaic mutations and to not limit the investigation of the variants too stringently before a manual inspection. After Sanger Sequencing was performed, it was clear that these four mutations were likely sequencing or alignment errors which have occurred during the preparation or sequencing protocol.

Of the nine variants which were shown to be inherited after Sanger sequencing, it can be seen that five had an MAF of 5 % – 9 % in the affected parent. Due to the inclusion of only variants with MAF of 10 % or higher, these variants were not identified in the parents and thus when the proband and parental files were cross examined, appeared *de novo*. For those variants where low coverage is not the issue, it is possible that a bias was introduced during the PCR stage for reads not containing the variant. The cause of this would likely be due to poor sequencing

quality in these regions, with a number being covered only by one smMIP or the variant lying closely to the end of a read.

Whilst no pathogenic rare *de novo* mutations were highlighted in these genes in this 75-trio cohort, it must be remembered that the human genome contains over 20,000 genes (Pertea et al., 2018) and this smMIP panel has only investigated 54 of those, all of which had been found to have a potentially causative *de novo* variant in only one other patient in our prior male infertility WES cohort. It is not surprising that in a relatively small sample size no additional *de novo* mutations were identified. In fact, when performing WES on these samples later in Chapter 4, as well as additional samples, it was confirmed that no *de novo* mutations were found in these genes. This does not, however, mean that these genes are now unlikely to be MI genes. This work highlights the necessity for significantly larger cohorts when performing more targeted research. The majority of the genes involved in this smMIP panel are still considered to be of interest and require further investigation. One example of this is the gene *U2AF2* which is noted to be a gene associated with pre-mRNA splicing whereby both the RNA and protein are highly expressed within the testis and present during spermatogenesis (Glasser et al., 2022). This gene still has potential to become a candidate male infertility gene, however, more work is yet to be done to prove this. It has been well documented that in complex heterogenic diseases such as male infertility as well as Intellectual Disability an exceptionally large sample size is required in order to accurately and confidently identify disease genes (Kaplanis et al., 2020).

Our study indicates that smMIP based targeted sequencing may not be the best technique to be used at this stage in the discovery of MI genes whilst there are only single patients with mutations in all the potential candidate genes. The panel provides a very narrowed approach and whilst more genes can indeed be added to the smMIP pools over time, it is better suited to later stages of disease gene research and in genetic diagnostics rather than replication studies as applied in this chapter. If there is a well-established causative gene (or set of genes) which may often be mutated in a certain disease, the smMIP panel is an excellent, cost-effective, and time-conserving method to achieve high quality results. If the smMIP panel is being used as a diagnostic tool rather than an exploratory tool, more rounds of rebalancing could be performed on the pool itself to ensure that all smMIPs are performing at their optimal, and that all regions of the evaluated genes are fully covered. An automated variant calling, and filtering pipeline could then easily be employed in order to streamline interpretation and reporting. This could be of great use in future when a larger number of male infertility genes have been confidently identified in order to aid with investigations in a clinical setting.

One example of the clinical application of a smMIP assay is seen in (Gallon et al., 2020) where a smMIP panel was designed to target specific microsatellites within colorectal cancer samples to determine whether Lynch Syndrome is present. The smMIP panel may be used in this specific clinical setting to help predict the patient's response to particular types of immunotherapies in their cancer treatments. Several smMIP-based panels have also been shown to greatly increase turnaround time in identifying mutations in the *BRCA1* and *BRCA2* genes where, in a 2017 study, a smMIP panel was shown to yield a 100 % analytical sensitivity and specificity in identifying true positive variants in 152 samples whilst decreasing the turnaround time of testing by 60 % (Neveling et al., 2017). Both studies highlight how smMIP based panels allow for mass testing of multiple samples and demonstrate how the panels may be optimally used in a clinical diagnostic setting, where multiple patients will be tested regularly for mutations in a set of known disease-causing genes.

From this work, it was clear that further studies were required to identify and consolidate the evidence for more candidate MI genes using a larger cohort of patient-parent trios, before utilising a highly targeted method such as the smMIPs again. This led onto the work in Chapter 4 where WES was utilised in a larger cohort of patient-parent trios for disease gene discovery.

Chapter 4: Identification of Putative Pathogenic *de novo* Mutations in Severe Forms of Male Infertility

Preface: The work being presented in this Chapter is an adaptation and expansion on my Co-First Author manuscript "A *de novo* paradigm for male infertility" published in Nature Communications (January 2022). As co-first author and the main writer of the manuscript, I have adapted the text and figures from this paper throughout the entirety of this chapter as well as having expanded on it by adding additional information and analyses. Any results that are not my own have been appropriately acknowledged within each respective section.

4.1 Introduction

Male infertility (MI) contributes to approximately half of all cases of infertility and affects around 7 % of the male population. For the majority of these men the cause of their infertility remains unexplained (Krausz & Antoni Riera-Escamilla, 2018). Despite a clear role for genetic causes in MI, there is a distinct lack of diagnostically relevant genes and at least 40 % of all cases are classified as idiopathic (Krausz & Antoni Riera-Escamilla, 2018; Kasak & Laan, 2021; Oud et al., 2019; Tüttelmann et al., 2018). Severe Oligozoospermia (<5 million sperm per ml ejaculate) and Azoospermia (complete lack of sperm in the ejaculate) are the most severe forms of qualitative MI with the more severe Azoospermia being diagnosed in 10 % - 20 % of all MI cases (Tüttelmann et al., 2011; Choy & Amory, 2020). It has been shown that men with this most severe case of infertility are at the highest risk of being carriers of genetic abnormalities (25 % in Azoospermia cases) and this risk progressively decreases as the severity of the infertility decreases. With over 2000 genes known to be involved in spermatogenesis and the complex multifactorial pathology of MI itself, the number of genetic mutations being identified across all subcategories of the condition are steadily increasing (Kasak & Laan, 2021; Houston et al., 2021).

4.1.1 De novo Mutations as a Cause for Male Infertility

With such a high percentage of all MI cases remaining idiopathic, the need for further research into the genetic causes of this condition is significant. Previous studies in other conditions with reproductive lethality, such as neurodevelopmental disorders, have demonstrated a significant role for *de novo* mutations (DNMs) in their aetiology (Veltman & Brunner, 2012). It is suggested that due to the spontaneous nature of these mutations, they are likely more deleterious and thus provide an explanation for severe forms of MI, which cannot logically have been inherited and are associated with reduced fitness.

In line with this, it is well known that recurrent *de novo* chromosomal abnormalities play a significant role in MI (Xavier et al., 2021). Both azoospermia factor (AZF) microdeletions on the Y chromosome as well as an additional X chromosome, resulting in Klinefelter syndrome, occur *de novo*. Collectively, these *de novo* events explain up to 25 % of all cases of non-obstructive azoospermia (NOA) and are commonly screened for during diagnostic tests in infertile males (Kasak & Laan, 2021; Krausz & Riera-Escamilla, 2018). Interestingly, in 1999, the first instance of a potentially damaging *de novo* point mutation in the Y chromosomal gene *USP9Y* was reported in a man with azoospermia (Sun et al., 1999). It has now been shown that deletion of this gene does not lead to infertility and the gene is more likely to be involved in improving the efficiency of spermatogenesis (Jedidi et al., 2019; Krausz et al., 2006; Luddi et al., 2009). Until now, a large-scale systematic analysis of the role of DNMs in MI had not been attempted. This is partly explained by a lack of basic research in male reproductive health in general (De Jonge & Barratt, 2019; Kasak & Laan, 2021), but also by the practical challenges of collecting parental samples for this disorder, which is typically diagnosed in adults.

Recently, a pilot exome sequencing study in 13 infertile men and their parents published (Hodžić et al., 2021). Within this study, five genes were highlighted as potential candidate genes for MI: *NEURL4*, *BRD2*, *SEMA5A*, *CD1D*, and *CD63*. This study showed a promising start to research into this mode of inheritance for idiopathic MI, however the work was limited by its small sample size and thus lack of replications, highlighting the need for further studies in larger cohorts.

4.1.2 Autosomal dominant male infertility disease genes

In recent decades, the majority of research into MI has been focussed on recessive or X and Y linked causes of MI with all common screening methods identifying these types of mutations. Currently, only four autosomal dominant (AD) genes have been confidently linked to isolated male infertility in humans; *DMRT1*, *KLHL10*, *SYCP2* and *SYCP3* (Houston et al., 2021; Oud et al., 2019; Schilit et al., 2020). *Doublesex- and mab-3-related transcription factor 1* (*DMRT1*) is a transcription factor which plays roles in male sex determination and differentiation by controlling male germ cell proliferation. Genes *SYCP2* and *SYCP3* (Synaptonemal Complex Protein 2 and 3) both produce proteins which are heavily involved in the synaptonemal complexes during meiotic prophase, with affected patients showing a variety of male infertility phenotypes including severe oligozoospermia, cryptozoospermia and meiotic arrest. *Kelch like family member 10* (*KLHL10*) is thought to be involved in the process of ubiquitination and proteasomal degradation of proteins during spermatogenesis (Yatsenko et al., 2006). Whilst

these genes have been determined to play a role in AD-related male infertility, no parental samples were investigated of patients with mutations in these genes and thus it remains unknown whether any of the variants were maternally inherited or occurred *de novo*. The pathways and mechanisms that these four genes act within differ functionally yet ultimately all produce an infertile phenotype in the affected males, highlighting the heterogeneity of severe male infertility and emphasising the need for large scale studies into disease gene identification in this disorder.

4.1.3 De novo Mutation Detection in Whole Exome Sequencing Data of 185 Patients

The work in the previous Chapter of this thesis focussed on identifying *de novo* mutations in a cohort of infertile patients using a targeted sequencing approach. The targeted smMIP sequencing technique has the ability to sequence specific regions of DNA extracted from both blood and saliva samples. This method often achieves a high read depth, with our own study showing a mean of 240 reads per smMIPs covering each region (Khan et al., 2019). A high read depth allows for greater detection of low frequency variation and, alongside the potential for massively parallel sequencing of samples per run, makes the smMIP targeted gene panel an attractive method for investigating variation in known disease-causing genes (Hiatt et al., 2013). When used in a systematic approach to sequence large numbers of samples, this method boasts an approximate cost of $1 per gene per sample despite the initial high start-up costs associated with designing a smMIP panel (O'Roak, Vives, Fu, et al., 2012). Additionally, the ability to add or remove genes from the smMIP pool gives it great flexibility. This Chapter moves on to make use of Whole Exome Sequencing (WES) to investigate mutations across the entirety of the coding portion of our infertile patients' genomes. The human exome accounts for approximately 1% of the human genome and is known to encompass 85 % of all known disease-related variation (van Dijk et al., 2014). The initial stage of WES differs from smMIP targeted sequencing as a commercially available target set is ready to be used, eliminating the need to design, pool and balance probe sets before patient samples can be sequenced. Both methods require sample preparation to capture the specific regions (either targeted genes or the whole exome) being sequenced. Downstream processing of samples for both smMIPs and WES are compatible with identical alignment and annotation tools utilised in both instances. The major difference at this stage is the substantially larger amount of data produced for processing and analysis per patient after WES. Due to the considerable amount of data produced in a single exome sequencing experiment, fewer samples can be sequenced in a single run and the overall read depth is also reduced. The increased number of runs as well as a need for larger data storage

and processing capabilities all contribute to higher cost per patient for exome sequencing in comparison to targeted sequencing. For disorders such as male infertility where there are very few associated disease genes and large numbers of idiopathic patients, these costs are arguably justified by the discovery potential unlocked when assessing a patient's entire exome.

Discussed here is the first large scale study on the role of DNMs in severe male infertility aiming to address the lack of knowledge in this subject. A total of 185 patients suffering from unexplained cases of severe oligozoospermia (n = 74) and azoospermia (n = 111), with both the patients and their parents having underwent WES in order to identify genes containing DNMs which could potentially explain the infertility observed in these patients.

4.2 Aims

1. To analyse Whole Exome Sequencing data of 185 infertile patient-parent-trios and identify all rare protein altering *de novo* variants.
2. To assess the pathogenicity of all rare coding *de novo* variants identified and assess the likelihood of these DNMs to explain the infertility phenotype observed in the patient.
3. Perform statistical analyses to investigate an enrichment for loss of function (LoF) and missense mutations in LoF and missense intolerant genes.
4. Assess a replication cohort of infertile men (as well as fertile controls) for rare heterozygous mutations in genes found to contain DNMs in the 185 infertile patients.

4.3 Methods

Preparation of the samples for sequencing in this chapter is described in Chapter 2 section 2.3 either in the Netherlands at Radboud University (n = 170) or by Dr Bilal Alobaidi at Newcastle University (n = 15). WES for all samples was conducted at Newcastle University by the Genomics Core Facility, also described in section 2.3 of this thesis. Post Sequencing processing was performed by the BSU at Newcastle University and by Dr Miguel Xavier. Analysis of sequencing data, including variant identification and interpretation, as well as primer design for Sanger validation was performed by myself. Approximately half of the samples (n = 99) were initially processed and had variants identified and validated at Radboud University by Dr Manon Oud, however I then re-processed them all after joint variant calling was performed including the entire cohort of patients and parents to improve the overall detection of variants and specifically of DNMs. Sanger Sequencing validation was performed both at Newcastle University by Dr Bilal Alobaidi and me, and at Radboud University by Petra de Vries.

Preparation of histological samples and figures was conducted by Dr Godfried van der Heijden in Radboud University, information presented previously in Chapter 2. Copy Number Variant (CNV) detection and related figure creation was performed by Dr Kumara Mastrorosa and for the purpose of this Chapter and the aforementioned publication, the genes identified within these *de novo* CNVs were analysed by me. Information regarding the processing of all additional patient and control samples can be found in Chapter 2 of this thesis in section 2.1. Figure 2.3 in Chapter 2 describes the workflow for variant filtration and prioritisation of samples within this chapter.

4.4 Results

4.4.1 Comparison of Whole Exome Sequencing Kits

Whole Exome Sequencing of the 185 trios was performed on the NovaSeq 6000 Sequencing System (Illumina) with the initial 99 trios being sequenced using Illumina's Nextera DNA Exome Capture kit and the subsequent 86 using the Twist Bioscience's Human Core Exome Kit. Comparable results were achieved with both kits similarly covering more than 99.4 % of all exon regions according to RefSeq, CCDS, Ensembl and GENCODE databases (Table 4.1). Due to the greater target size covered using the Illumina Nextera kit (45,326,818bp compared to 36,539,805bp for the Twist Bioscience kit), a considerable difference was seen between the overall target coverage of samples prepared with the two different kits, with samples prepared using the Illumina Nextera DNA Exome Capture kit reaching an average read depth of 72x across the entire exome sequenced and samples prepared with the Twist Bioscience's Human Core kit achieving a greater average depth per sample of 99X (Figure 4.1, Table 4.2). When comparing the saliva vs blood samples, a small but not statistically significant difference was detected between the coverage of the blood vs saliva samples from either of the two kits. The blood samples performed slightly better than the saliva samples in both cases with a coverage of 79X in the Illumina Probands vs 74X and 73X in the Illumina parental saliva samples and 105X in the Twist probands vs 96X and 91X in the Twist parental saliva samples. Despite the slightly increased coverage in blood samples the number of SNPs called in each different DNA extraction group was not affected and was in fact highly comparable (n = 99,821 in Twist blood samples vs n = 99,861 and n = 99,170 in Twist saliva, n = 101,102 in Illumina Blood vs n = 100,850 and n = 100,749 in Illumina saliva samples).

	Illumina's Nextera DNA Exome Capture kit		*Twist Bioscience's Twist Human Core Exome Kit*	
	Mean	Coefficient of Variation	Mean	Coefficient of Variation
Total Reads (M)	77.4	0.19	72.3	0.23
Average Read length (bp)	97.0	0.01	98.9	0.01
Uniquely mapped reads (M)	49.8	0.18	67.1	0.23
Bases mapped to target (Mb)	4648.9	0.19	4755.8	0.22
Bases mapped to target ± 150bp (Mb)	7550.1	0.19	7281.7	0.23
Target size (bp)	45,326,818	-	36,539,805	-
Exonic content covered	99.4 %	-	99.5 %	-

Table 4.1: Statistics of reads and alignment to reference genome per exome kit. Abbreviations: M = Million; bp = base pairs; Mb = Megabases. Exonic content covered refers to RefSeq, CCDS, Ensembl, GENCODE databases.

Exome Kit	Type	Mean Sample Coverage	SNPs	Indels	Het/Hom Ratio	Insertion/Deletion Ratio	ti/TV Ratio
Illumina's Nextera DNA Exome Capture kit	Probands (Blood)	79X	101,102	11,274	1.6	0.85	2.25
	Mothers (Saliva)	74X	100,850	11,320	1.6	0.85	2.25
	Fathers (Saliva)	73X	100,749	11,318	1.59	0.85	2.25
Twist Bioscience's Twist Human Core Exome Kit	Probands (Blood)	150X	99,821	11,785	1.72	0.84	2.24
	Mothers (Saliva)	96X	99,861	11,682	1.73	0.85	2.24
	Fathers (Saliva)	91X	99,170	11,698	1.72	0.85	2.24

Table 4.2: Comparison of variants called for the different sample types per exome kit.

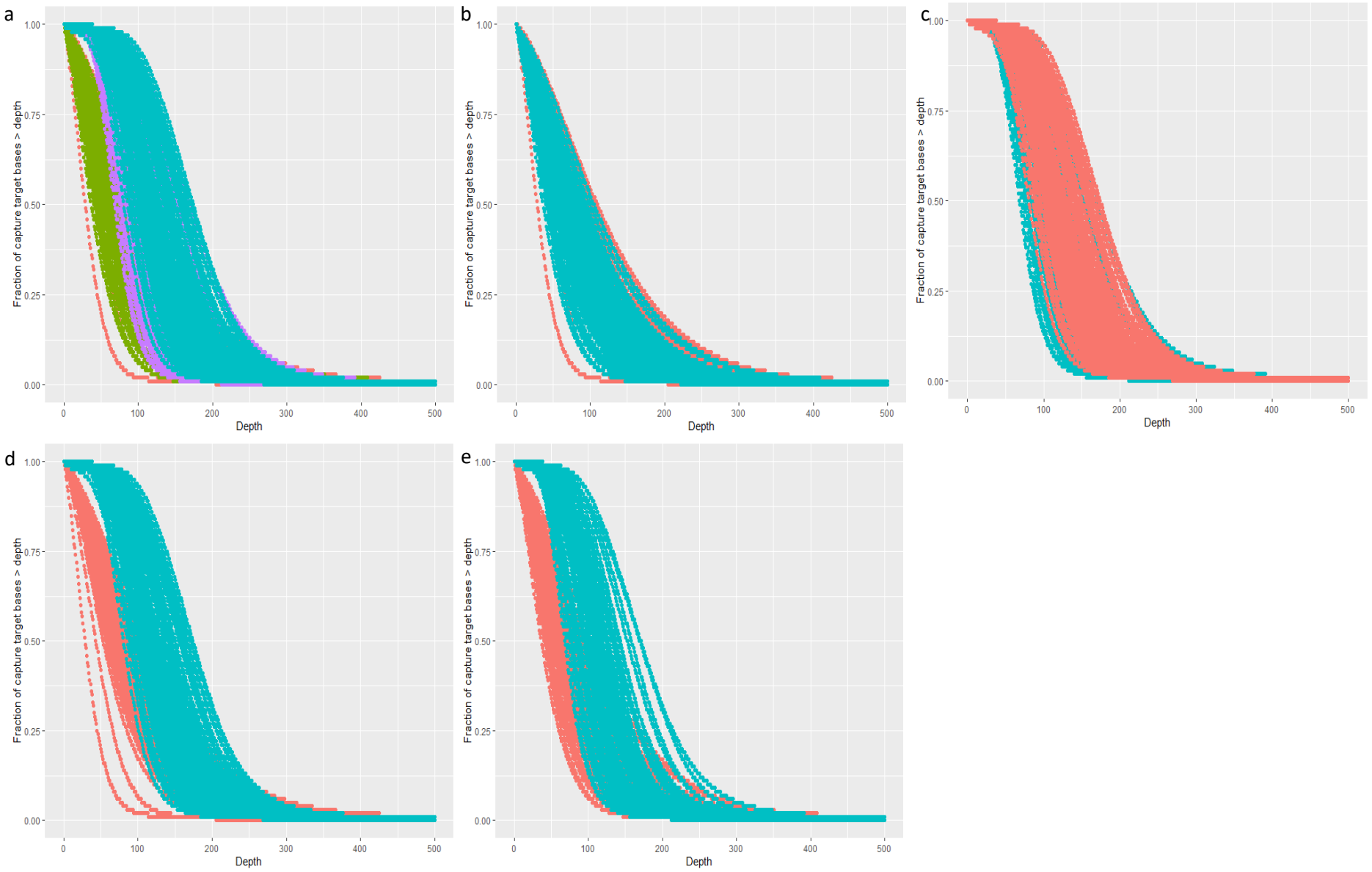

a
b
c
d
e
Fraction of capture target bases > depth
Depth
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
500

Figure 4.1: Target coverage on all 552 samples included in experiment. *a)* Fraction of capture target bases per depth for all samples by cell type of origin and exome kit used. Legend: Red dots - Illumina Blood samples; Green dots - Illumina Saliva Samples; Purple dots - Twist Saliva Samples; Blue dots - Twist Blood Samples; *b)* Fraction of capture target bases per depth of samples sequenced with Illumina's Nextera DNA Exome Capture kit. Legend: Red dots - Illumina Blood samples; Green dots - Illumina Saliva Samples; *c)* Fraction of capture target bases per depth of samples sequenced with Twist Bioscience's Twist Human Core Exome Kit. Legend: Blue dots - Twits Saliva Samples; Red dots - Twist Blood Samples; *d)* Fraction of capture target bases per depth of samples with DNA extracted from Blood cells for both exome kits. Legend: Red dots - Illumina Blood samples; Blue dots - Twist Blood Samples; *e)* Fraction of capture target bases per depth of samples extracted from saliva cells for both exome kits. Legend: Red dots - Illumina Saliva Samples; Blue dots - Twist Saliva Samples.

4.4.2 Identification of Rare Potentially Pathogenic de novo Mutations

After WES was performed, patient annotated VCF files were analysed to filter down all DNMs. I only included DNMs which were covered by at least 10 sequencing reads with at least 15 % of all sequencing reads containing the variant, found in protein coding regions of the exome and either rare (present in less than 0.1 % of the population in gnomAD) or previously not reported at all. After applying these filters, a total of 192 rare DNMs were identified in 185 patients at a rate of 1.04 coding DNMs per patient. These 192 variants were then further filtered down to remove any non-protein altering DNMs to give a total of 145 rare protein altering DNMs (Figure 2.3). The distribution of these DNMs within the 185 unexplained cases of both azoospermia and severe oligozoospermia can be seen in Figure 4.2. A total of 76 patients were found to have no rare coding DNMs at all with this number increasing to 90 when non-protein altering DNMs were removed. A further 55 patients (n = 58 when looking only at protein altering DNMs) presented with one DNM and 54 patients (n = 37 when protein altering) had two or more DNMs. One patient in the cohort was found to have four rare protein altering DNMs and an additional patient presented with five rare protein altering DNMs.

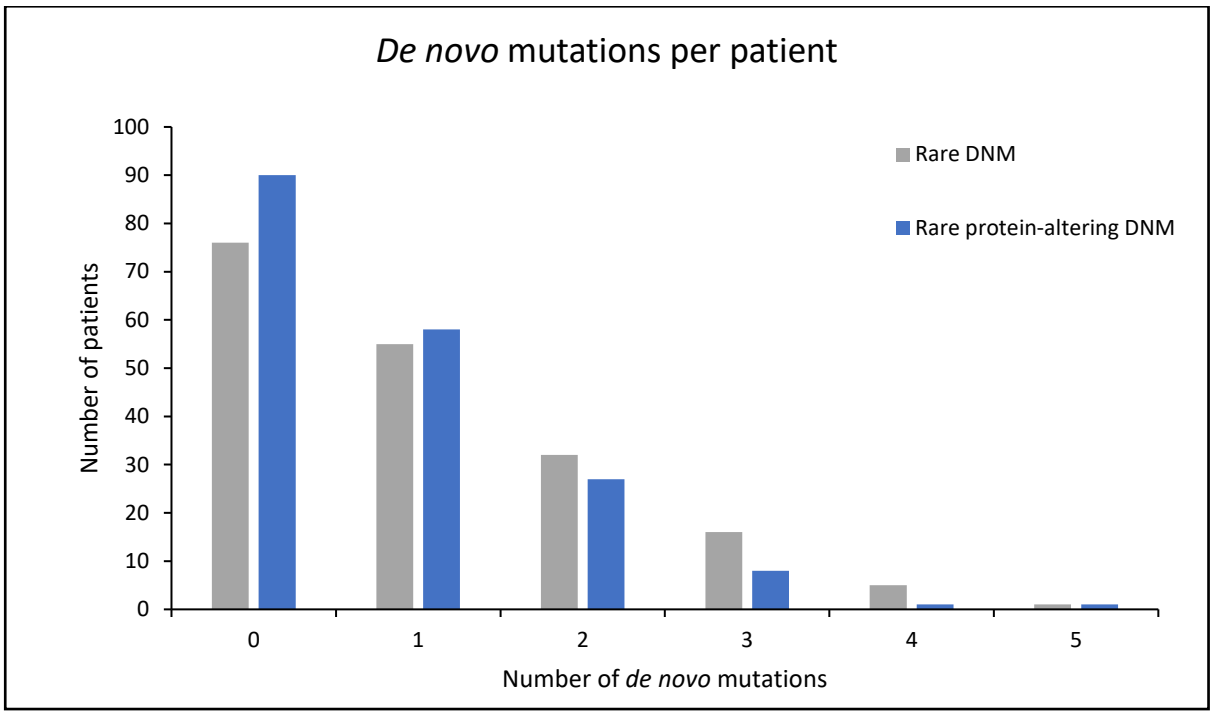


Figure 4.2: Distribution of 192 *de novo* mutations in 185 patients.

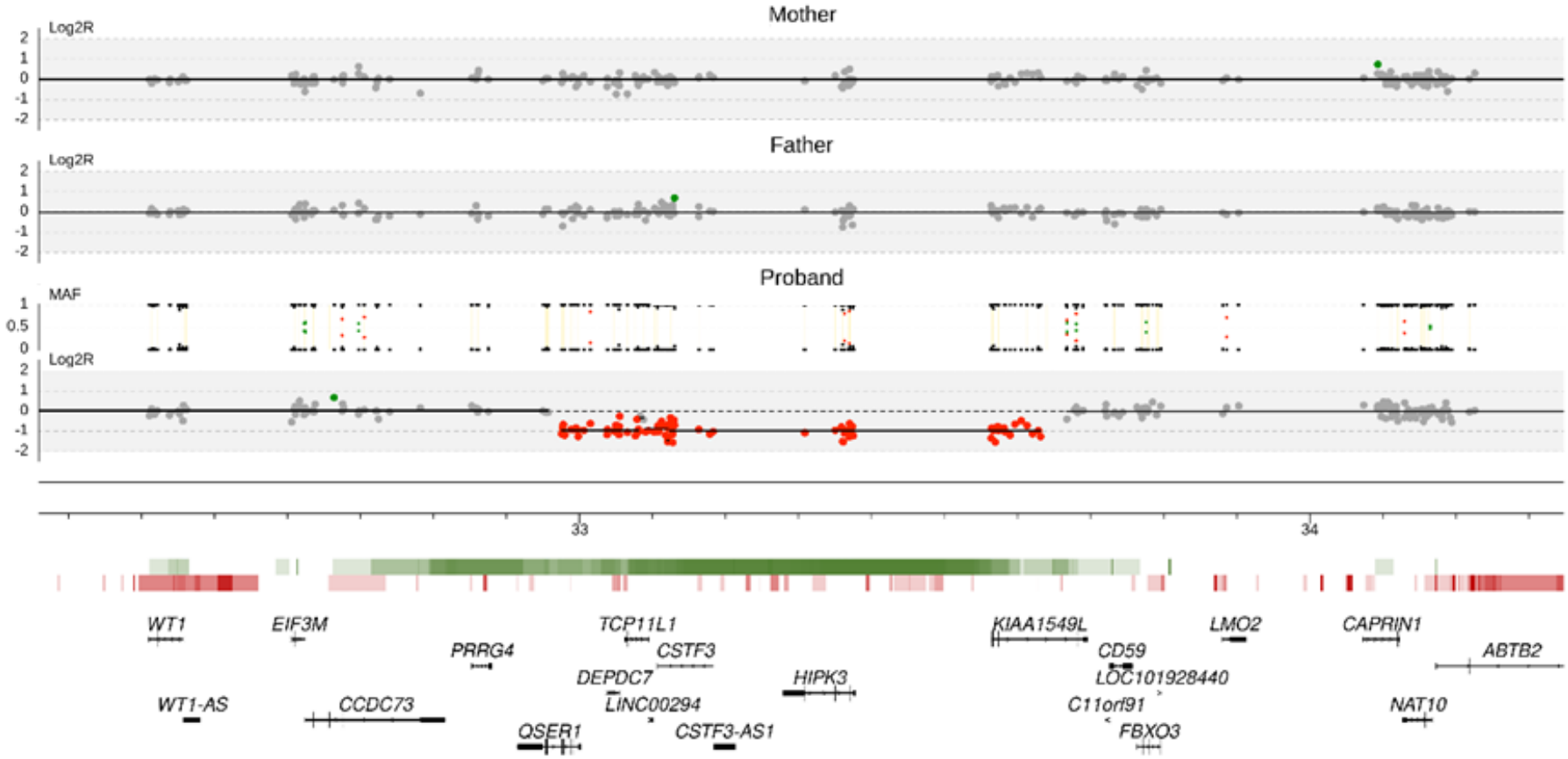
All of these *de novo* point mutations were autosomal apart from a singular DNM occurring on chromosome X, and all 145 rare, protein altering DNMs exist within different genes. At initial glance, none of these 145 DNMs occurred in a gene already known for its involvement in autosomal dominant human male infertility.

To systematically evaluate and predict the likelihood of these DNMs causing male infertility and identify novel candidate disease genes, the predicted pathogenicity of all DNMs was assessed using three prediction tools: Sorting Intolerant From Tolerant (SIFT) (Vaser et al., 2016), MutationTaster (Schwarz et al., 2010), and PolyPhen2 (Adzhubei et al., 2010). To define a variant as pathogenic, a minimum of two out of three of these prediction tools were required to classify the DNM as damaging. Using this approach, 84 of 145 rare protein altering DNMs were predicted to be pathogenic, while the remaining 61 were predicted to be benign.

Whilst my own work focused on identifying *de novo* point mutations present in these patients, two rare *de novo* CNVs were also identified in the same WES data by Dr Kumara Mastrorosa (Figure 4.3). The seven genes affected by these two rare *de novo* CNVs were further investigated within this Chapter and the findings are presented in Table 4.7 along with all other DNMs found in this cohort.

After the initial identification of the 145 rare, protein altering DNMs, a series of unbiased analyses were performed to investigate the variants and genes affected.

a


Mother
Log2R
Father
Log2R
Proband
MAF
Log2R
33
34
WT1
EIF3M
PRRG4
TCP11L1
CSTF3
DEPDC7
HIPK3
LINC00294
CSTF3-AS1
QSER1
KIAA1549L
CD59
LOC101928440
C11orf91
FBXO3
LMO2
CAPRIN1
ABTB2
NAT10
WT1-AS
CCDC73

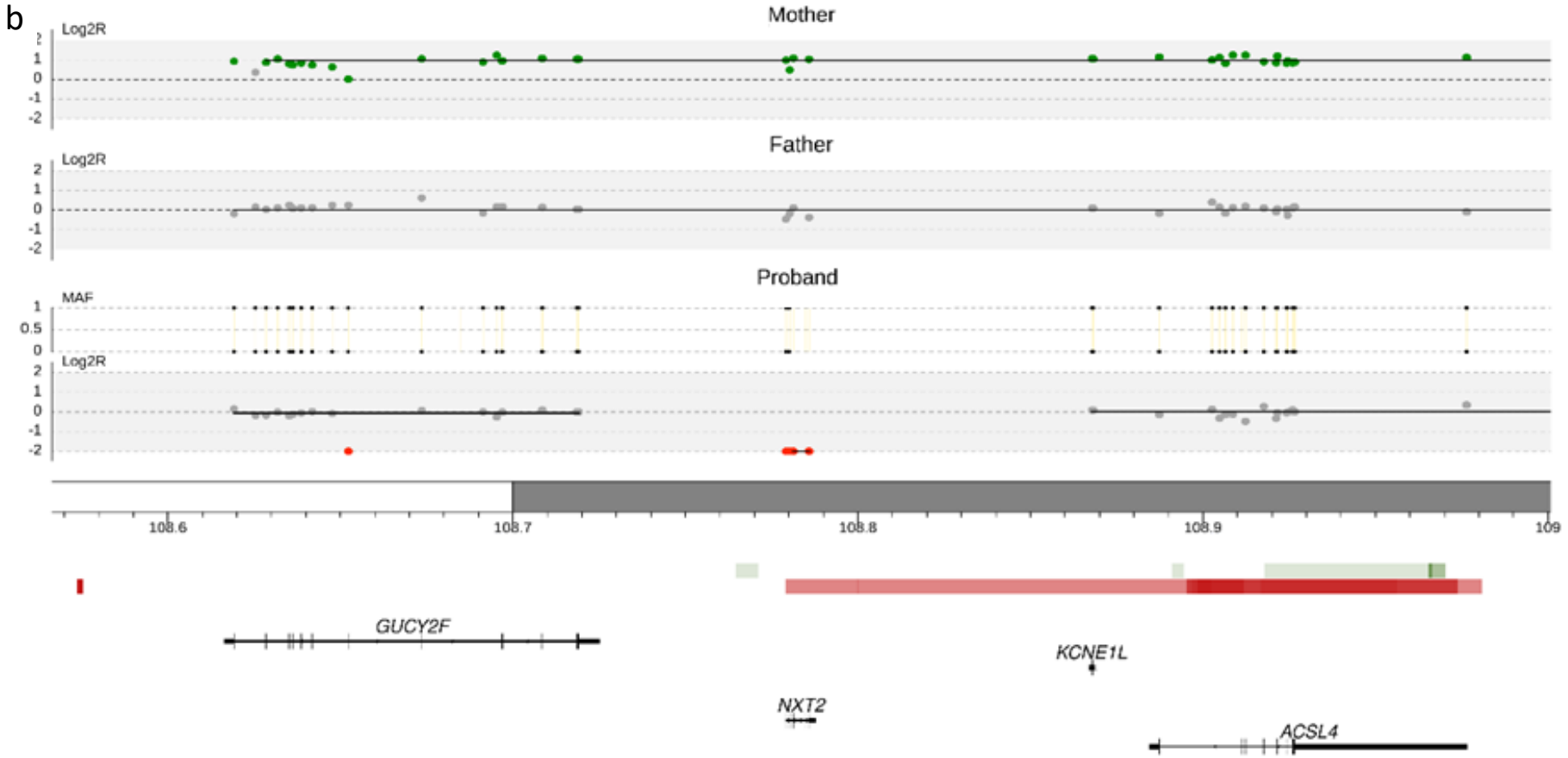

b
Mother
Log2R
Father
Log2R
Proband
MAF
Log2R
108.6
108.7
108.8
108.9
109
GUCY2F
KCNE1L
NXT2
ACSL4

Figure 4.3: *De novo* Copy Number Variations (CNVs) identified in infertile men. *A) De novo* deletion of ca. 656kb (chr11:32975325-33631588) identified in patient Proband_066 affecting genes *QSER1* (partially), *DEPDC7*, *TCP11L1*, *LINC00294*, *CSTF3*, *HIPK3* and *KIAA1549L* (partially). *B)* Deletion of ca. 6kb (chrX:108779109-108785919) identified in Proband_039 affecting gene *NXT2*. Log2Ratio tracks show the number of alleles in the specific region of chromosome in mother, father and proband, which also include the Minor Allele Frequency (MAF) plot showing SNP zygosity. Below the Log2R plots are the cytobands of the specific region of the chromosome. Horizontal red and green bars indicate respectively the deletions and duplications in the Database of Genomic Variants (DGV). The more intense the colour, the more CNVs are present in the region. Below the DGV track, RefSeq genes are indicated. (Oud et al., 2022)

4.4.3 Loss-of-Function de novo Mutations in Loss-of-Function Intolerant Genes

Broadly speaking, across genetic disorders, dominantly acting disease genes are usually intolerant to Loss-of-Function (LoF) mutations, as represented by a high Loss Intolerance probability (pLI) score (Lek et al., 2016) or a low Loss-of-function Observed over Expected upper bound Fraction (LOEUF) score (Karczewski et al., 2020). Prior to any further filtration or investigations into gene functions, these two scoring methods were applied to all genes containing LoF *de novo* mutations (n = 17). In the cohort of infertile men, a significant enrichment was detected in the number of LoF-intolerant genes with a LoF DNM with no such enrichment being identified in a cohort of 1,941 control cases from denovo-db v1.6.11 (http://denovo-db.gs.washington.edu) (median pLI in patients with male infertility = 0.80, median pLI in controls = 3.75 × 10−5, p value = 1.00 × 10−5, N simulations = 100,000) (Figure 4.4a). Comparable results were obtained using the LOEUF scores (median LOEUF in patients with male infertility = 0.34, median LOEUF in controls = 0.59, p value = 1.00 × 10−5, N simulations = 100,000) (Figure 4.4b). A list of all genes containing rare protein altering DNMs can be found in Table 4.3.

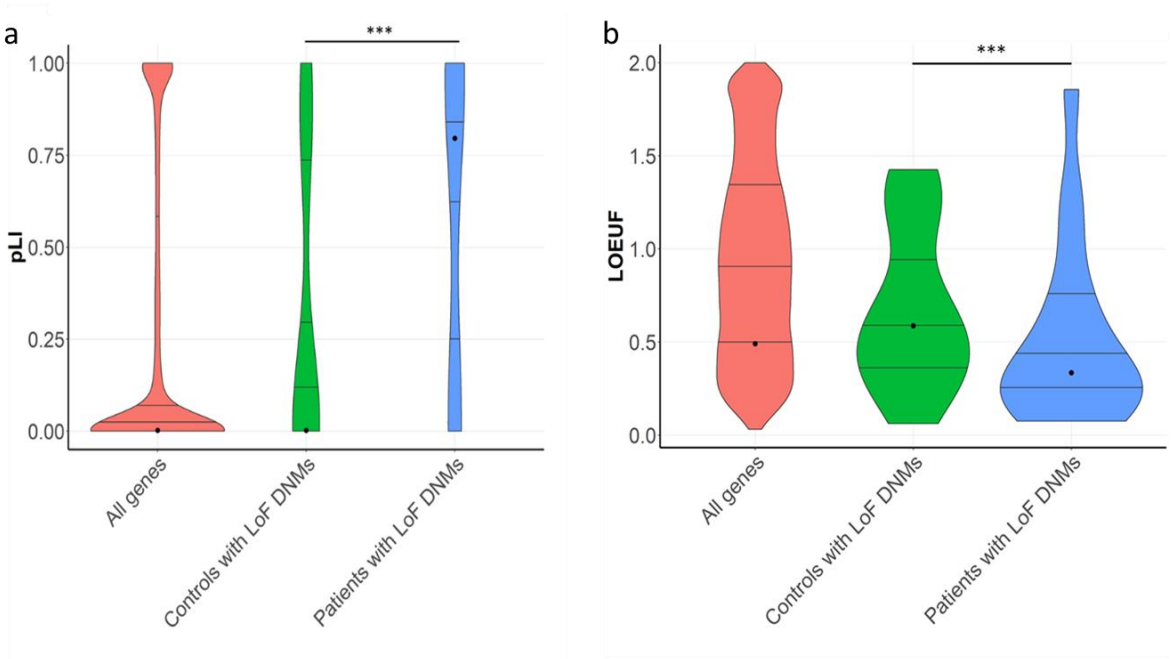


Figure 4.4: Analysis of the intolerance to LoF variants for DNM genes. *A)* Violin plot with quantile lines showing pLI scores in all genes in gnomAD (red), all genes affected by rare protein altering loss-of-function (LoF) *de novo* mutations (DNMs) in a control population (http://*de novo*-db.gs.washington.edu/*de novo*-db/) (green) and in all genes with a rare protein altering LoF DNM in our trio cohort (blue). Using the permutation-based, nonparametric test defined by Lelieveld et al. 64 a significant enrichment of LoF DNMs in LoF-intolerant genes in patient cohort was detected in comparison to the number of LoF in fertile control cohort (DNM LoF mutations in patients n = 17, median pLI in patients with male infertility = 0.80, DNM LoF mutations in controls n = 21, median pLI in controls = $3.75 \times 10-5$, p value = $1.00 \times 10-5$, N simulations = 100,000). The black dot indicates median pLI scores. *b)* Violin plot representing the distribution of the LOEUF scores of all genes in gnomAD, all genes affected by rare protein altering LoF DNMs in a control population (http://denovo-db.gs.washington.edu/denovo-db/) and in all rare protein altering LoF DNM in our trio cohort. The observed median LOEUF score is displayed for each category as a black circle. The lower the LOEUF score, the more intolerant to LoF variation a gene is. This analysis shows a significance enrichment for LoF DNMs in LoF-intolerant genes when compared to a fertile control cohort (p-value = $1.00x10^{-5}$).

Gene	Variant	Consequence	pLI	LOEUF	Conclusion
GREB1L	chr18:19019514-AAGGGC-A	Frameshift Variant	1	0.07	Possibly Causative
ATP1A1	chr1:116930014-CT-C	Frameshift Variant	1	0.12	Unclear
ZFHX4	chr8:77763486-CT-C	Frameshift Variant	1	0.14	Unclear
HTT	chr4:3213834-C-CG	Frameshift Variant	1	0.18	Possibly Causative
SOGA1	chr20:35438426-G-A	Stop Gained	1	0.19	Unclear
TENM2	chr5:167642269-GC-G	Frameshift Variant	1	0.19	Unclear
PPP1R7	chr2:242099831-CAATAA-C	Frameshift Variant	0.99	0.24	Possibly Causative
FBXO5	chr6:153293449-A-ATCAC	Frameshift Variant	0.97	0.32	Possibly Causative
RASAL2	chr1:178435121-G-T	Stop Gained	0.8	0.34	Possibly Causative
ZNF469	chr16:88494628-T-TC	Frameshift Variant	0.72	0.37	Unclear
KLC1	chr14:104129206-TC-T	Frameshift Variant	0.37	0.41	Unclear
OSBPL3	chr7:24874131-G-A	Stop Gained	0	0.54	Unclear
MSH5	chr6:31721100-CAT-C	Frameshift Variant	0	0.7	Unlikely Causative
EXOSC10	chr1:11136965-A-AT	Frameshift Variant	0	0.74	Unclear
EVC	chr4:5743515-C-T	Stop Gained	0	1.06	Unclear
PCDHB1	chr5:140431405-T-TCCGGG	Frameshift Variant	0	1.31	Unclear
ZNF709	chr19:12575362-A-AC	Frameshift Variant	0	1.86	Unclear

Table 4.3: pLI and LOEUF scores of all 17 genes containing rare protein altering LoF de novo mutations predicted to be pathogenic at a variant level. The pLI and LOUEF scores of a gene represent the likelihood of a gene being intolerant to LoF mutations with a pLI of 1 indicating a gene with the highest level of intolerance to LoF mutations and pLI of 0 indicating a greater tolerance for LoF mutations. The parameters for the LOEUF score differ with a LOEUF score of 0 representing the most LoF intolerant genes and a value of 9 representing the least intolerant.

Due to their potentially deleterious nature, the damaging effects of the two rare *de novo* CNVs were also investigated by looking at the pLI score of the genes involved. Proband_066 presented with a large 656 kb *de novo* deletion on chromosome 11, spanning six genes in total. This deletion partially overlapped with a deletion reported in 2014 in a patient with cryptorchidism and non-obstructive azoospermia (NOA) (Seabra et al., 2014). Two genes affected in both patients, *QSER1* and *CSTF3*, are LoF-intolerant with pLI scores of 1 and 0.98, respectively. In particular, *CSTF3* is highly expressed within the testis and is known to be involved in pre-mRNA 3′-end cleavage and polyadenylation (Grozdanov et al., 2018).

4.4.4 Missense Mutations in Missense Intolerant Genes

To further analyse the impact of the variants on the genes affected, the missense Z-score of all 122 genes affected by a missense DNM was investigated. This score indicates the tolerance of genes to missense mutations based on the number of observed vs expected missense variants within the given gene where a positive score indicates a missense depletion (Samocha et al., 2014; Deák & Cook, 2022). When comparing missense mutations found in the infertile cohort to those in a control cohort, no significant enrichment was found for missense DNMs present in missense-intolerant genes (median Z-score in male infertility patients = 0.83, median Z-score in controls = 1.04, p value = 1, N simulations = 100,000) (Figure 4.5a). Interestingly, however, a significantly higher median missense Z-score was observed in genes affected by a missense DNM predicted as pathogenic using the method described in section 4.4.1 (median Z-score = 1.21, n = 63) when compared to genes affected by predicted benign (median Z-score = 0.98, n = 59) missense DNMs in our cohort (p value = $5.01 \times 10{-}4$, Figure 4.5b). It should be noted that the same analysis in controls showed no such significant difference (p value=0.95, Figure 4.5c).

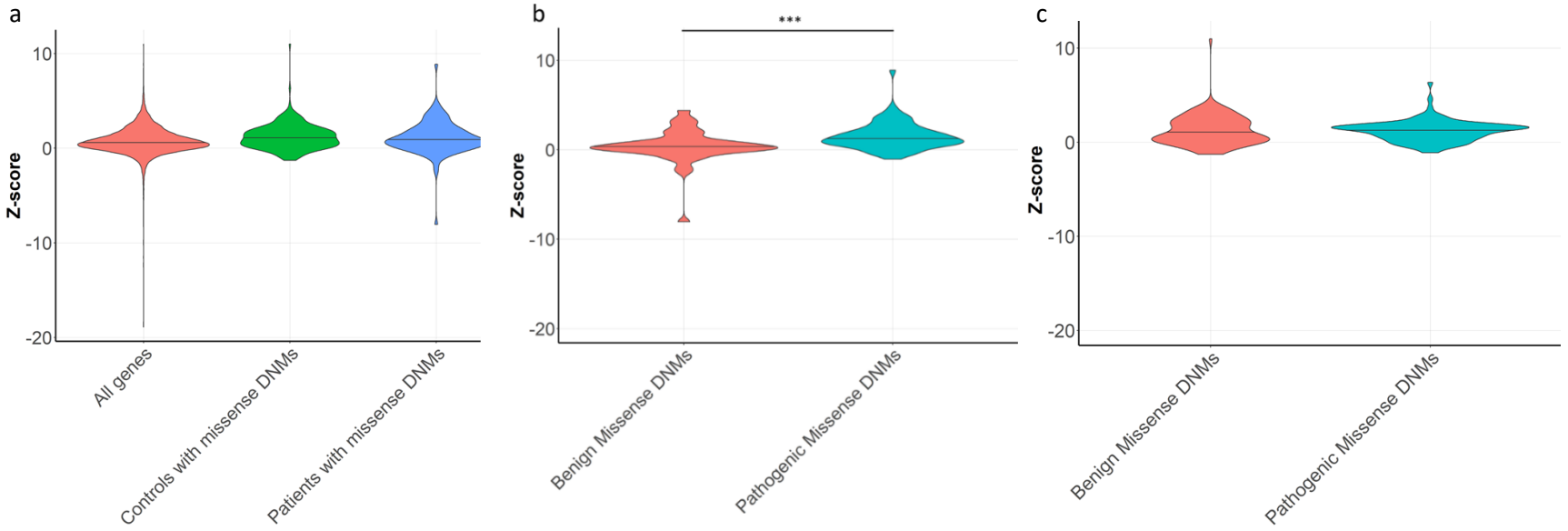
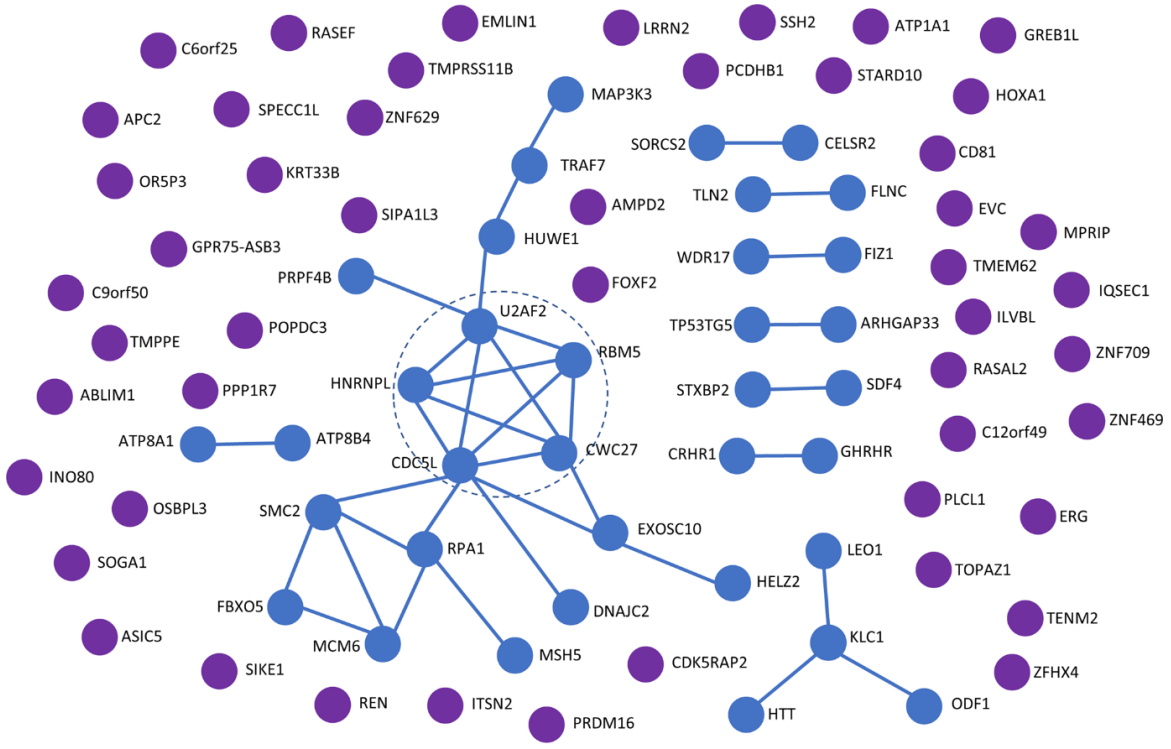


Figure 4.5: Analysis of the intolerance to missense variants for DNM genes. *A)* Violin plots with quantile lines represent the distribution of the z-scores of all genes in gnomAD (red), all genes affected by rare protein altering missense DNMs in this study (blue) and in the previously mentioned control population (green). The higher the Z-score, the more intolerant the gene is to missense variants. Comparison between overall missense DNMs in our study (n = 122) and control population (n = 262) shows no significant difference (Two-sided Mann-Whitney U test, p-value=1 *b)* Violin plot with quantile lines showing the distribution of Z-scores for genes with predicted benign (n = 59) and pathogenic missense DNMs (n = 63) in infertile patients. A significant increase in predicated pathogenic DNMs in missense-intolerant genes was detected compared to benign missense DNM (Two-sided Mann–Whitney U test, p value of $3.44 \times 10{-}4$). *C)* Violin plot with quantile lines showing the distribution of Z-scores of genes containing a rare protein altering missense DNM in the previously mentioned control cohort, where no significant enrichment (Two-sided Mann-Whitney U test, p-value=0.95) can be observed in the median missense-z score of genes containing rare protein altering benign missense DNMs (n=39, red) vs pathogenic missense DNMs (n=68, blue). (***$p < 0.001$).

4.4.5 Protein–Protein Interactions Reveal Link to mRNA Splicing

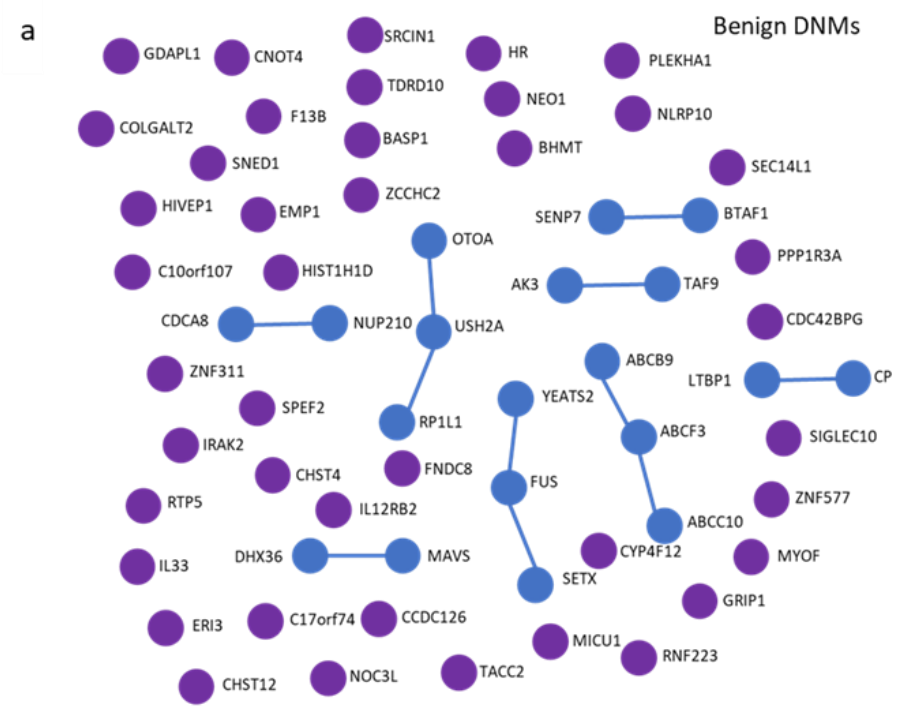
The final investigation into all 145 rare, protein altering DNMs prior to a functional assessment of the genes in relation to male infertility involved a network analysis to identify any recurrent pathways effected by DNMs in the infertile cohort.

An analysis using the Search Tool for the Retrieval of Interacting Genes (STRING) database (Szklarczyk et al., 2017), revealed a significant enrichment of protein interactions amongst the 84 genes affected by a protein altering DNM predicted to be pathogenic (PPI enrichment p value = 2.35 × 10−2, Figure 4.6). No such enrichment was observed for the genes highlighted as likely benign (n = 61, PPI enrichment p value = 0.206) or those affected by synonymous DNMs (n = 35, PPI enrichment p value = 0.992, Figure 4.7a, 4.7b).

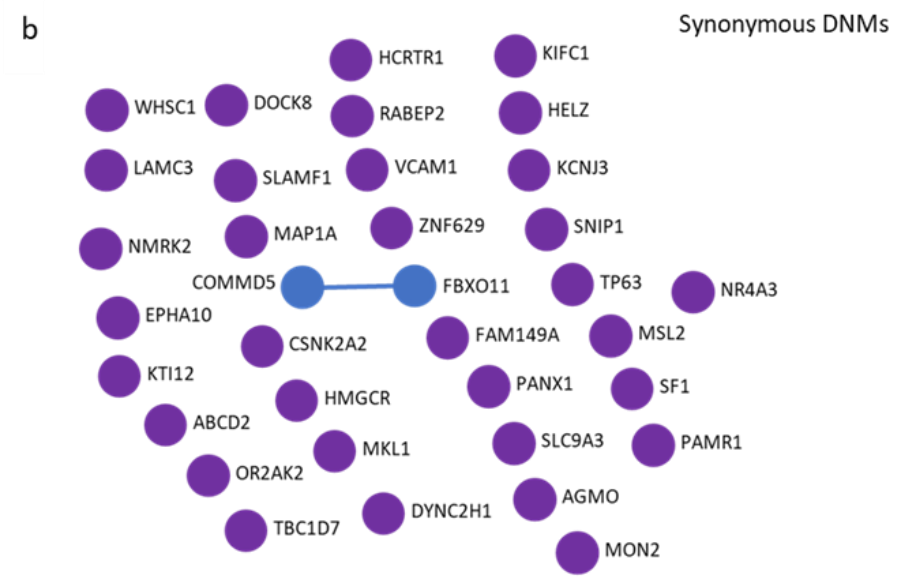


Number of nodes: 84	Average local clustering coefficient: 0.336
Number of edges: 36	Expected number of edges: 25
Average node degree: 0.857	PPI enrichment p-value: 0.0235

Figure 4.6: Protein–protein interactions predicted for proteins affected by pathogenic de novo mutations: Significantly larger number of interactions were observed in proteins affected by *de novo* mutations than expected for a similar sized dataset of randomly selected proteins (PPI enrichment p value = 2.35 × 10−2). The central module of the main interaction network (blue dashed circle) contains five proteins involved in mRNA splicing as well as a secondary branch containing five proteins involved in the mitotic cell cycle (Figure 4.8a).



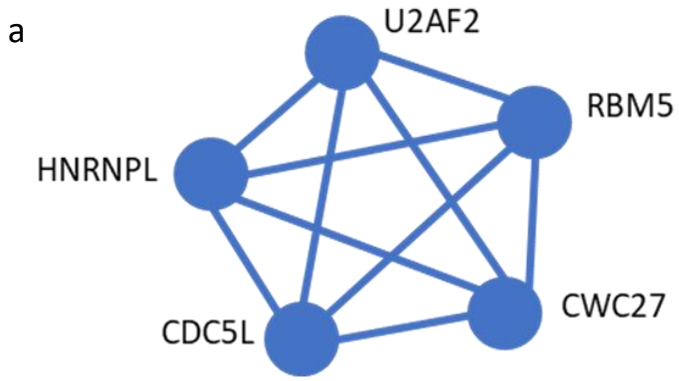
Number of nodes: 61	Average local clustering coefficient: 0.262
Number of edges: 11	Expected number of edges: 8
Average node degree: 0.361	PPI enrichment p-value: 0.206



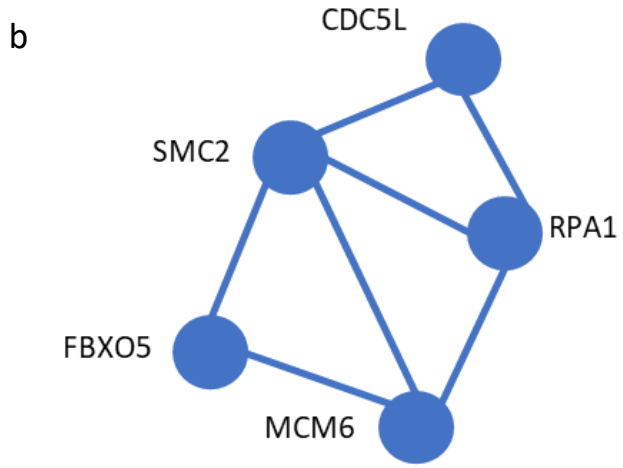
Number of nodes: 35	Average local clustering coefficient: 0.0571
Number of edges: 1	Expected number of edges: 5
Average node degree: 0.0571	PPI enrichment p-value: 0.992

Figure 4.7: Protein-protein interactions between synonymous DNM. *A)* A protein-protein interaction analysis was performed on all protein altering benign DNM (n = 61) which classified as benign in two thirds of the pathogenicity scores. No significant interaction is seen between the different proteins (PPI enrichment p-value = 0.206). *b)* A protein-protein interaction analysis

was performed on all synonymous DNM (n = 36) which are known to not affect the gene. Here no significant interaction is seen with fewer edges than would be expected between this number of genes (PPI enrichment p value 0.992).



Number of nodes: 5	Average local clustering coefficient: 1
Number of edges: 10	Expected number of edges: 1
Average node degree: 4	PPI enrichment p-value: 8.44×10^{-9}



Number of nodes: 5	Average local clustering coefficient: 0.767
Number of edges: 7	Expected number of edges: 1
Average node degree: 2.8	PPI enrichment p-value: 1.56×10^{-5}

Figure 4.8: Functional modules of protein-protein interaction network. *A)* The five proteins: RBM5, HNRNPL, CWC27, CDC5L and U2AF2 all interact highly with one another (PPI enrichment p-value = 8.44×10^{-9}) and are all seen to be involved in the biological process of mRNA splicing, via the spliceosome (False discovery rate: 1.72×10^{-7}). *b)* The five proteins: FBXO5, SMC2, RPA1, CDC5L and MCM6 are all interconnected interacting with at least two other proteins within this module (PPI enrichment p-value = 8.44×10^{-9}) and are all seen to have an involvement with the biological process of the mitotic cell cycle.

The STRING network analysis also highlighted a central module of interconnected proteins (U2AF2, HNRNPL, CDC5L, CWC27, and RBM5) with a significant enrichment of genes required for mRNA splicing (Figure 4.8a). Pre-mRNA splicing allows gene functions to be expanded by creating alternative splice variants of gene products and is highly elaborated within the testis (Song et al., 2020). One of these genes, *RBM5* has been previously highlighted as an essential regulator of haploid male germ-cell pre-mRNA splicing and MI in mice (O'Bryan et al., 2013). Of note, RBM5 is a tumour suppressor in the lung (Jamsai et al., 2017), with reduced expression affecting RNA splicing in patients with non-small cell lung cancer (Liang et al., 2012). HNRNPL is another splicing factor affected by a possible pathogenic DNM in the study. One study implicated a role for *HNRNPL* in patients with Sertoli cell-only phenotype (Li et al., 2012). The remaining three mRNA splicing genes have not yet been implicated in human MI. However, mRNA for all three is expressed at medium to high levels in human germ cells and all are widely expressed during spermatogenesis (Wang et al., 2018). Specifically, CDC5L is a component of the PRP19-CDC5L complex that forms an integral part of the spliceosome and is required for activating pre-mRNA splicing (Ajuh, 2000), as is CWC27 (Brea-Fernández et al., 2019). U2AF2 plays a role in pre-mRNA splicing and 3′-end processing (Millevoi et al., 2006). Interestingly, *CSTF3*, one of the genes affected by a *de novo* CNV in Proband_066, affects the same mRNA pathway (Seabra et al., 2014).

A secondary functional module was also identified including genes *CDC5L*, *SMC2*, *MCM6*, *FBXO5* and *RPA1* which all proved to be involved in the biological process of the mitotic cell cycle. None of these five genes had been previously implicated in MI, however, all the encoded proteins show medium to high levels of expression within the testis and four of the five show RNA present in the testis at medium levels during spermatogenesis. Whilst also playing a role in pre-mRNA splicing, *CDC5L* has been shown to be an essential regulator of mitotic progression with knockdown experiments in tumour cells showing mitotic arrests and chromosomal misalignments (Mu et al., 2014). Structural Maintenance of Chromosomes 2 or *SMC2* encodes a protein which is a principal component of the condensing complex, required for mitotic chromosome condensation (Dávalos et al., 2012). *Minichromosome Maintenance Complex Component 6* (*MCM6*) has been identified as an essential regulator in the DNA replication system as part of the MCM complex (Zeng et al., 2021). The protein encoded by *RPA1* plays an essential role in DNA replication as a part of the RPA heterotrimeric complex (Lin et al., 1998). It is also involved in DNA metabolism and the cellular response to DNA damage (Maréchal et al., 2014). Interestingly, all five of these genes were determined to be

Possibly Causative and the best candidate gene for the individual patient's infertility after further analyses were performed as described in section 4.4.6.

4.4.6 Classification of All de novo Mutations Within the Cohort of 185 Infertile Males

After the completion of the unbiased assessments presented in the previous sections, all 145 rare, protein altering DNMs were further classified using information on the function of the gene within which the DNMs lay. A total of six categories were investigated to give a conclusion on the likelihood of the DNM causing the patient's infertility, these included: presence of an infertile mouse model, protein function, interactions with any known infertility genes, RNA expression in the testis, RNA specificity to the testis and protein expression in the testis. Of all 192 initial DNMs, 29 affected genes were linked to male reproduction and classified as Possibly Causative, with a further 50 as Unclear, 52 as Unlikely Causative and 61 Not Causative (Figure 2.3, Table 4.4). A final list of all DNMs and their classifications can be found in Appendix B.

	Possibly Causative	**Unclear**	**Unlikely Causative**	**Not Causative**	**Total**
Missense	21	38	50	13	**122**
Frameshift	4	8	1	0	**13**
Stop gained	1	3	0	0	**4**
In-frame indels	3	1	1	1	**6**
Splice site variant	0	0	0	11	**11**
Synonymous	0	0	0	36	**36**
Total	**29**	**50**	**52**	**61**	**192**

Table 4.4: *De novo* mutation (DNM) classification summary. Rare DNMs were classified based on pathogenicity, based on pathogenicity prediction, American College of Medical Genetics (ACMG) classification, number of cases in gnomAD and presence of exact mutation in fertile control cohort, as well as functional data taking into account RNA expression in testis, RNA enrichment in the testis or involvement in spermatogenesis, protein expression in the testis, model organisms, the protein function in relation to spermatogenesis and interactions with known fertility genes.

4.4.7 Identification of novel de novo disease genes replicated across multiple cohorts

In order to formally evaluate the potential link of DNM genes to MI, a replication study was required. After searching relevant literature for additional studies on *de novo* mutations in male infertility, only one pilot WES study including 13 trios was recently published in MI (Hodžić et al., 2021). None of the DNM genes reported in this study showed DNMs in our cohort and the numbers were too small to be considered a useful replication. To further study the DNM genes identified in our cohort, an investigation took place into the presence of rare predicted

pathogenic mutations in these genes in exome datasets of infertile men (n = 2,506), in collaboration with members of the International Male Infertility Genomics Consortium and the Geisinger-Regeneron DiscovEHR collaboration (Dewey et al., 2016). For comparison, an exome dataset from a cohort of 11,587 fertile men and women from Radboudumc was included.

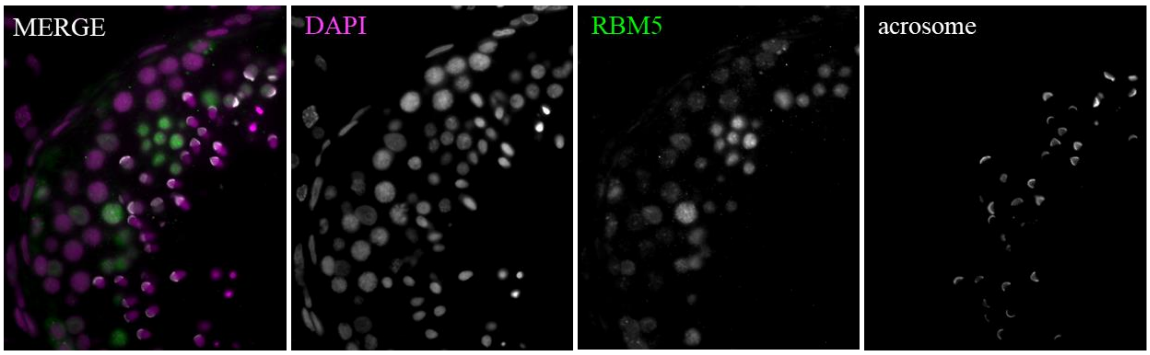
In the additional infertile cohorts, 17 LoF mutations were identified in our DNM LoF-intolerant genes (pLI ≥ 0.9), but no statistical enrichment was detected in the number of LoF mutations in these genes compared to those in fertile men (Two-tailed Fisher's Exact test with Bonferroni correction adjusted p-values > 0.05). Next, the enrichment of rare predicted pathogenic missense mutations was investigated in these cohorts (Table 4.5). A total of 11 genes showed an enrichment of pathogenic missense mutations in infertile men compared to fertile men (Two-tailed Fisher's Exact test, p-value < 0.05, Table 4.5). After applying the Bonferroni correction to counteract the effects of multiple testing, the only significant enrichment was observed in the *RBM5* gene (adjusted p-value = 0.03). In this gene, six infertile men were found to carry a rare pathogenic missense mutation, in addition to the proband with a *de novo* missense mutation (Figure 4.9, Table 4.6). These mutations were found to lie throughout the entirety of the gene with five of the seven lying within the RRM1, OCRE and G-Patch functional domains (Figure 4.9b). The sample information for all missense mutations found in the gene *RBM5* can be found in Figure 4.9c. Importantly, no such predicted pathogenic mutations were identified in men in the fertile cohort. In line with these results, *RBM5*, already highlighted above as an essential regulator of male germ cell pre-mRNA splicing and male infertility (O'Bryan et al., 2013), is highly intolerant to missense mutations (missense Z-score 4.17). In addition to the comparison between fertile and infertile men, it was investigated whether there was any difference between the number of predicted pathogenic mutations carried in fertile men compared to fertile women. However, none of the DNM genes showed a significant difference between the sexes in that aspect (Two tailed Fisher's Exact test with Bonferroni correction adjusted p values = 1, data not shown). Another gene of interest highlighted from this replication study is *HUWE1*. Whilst the statistical tests did not show these results to be significant, the adjusted p-value of 0.13 gives reason to consider this gene further, with six infertile males all having rare pathogenic missense mutations within *HUWE1* whereas none such mutation was observed in the fertile Dutch men. In line with this, this gene shows a high level of intolerance to missense mutations (missense Z-score 8.87). During the functional assessment of genes, HUWE1 showed medium protein and RNA expression in both the testis and epididymis as well as confirmed RNA expression during spermatogenesis. A 2017 paper determined in mice that *Huwe1* is critical for

maintenance of genomic stability during proliferation of spermatogonia through the vital role it plays in DNA repair regulation. (Bose et al., 2017).

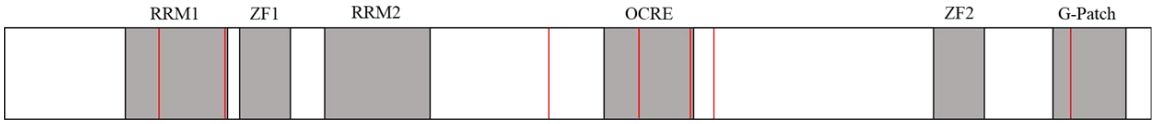
Gene	Missense Z-score	Total Infertile Men (n=2,506)	Fertile Dutch Men (n=5,784)	Burden test Infertile vs Fertile Men	
				p value	Adjusted p value
RBM5	4.17	7	0	0.0002	0.03
HUWE1	8.87	6	0	0.001	0.12
REN	0.80	7	1	0.001	0.21
HIST1H1D	-8.06	10	5	0.004	0.59
ABLIM1	1.62	6	1	0.004	0.60
FUS	2.21	4	0	0.01	1
CNOT4	3.49	5	1	0.01	1
CDC5L	2.78	6	2	0.01	1
ZNF629	3.86	6	2	0.01	1
PCDHB1	1.02	11	8	0.01	1
AK3	-1.97	10	7	0.02	1

Table 4.5: Rare pathogenic missense mutations in exome data from various cohorts of infertile men and fertile control cohorts. Genes affected by a rare protein altering DNM were investigated in additional cohorts of infertile patients and a cohort of verified fertile men and women to identify other individuals carrying rare protein altering missense mutations. A burden test was used to compare the total number of predicted pathogenic missense mutations observed in all infertile vs. fertile men. A two-tailed Fisher's Exact test was performed with and without Bonferroni correction applied to adjust p-values for multiple testing of all 152 genes of interest. The total male infertile cohort includes both the initial discovery cohort of 185 patients and an additional 2,321 infertile males from multiple replication cohorts detailed in section 2.1.3.

a



b



c

Cohort	Sample_ID	Genetic Coordinates (GRCh37)	Refseq ID	HGVS
MERGE Cohort of Infertile Men	M248	chr3:50137446	NM_005778.4:c.371C>T	p.Pro124Leu
NIJ/NCL Cohort of Patient-Parent Trios (DNM)	Proband_108	chr3:50140556	NM_005778.4:c.524A>G	p.Tyr175Cys
MERGE Cohort of Infertile Men	M2086	chr3:50147036	NM_005778.4:c.1193G>T	p.Gly398Val
MERGE Cohort of Infertile Men	M325	chr3:50148177	NM_005778.4:c.1429G>A	p.Gly477Arg
NIJ/NCL Cohort of Infertile Men (Singleton)	Proband00524	chr3:50150877	NM_005778.4:c.1517T>A	p.Val506Glu
MERGE Cohort of Infertile Men	M2013	chr3:50150955	NM_005778.4:c.1595C>T	p.Pro532Leu
NIJ/NCL Cohort of Infertile Men (Singleton)	Proband00282	chr3:50154750	NM_005778.4:c.2260A>G	p.Met754Val

Figure 4.9: RBM5 pathogenic mutations found in multiple infertile men from four different international cohorts: *a)* Localisation of *RBM5* in human testis. DAPI in magenta, *RBM5* in green and the acrosome in white. Staining was performed on testis sections from two control patients. For each patient, 50 seminiferous tubules were stained and analysed. *RBM5* is expressed in most stages of germ cell development albeit at various levels. Expression in Sertoli cells is also observed. *b)* Schematic representation of *RBM5* protein domains and the location of rare pathogenic mutations found in infertile males. *c)* Details of rare protein altering pathogenic variants found in *RBM5* in seven infertile males.

Cohort Name	Patient ID	Age of Patient	Karyotype	Y deletions	Conclusion Semen Analysis	Conclusion Testis Histology	Testicular Volume Left (ml)	Testicular Volume Right (ml)	Semen Conc. (x106)	Semen vBlume (ml)	Semen pH	FSH (U/L)	Testicular Sperm Retrieved	Urological History
NIJ/NLC Cohort of Patient-Parent Trios	Proband_108	40	46,XY	None	Severe oligo-zoospermia	N/A (No biopsy)	N/A	N/A	0.5	3.2	7.5	N/A	N/A	Unknown
NIJ/NLC Cohort of Infertile Men	Proband00282	32	46,XY	None	Azoospermia	N/A (Biopsy compromised)	15	15	0	3.1	7.7	6.7	Yes	Cryptorchism with orchidopexy (unknown if unilateral/bilateral)
NIJ/NLC Cohort of Infertile Men	Proband00524	36	46,XY	None	Azoospermia	Hypo-spermatogenesis	11	6	0	3.4	7.5	47	Yes	None
MERGE Cohort of Infertile Men	M248	43	46,XY	None	Azoospermia	Complete SCO	14	27	0	7.3	7.7	31.7	No	None
MERGE Cohort of Infertile Men	M325	36	46,XY	None	Azoospermia	Predominant SCO with some focal arrest up to round spermatids but no sperm	8	11	0	2.5	7.9	9.5	No	None
MERGE Cohort of Infertile Men	M2013	28	46,XY	None	Azoospermia	N/A (No biopsy)	7	8	0	3.2	8.3	27.9	N/A	None
MERGE Cohort of Infertile Men	M2086	33	46,XY	None	Azoospermia	N/A (No biopsy)	23	27	0	4.7	7.9	11.3	N/A	None

Table 4.6: Clinical details of individuals with RBM5 pathogenic mutations described in this study. Multiple infertile men from different cohorts were found with a rare pathogenic mutation in *RBM5* in addition to the Proband_108 where a DNM in *RBM5* was initially identified.

4.4.8 Recessive and Maternally Inherited Assessment of DNM Genes

The focus of the experiments in this thesis was to identify likely causative DNMs in patients with idiopathic male infertility. This would normally occur by causing dominant disease, but DNMs can also cause recessive disease if combined with an inherited mutation on the other allele. To assess this, all genes found to contain a rare DNMs were systematically evaluated through a literature review as well as a study of databases such as ClinVar to assess whether they had been implicated as a recessive male infertility gene. Any genes which had been implicated as recessive were then classified as unlikely causative if no additional mutation was identified on the alternate allele. One such mutation was found in the *MutS Homolog 5* (*MSH5*) gene. MSH5 is known to be involved in DNA mismatch repair and meiotic recombination in its heterodimeric form with MSH4 (Wyrwoll et al., 2021; Guo et al., 2017; Chen et al., 2022). Recently, two separate studies identified a total of six individual patients with likely pathogenic homozygous variants in *MSH5* (Kherraf et al., 2022; Wyrwoll et al., 2021). Due to the lack of a secondary inherited mutation within our infertile patient and the strong evidence for a recessive inheritance model for the gene MSH5, the heterozygous DNM in our cohort was classified as unlikely causative. Alternatively, in Proband_060, who carried a DNM in *Testis and Ovary Specific PAZ Domain Containing 1* (*TOPAZ1*) on the paternal allele, a maternally inherited variant predicted to be pathogenic was identified. *TOPAZ1* is a germ cell-specific gene which is highly conserved in vertebrates (Baillet et al., 2011). Studies in mice revealed that *TOPAZ1* plays a crucial role in spermatocyte, but not oocyte, progression through meiosis (Luangpraseuth-Prosper et al., 2015). In humans, *TOPAZ1* is expressed in germ cells in both sexes (Wang et al., 2018; Guo et al., 2015; Li Li et al., 2017). Analysis of the testicular biopsy of this patient revealed a germ cell arrest in early spermiogenesis (Figure 4.10). The DNM found in this patient alongside the maternally inherited variant was therefore classified as likely causative, with *TOPAZ1* noted as a likely recessive candidate gene for male infertility.

Maternally inherited mutations can also result in dominant causes of male infertility if not affecting female fertility. All DNM genes were examined for the presence of maternally inherited mutations in the entire cohort and compared this to the presence of paternally inherited mutations in the same genes. A total of four maternally inherited variants predicted to be pathogenic were identified in DNM genes (*TENM2* (2×), *CWC25*, and *EVC*). All of these variants, however, were also observed multiple times in an exome dataset from a cohort of 5,784 fertile men suggesting that these maternally inherited variants are not causative of MI.

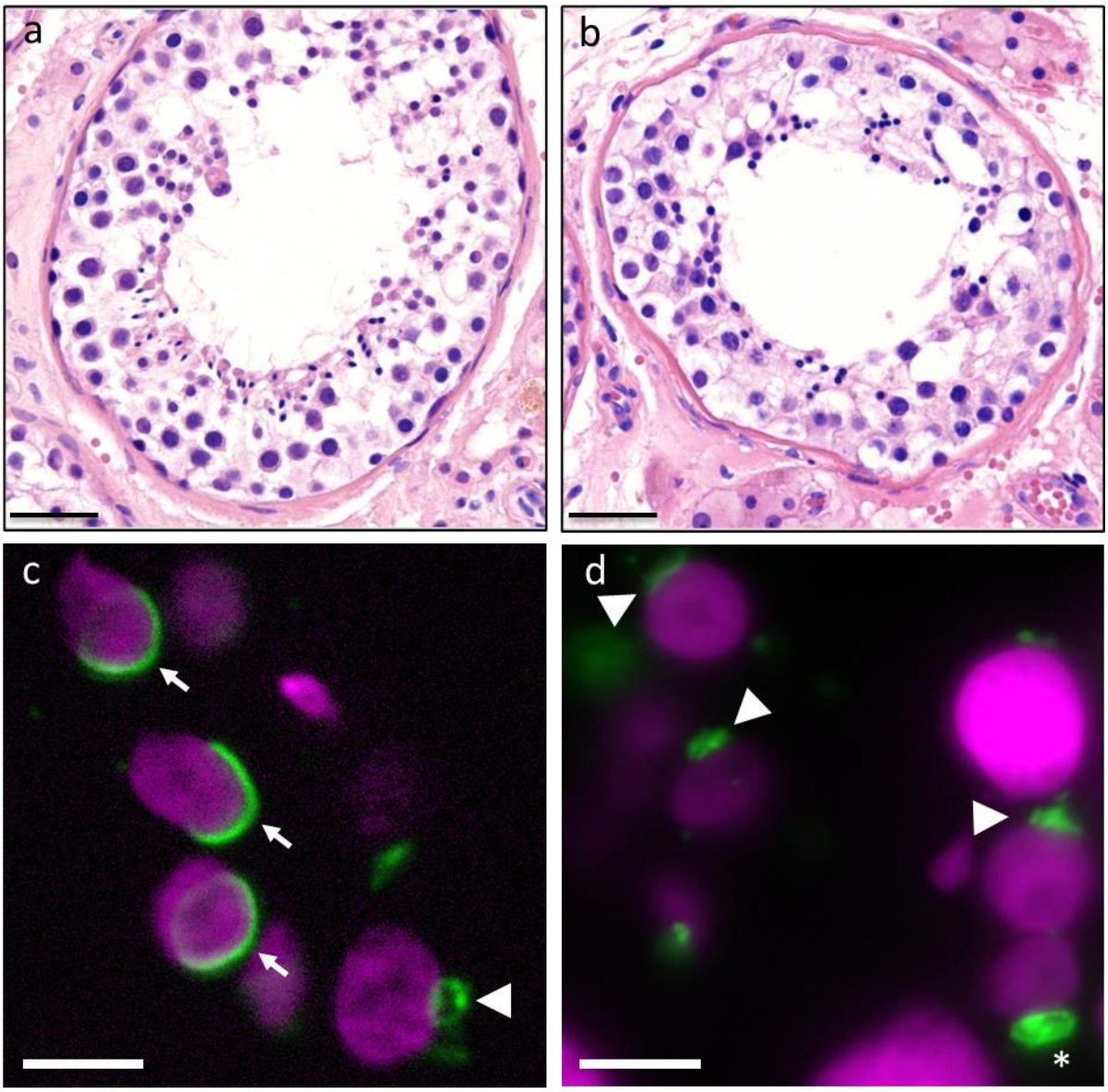


Figure 4.10: Description of control and TOPAZ1 proband testis histology and aberrant acrosome formation. *a, b)* H&E stainings of (**a**) control and (**b**) Proband_060 with pathogenic mutations in *TOPAZ1* gene. The epithelium of the seminiferous tubules in the *TOPAZ1* proband show reduced numbers of germ cells and an absence of elongating spermatids based on the analysis of 150 seminiferous tubules in control and patient. *c, d)* immunofluorescent labelling of DNA (magenta) and the acrosome (green) in (**c**) control sections and (**d**) *TOPAZ1* proband sections. (**c**) The arrowhead indicates the acrosome in an early round spermatid and the arrows the acrosome in elongating spermatids. Spreading of the acrosome and nuclear elongation are hallmarks of spermatid maturation. (**d**) No acrosomal spreading (see arrowheads) or nuclear elongation is observed in the *TOPAZ1* proband. The asterisk indicates an example of progressive acrosome accumulation without spreading. Scale bar: (**a**, **b**) 40 µm and (**c**, **d**) 5 µm. (Oud et al., 2022)

4.5 Discussion

In this Chapter, I aimed to address the lack of research and knowledge surrounding the topic of isolated male infertility by analysing exome sequencing data of 185 infertile males and their parents. After filtration of raw data files, 192 rare *de novo* mutations were identified in the cohort of infertile men with 145 of these being protein altering. Currently there are no large-scale studies into DNMs in cases of male infertility and therefore there is no direct comparison group when assessing the DNM rate in this cohort. A large amount of research has been performed looking at diseases with a similar reduced reproductive fitness in which DNMs were assessed such as ID and Autism Spectrum Disorder (Gilissen et al., 2014; Acuna-Hidalgo et al., 2016; Michaelson et al., 2012). Whilst the rates vary slightly within each study, it is widely accepted that on average within the genome of each individual, one to two *de novo* point mutations will be identified within the coding regions of DNA (Acuna-Hidalgo et al., 2016). Our results are highly compatible with these boundaries with a rate of 1.04 DNMs per patient in coding regions. Of these 145 protein altering DNMs, 122 were missense variants, 17 were loss-of-function (LoF) mutations (13 frameshift, four stop-gained) and six were in-frame indels. An assessment of the pathogenicity of each variant was then carried out using three different tools (SIFT, PolyPhen and MutationTaster) which determined the likely effect of the DNMs on the overall structure and function of the gene they lay within. Following this assessment, 84 of the rare protein altering DNMs were found to be likely pathogenic at a variant level, whilst the remaining 61 were likely to be benign. Two additional *de novo* CNVs were detected, effecting a total of seven genes, two of which have been previously reported in a partially overlapping deletion reported in a patient with cryptorchidism and non-obstructive azoospermia (Seabra et al., 2014).

Of all 192 genes identified in our infertile male cohort, one was determined to lie in a previously highlighted AR MI gene (*MSH5*). According to a 2021 update, there are currently 68 genes linked to MI in a definitive, strong, or moderate manner (Houston et al., 2021). These 68 genes represent 0.34 % of the total genes found in the human genome. Applying this to our own dataset of 192 genes, we would expect around one DNM to lie in a gene known for a recessive form of MI which is consistent with our results. When factoring in all known AR genes, 22/192 DNMs lay in genes associated with some form of AR disorder. A 2021 study focussing on AR variation determined a list of 1,929 currently known genes with confirmed recessive inheritance representing around 9.65% of all known genes (Fridman et al., 2021). Of the 192 genes found in our cohort, we would therefore expect around 19 DMNs in known AR genes, again highly consistent with our own results. Despite being linked to recessive conditions, these DNMs may

still play a role in dominant male infertility with many of these genes being expressed ubiquitously, playing a role in multiple important biological processes within the human body. A DNM was highlighted in the gene *CWC27* which is associated with AR Retinitis Pigmentosa (Bertrand et al., 2022), this gene however shows low tissue specificity and enhanced RNA expression as well as enrichment in early spermatids. This information, alongside its role in pre-mRNA splicing highlights it as a gene of interest for male infertility (Brea-Fernández et al., 2019; Busetto et al., 2020).

Prior to a functional assessment of the genes containing DNMs and a final classification for each DNM being provided, unbiased analyses were then performed to determine whether any significance was seen in the data. An initial test emphasized an enrichment for LoF DNMs in LoF-intolerant genes when the pLI and LOEUF scores were investigated for all genes containing LoF DNMs in the infertile male cohort, compared to LoF mutations in a healthy control male cohort (DNM LoF mutations in patients n = 17, median pLI in patients with male infertility = 0.80, DNM LoF mutations in controls n = 21, median pLI in controls = $3.75 \times 10{-}5$, p value = $1.00 \times 10{-}5$, N simulations = 100,000). This observation indicates that LoF DNMs likely play an important role in MI as is seen in developmental disorders and severe intellectual disability (Y. Gu et al., 2020; Fritzen et al., 2018). One example from the patient cohort is a heterozygous likely pathogenic frameshift DNM which was observed in the LoF-intolerant gene *GREB1L* (pLI = 1) of Proband_076. Homozygous *Greb1l* knockout mice appear to be embryonic lethal, however, typical MI phenotypic features such as abnormal foetal testis morphology and decreased foetal testis volume are observed (De Tomasi et al., 2017). Interestingly, this Proband_076 presents with reduced testis volume and severe oligozoospermia. Nonsense and missense mutations in *GREB1L* in humans are known to cause renal agenesis (Brophy et al., 2017) (OMIM: 617805), not known to be present in our patient, however, all previously reported damaging mutations in *GREB1L* causing renal agenesis are either maternally inherited or occurred *de novo* which led the authors of a renal agenesis studies to speculate that disruption to *GREB1L* could cause infertility in males (De Tomasi et al., 2017). A recent WES study involving a cohort of 285 infertile men also noted several patients presenting with pathogenic mutations in genes with an associated systemic disease where male fertility is not always assessed, highlighting the need for broader phenotyping to occur during clinical studies in order to progress research forward in rare diseases (Alhathal et al., 2020).

The work in this Chapter also uncovered an enrichment for likely pathogenic missense *de novo* mutations (at a variant level) in missense intolerant genes by comparing those pathogenic missense DNMs present in the infertile male cohort (median Z-score = 1.21, n = 63) to the

missense DNMs categorised as likely benign at a variant level (median Z-score = 0.98, n = 59) within the same cohort (p value = 5.01×10^{-4}). No such difference was seen when comparing these two categories within the control cohort (p value = 0.95, Figure 4.4c). This indicates that, missense mutations that are determined to be likely pathogenic occur more frequently in genes which are intolerant to these types of mutations in this infertile cohort than in a control cohort. This suggests that more damaging *de novo* missense mutations are occurring in genes which are intolerant to these missense mutations and that may contribute to MI.

A protein-protein interaction network was created to investigate the functional pathways containing genes with predicted pathogenic protein altering DNMs present. This investigation highlighted a significant enrichment for interactions amongst these 84 genes with a PPI enrichment p value of 2.35×10^{-2} (Figure 4.5). This level of enrichment did not exist in either genes containing predicted benign protein altering DNMs or those affected by synonymous DNMs (Figure 4.6). This suggests that the proteins affected by predicted pathogenic DNMs in the cohort of infertile males likely share common biological functions. Two examples of this can be seen in Figure 4.7 where two functionally distinct modules of interacting proteins lie within a central network displayed in Figure 4.5. The two modules highlight numerous genes involved in the processes of mRNA splicing and mitotic cell cycle. Due to the high levels of cellular division and alternative splicing within the testis and during spermatogenesis, damaging mutations in genes essential to the successful maintenance of these pathways are likely to have catastrophic effects during spermatogenesis itself (Yeo et al., 2004; Song et al., 2020).

De novo mutations are highly researched in the field of ID and, within this, they have been observed to occur more frequently in regions of the genome which are more likely to cause disease when compared to healthy cohorts (Lim et al., 2017; De Vas et al., 2023). Our research is the first of its kind to show statistically significant test results highlighting pathogenic DNMs occurring in more damaging regions in patients with infertility. Given the genetic heterogeneity of the disease, it is unsurprising that we see no recurrently mutated genes. The biological pathways in which the likely pathogenic DNM genes work, and the intolerant genes in which they lie, mirror the results from previous studies into the distribution of DNMs in rare disease cohorts.

After performing replication studies to identify any further heterozygous mutations in genes containing rare protein altering DNMs, the gene *RBM5* was highlighted as a strong candidate gene for MI with a possibly causative DNM being present in the infertile male trio cohort and six additional heterozygous variants found in the larger infertile male cohort. The gene *RBM5*

was previously highlighted in the protein-protein interaction network as involved in mRNA splicing gene *RBM5* and is well known to be an essential regulator of haploid male germ cell pre-mRNA splicing. Mice with a homozygous ENU-induced allele point mutation in *RBM5* present with azoospermia and germ cell development arrest at round spermatids (O'Bryan et al., 2013). (O'Bryan et al., 2013). Despite the abundance of evidence for the role of this gene in mouse infertility, little research has been performed on this gene in relation to human MI (Jamsai et al., 2017). Due to its strong links to infertility in mice and its essential role in the pre-mRNA splicing pathway, alongside the presence of damaging heterozygous mutations in seven patients suffering from severe male infertility, the gene *RBM5* presents as a strong candidate gene for human male infertility with a need for further investigation into the function of this gene within the process of spermatogenesis itself. As well as its role in pre-MRNA splicing, *RBM5* is a known tumour suppressor with multiple studies linking dysregulation of the gene to breast, lung and renal cancers (Oh et al., 2006; Scanlan et al., 1999). A 2017 study investigated the role of *RBM5* as a tumour suppressor in mice by downregulating the gene and exposing them to a known tobacco smoke carcinogen. The main finding of this study was that whilst *Rbm5* LOF mice did not develop cancer at a higher rate than healthy mice, the mutant mice displayed a much more aggressive lung cancer pathology (Jamsai et al., 2017). This study highlights how genetics, along with environmental factors such as smoking can lead to greater risk of comorbidities, particularly in patients with DNMs in genes which are ubiquitously expressed throughout the body and play a role in several different biological pathways. All patients who were sequenced for the DNM and singleton cohorts in both the Netherlands and Newcastle underwent assessment for suitability for the study and during this were questioned as to any underlying medical conditions either potentially related or not to their infertility. This is standard practice when patients attend clinics. According to our patient data, there was no currently known comorbidities in the *RBM5* DNM patient. This would be an important subject to broach with patients in future once more genes have been more confidently linked to male infertility during genetic counselling sessions. As well as its role in pre-MRNA splicing, *RBM5* is a known tumour suppressor with multiple studies linking dysregulation of the gene to breast, lung and renal cancers (Oh et al., 2006; Scanlan et al., 1999). A 2017 study investigated the role of *RBM5* as a tumour suppressor in mice by downregulating the gene and exposing them to a known tobacco smoke carcinogen. The main finding of this study was that whilst *Rbm5* LOF mice did not develop cancer at a higher rate than healthy mice, the mutant mice displayed a much more aggressive lung cancer pathology (Jamsai et al., 2017). This study highlights how genetics, along with environmental factors such as smoking can lead to greater risk of comorbidities, particularly in patients with DNMs in genes which are ubiquitously expressed

throughout the body and play a role in several different biological pathways. All patients who were sequenced for the DNM and singleton cohorts in both the Netherlands and Newcastle underwent assessment for suitability for the study and during this were questioned as to any underlying medical conditions either potentially related or not to their infertility. This is standard practice when patients attend clinics. According to our patient data, there was no currently known comorbidities in the *RBM5* DNM patient. This would be an important subject to broach with patients in future once more genes have been more confidently linked to male infertility during genetic counselling sessions.

Only the gene *RBM5* showed statistical significance after burden testing but five singleton men were also identified as harbouring likely pathogenic heterozygous mutations in the X chromosome gene *HUWE1*. The *DNM* initially highlighted in this gene was classified as "possibly causative" due to the multiple lines of evidence suggesting a role for *HUWE1* as a candidate male infertility gene. Expanded studies into larger cohorts of infertile males and their parents are necessary to potentially identify more patients with mutations in both this gene and all others identified in this study. The use of singleton cohorts as a replication in this study has provided interesting and useful results, leading directly to the investigation of *RBM5* as a cause of male infertility in Chapter 5 of this thesis. It must be noted that all additional variants highlighted in the additional infertile cohorts are of unconfirmed inheritance with only their status as likely pathogenic rare heterozygous mutations being confirmed. Whilst the results were constrained by confirming the absence of these mutations in large databases such as gnomAD, as well as our control "fertile" male cohort, there is a chance they could be paternally inherited variants. It is because of this that we emphasise the necessity of validation of these results in further large-scale studies in patient-parent cohorts to give true replication. Additionally, the five infertile singleton cohorts made use of different enrichment kits during WES leading to potential variation in the regions covered, with some gene data unavailable for certain cohorts.

In 2010, a pilot study was published pointing to a *de novo* paradigm for mental retardation (Vissers et al., 2010) (now more appropriately termed developmental disorders or intellectual disability). This work contributed to the widespread implementation of patient–parent WES studies in research and diagnostics for neurodevelopmental disorders (Vissers et al., 2016), accelerating disease gene identification and increasing the diagnostic yield for these disorders. The data presented here suggest that a similar benefit could be achieved from trio-based exome sequencing in MI. In order to achieve this there is an urgent need to expand on this work as larger studies are essential to identify recurrently mutated DNM genes and further demonstrate

the exact contribution of DNMs to MI. Modelling studies recently performed for developmental disorders showed that more than 350,000 trios may be required to have approximately 80 % power to detect all haplo-insufficient genes causing this disorder (Kaplanis et al., 2020). Evidently, these numbers can only be reached by implementing trio-based exome sequencing as a routine diagnostic test and by sharing these diagnostic data with the international research community. The work required to functionally validate the impact of these DNMs on spermatogenesis is also no small task and will require extensive work throughout the research community. Altogether, this will not only help to increase the diagnostic yield for men with infertility but will also enhance our fundamental biological understanding of human reproduction and natural selection. In addition, it will indicate whether male infertility follows a dominant inheritance pattern, and this has impact for disease transmission. Couples that seek treatment for male infertility should be counselled on the risk of transmitting this condition to their offspring, something that is now limited to couples receiving fertility treatment due to Y chromosome deletions. Male infertility is also increasingly seen as the most visible symptom of a more complex disease with associated comorbidities (Kasman et al., 2020). Studying the long-term health of men with DNMs in specific genes should help in identifying genotype–phenotype correlations that may impact more than the fertility of these men.

Chapter 5: mRNA Splicing in Male Infertility

5.1 Introduction

5.1.1 Novel Candidate Male Infertility Genes and Their Roles in mRNA Splicing

In Chapter 4 of this thesis, predicted pathogenic DNMs were identified in infertile patients in five genes which all play a role in the process of pre-mRNA splicing via the spliceosome. The proteins encoded by the genes *U2AF2*, *CWC27*, *CDC5L*, *HNRNPL* and *RBM5*, all interact with the spliceosomal machinery in differing circumstances, and all pose an interesting avenue for further functional investigation. All five genes show prominent levels of transcription across multiple human tissues, including the testis, according to data from the Human Protein Atlas (Uhlén et al., 2015). In order to further understand the role these genes and their transcriptional products play, a brief introduction into pre-mRNA splicing and the spliceosome is essential.

5.1.2 Pre-mRNA Splicing and the Spliceosome

Pre-mRNA splicing is an essential step in post-transcriptional gene regulation and allows significant expansion of the human proteome and increased functional complexity, with around 95 % of all multi-exonic human genes providing one or more alternative splicing products (Pan et al., 2008; Nilsen & Graveley, 2010). The total number of potentially protein coding genes in the human genome is approximately 20,000, however, the number of transcripts created from these genes has been suggested to be around 300,000(Pertea et al., 2018). It has been noted that different splice variants created during pre-mRNA splicing are typically differentially expressed in the different tissues of an organism or are produced during various stages of developmental processes and in response to changing physiological states (Trapnell et al., 2010).

The process of pre-mRNA splicing occurs via a large ribonucleoprotein complex known as the spliceosome, which is comprised of five small nuclear ribonucleoproteins (snRNP) as well as at least 300 non-snRNP protein factors (Ren et al., 2021). The snRNPs are RNA-protein complexes which recognise specific sites at intron/exon boundaries and bind to unmodified pre-mRNA, and other proteins, to form and engage the spliceosome (Figure 5.1). This facilitates intron excision and exon ligation, or pre-mRNA splicing. Each snRNP contains a uridine rich small nuclear RNA (snRNA), U1, U2, U4, U5 and U6 and its own unique group of proteins.

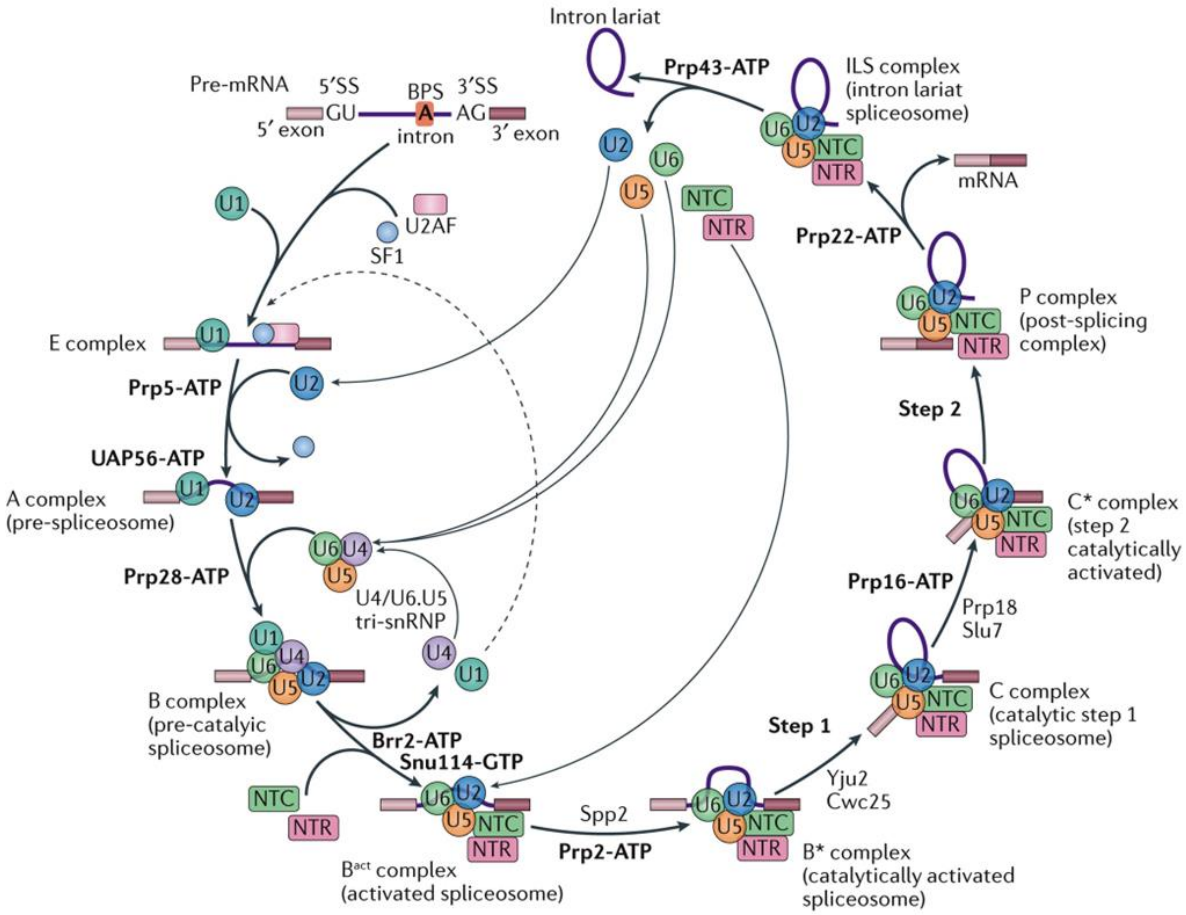


Figure 5.1: Assembly and activation of the yeast spliceosome and the complete splicing-reaction cycle. The 5′SS, BPS and 3′SS are first recognized by the U1 small nuclear ribonucleoprotein (snRNP), splicing factor 1 (SF1; also known as branchpoint-bridging protein) and U2AF, respectively, forming an early spliceosome (known as the E complex). SF1 is displaced by the U2 snRNP to form the pre-spliceosome (A complex), which associates with the U4/U6.U5 tri-snRNP to assemble into the pre-catalytic spliceosome (B complex). The B complex represents the first fully assembled spliceosome. There are at least six additional distinct spliceosome complexes: Bact, B*, C, C*, P and the intron lariat spliceosome (ILS). Each complex has a unique composition, and conversion between complexes is driven by highly conserved RNA-dependent ATPase/helicases (in bold). Notably, a spliceosomal complex can have distinct conformational states, which may also differ in composition. For example, the B and ILS complexes each have at least two distinct conformations (Shi, 2017) Reproduced with permission from Springer Nature.

To begin, the U1 snRNP recognises and binds to the 5′ splice site of the pre-mRNA by base-pairing within the intron of the conserved splice site. At the same time, Splicing Factor 1 (SF1/mBBP) binds to the conserved branch point sequence (BPS) within the intronic region of the pre-mRNA, U2FA2 (U2AF 65k Da subunit) binds to the polypyrimidine (PY) tract and U2AF1 (U2AF 35 kDa subunit) binds to the 3′ splice site (Berglund et al., 1998). This is referred to as Complex E of the Spliceosome and during its formation, the intron is bridged and the

splice sites are brought together (Kent & MacMillan, 2002). In an ATP-dependent reaction, the U2 snRNP displaces SF1/mBBP bound at the BPS, to form the A Complex. A pre-formed tri-snRNP U4/U6.U5, in which U4 and U6 are extensively base-paired to each other, then binds to the 5′ splice site to form the B complex (Fica & Nagai, 2017). Following the destabilization of the U1 and U4 snRNPs and association of PRP19/CDC5L (NTC complex) as well as a number of other RNA-RNA and RNA-Protein re-arrangements, the B complex then becomes activated (Grote et al., 2010). This stage initiates the splicing process whereby a cleaved 5′ exon is formed, and an intermediate intron-3′-exon lariat is also produced (Bessonov et al., 2008). The C complex is established during this time and catalyses the second step of splicing, this involves the full excision of the intron to create an intron lariat, and the 5′ and 3′ exons are then ligated to form mature mRNA which can then potentially be translated into a protein. (Wolf et al., 2009).

5.1.3 Candidate Male Infertility Genes and Their Roles in the Spliceosome

Whilst Figure 5.1 highlights a large number of proteins and complexes involved with the core spliceosome, a large number of additional splicing factors are involved in these steps including the five genes of interest highlighted in section 5.1.1.

U2AF2 encodes a core pre-mRNA splicing factor which is also known as the 65kDa subunit U2AF ($U2AF^{65}$) which binds initially to the PY tract of the transcript undergoing pre-mRNA splicing via the spliceosome. The $U2AF^{65}$ subunit then forms a complex with SF1 and $U2AF^{35}$ which in turn then ensures splice site fidelity at the 3′end within complex E of the spliceosome (Glasser et al., 2022; Maji et al., 2020). $U2AF^{65}$ is also known to recruit the NineTeen Complex (NTC) to the spliceosome to form the B complex (Grote et al., 2010). The NTC is made up of multiple proteins including both PRP19C and CDC5L, the latter being encoded by another gene highlighted in Chapter 4 of this thesis as containing a DNM in an infertile patient. It is thought that the NTC is likely to be a target for a number of spliceosomal ATPases which catalyse the transitions between and activation of complexes of the spliceosome (de Almeida & O'Keefe, 2015). The role of the splicing factor encoded by the gene *CWC27* is not as well understood as that of *U2AF2*. The gene *CWC27* encodes for a spliceosome associated cyclophilin that interacts with another similar protein called CWC22 and recruits the translation initiation factor EIF4A3 to the spliceosome (Busetto et al., 2020). This in turn leads to the formation of the exon junction complex (EJC) within the B-act and C complexes of the spliceosome (Bertrand et al., 2022). *HNRNPL* encodes the splicing factor hnRNPL which is an essential component of the nuclear hnRNP complex and can act either as an inhibitor or an activator during the pre-mRNA

splicing process. As an inhibitor, hnRNPL is involved in the inhibition of recessive exon inclusions by binding to CA repeat regions (Hui et al., 2003), maintaining a normal working splicing process. As an activator hnRNPL is involved retaining/activating as well as determining the splicing efficiency of various introns (J. Gu et al., 2020). The final gene of interest was *RBM5* or *RNA binding motif protein 5*. *RBM5* is a known tumour suppressor gene and has been implicated in cancers of both the lung and breast upon varying levels of gene expression (Jamsai et al., 2017). The protein produced by *RBM5* is responsible for the regulation of alternative splicing in multiple apoptosis related mRNAs during the spliceosomal A complex (Niu et al., 2012). The RBM5 protein and its function as a pre-mRNA splicing factor is the main focus of the work in this chapter due to the detection of one predicted pathogenic DNM as well as five predicted pathogenic heterozygous missense mutations in this gene in other cohorts of infertile males (none of which were observed in fertile control men, see Chapter 4).

5.1.4 Spliceosome Related Diseases

Mutations in genes involved in the spliceosome can lead to localised dysregulation of pre-mRNA splicing and disease states with around 9 % of all disease-causing mutations reported in the Human Gene Mutation Database (HGMD) being splicing mutations particularly in tissues where alternate splicing is rather prevalent (Anna & Monika, 2018). One such example occurs in Retinitis Pigmentosa (RP), the most common inherited retinal disease, whereby a progressive degradation of photoreceptors eventually results in blindness. In around 15 % – 20 % of all autosomal dominant cases of RP (adRP), a mutation occurs in the *RP-PRPF* gene family encoding human splicing factors (Yang et al., 2021). The most common mutations are found in *PRPF3*, *PRPF4*, *PRPF6*, *PRPF8*, *PRPF31*, SNRNP200/*Brr2* and *RP9* genes. All but *RP9*, a non snRNP splicing factor, are involved in the tri-snRNP U4/U6.U5 (Figure 5.2).

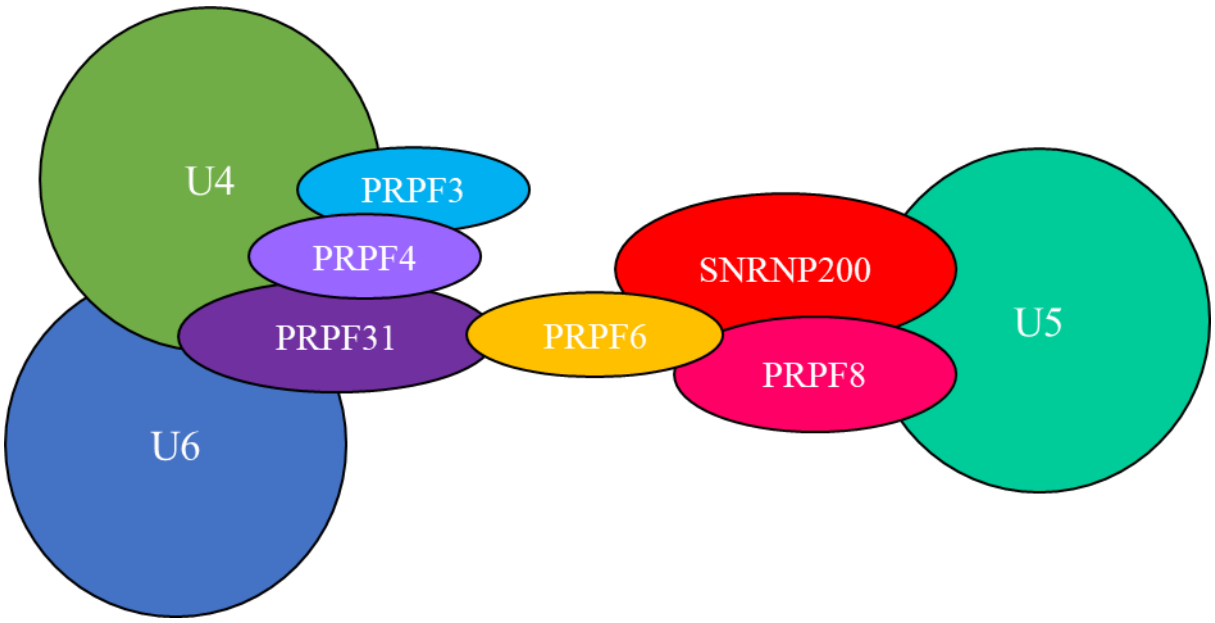


Figure 5.2: U4/U6.U5 tri-snRNP complex consisting of 3 major snRNPs U4,U5 and U6 and their associated snRNP specific factors. It is this tri-snRNP which binds to the 5′ splice site of the pre-mRNA forming the B complex of the Spliceosome via an ATP-dependant reaction.

PRPF3,*4* and *31* are U4/U6 snRNP specific factors, *PRPF6*,8 and *Brr2* are U5 snRNP specific factors (Růžičková & Staněk, 2017). Accounting for 8.9 % of all adRP cases, mutations in the gene *PRPF31* lead to a disruption of the normal function of the PRPF31 protein which constitutes part of the U4 snRNP and tethers the U5 snRNP to the U4/U6 snRNP by interacting with PRPF6, thereby inhibiting the formation of the U4/U6.U5 tri-snRNP (Yang et al., 2021). Over 220 mutations of distinct types have been recorded in this gene as resulting in either altering the sequence and structure of the protein or in a decreased level of expression which in turn then leads to retina-specific spliceosome dysregulation (Růžičková & Staněk, 2017). Mutations in the gene *PRPF31* lead to a specific form of RP with varied penetrance across patients known as Retinitis Pigmentosa 11 or RP11(Mordes et al., 2006). Furthermore, the gene *SMN1* produces the protein SMN which plays an indirect yet essential role in the generation of the pre-splicing machinery by recruiting and assembling the snRNP spliceosome components (Pellizzoni et al., 1998). Whilst the specific disease-causing pathways are not yet fully understood, it has been theorised that it is this specific gene function which is being disrupted in patients where *SMN1* mutations were found to be causing reduced levels of SMN protein, leading to Spinal Muscular Atrophy (Hensel et al., 2020).

It is noteworthy that mutations in genes encoding core splicing factors, as well as dysregulation of the function of these factors, can be seen to result in disruptions of specific tissues such as the retina (Mordes et al., 2006) despite being so integral in the generic process of pre-mRNA

splicing. One theory behind this is that the tissues which are affected by these diseases are those which show relatively high levels of alternative spicing and increased presence of snRNAs (Tanackovic et al., 2011). The fact that these processes occur much more frequently in specific tissues suggests that pre-mRNA splicing is essential and integral to the functioning of these tissues and mutations in genes associated with the spliceosome or other splicing factors are more likely to cause disruption and disease states (Cao et al., 2011).

5.1.5 The Role of mRNA Splicing During Spermatogenesis

Spermatogenesis is a transcriptionally taxing and complicated process during which mature spermatozoa are produced from spermatogonial sperm cells (Song et al., 2020) During each of these stages, multiple different genes are dynamically expressed to regulate and facilitate the differentiation and processing of the male germ cells. The testis and the brain both show the highest levels of alternative splicing (AS) occurring in any tissues throughout the body (Naro et al., 2021). The highest proportion of this AS is via exon skipping (Yeo et al., 2004).

Pre-mRNA splicing can influence the regulation of spermatogenesis in multiple ways, some genes involved in spermatogenesis will undergo AS and thus disruption of splicing factors effects their function downstream. An example lies within the gene *SPAG16*. *SPAG16* or Sperm Associated Antigen 16, is a protein coding gene which, via AS, produces two major protein isoforms, one which associates with the axoneme of the sperm tail and the second which localises in the nucleus of post meiotic germ cells (Zhang et al., 2007).

In some other cases, it is possible that the disruption of genes involved pre-mRNA splicing via knockout or genetic mutation, can potentially cause detrimental downstream effects to the splicing machinery within the testis and can even lead to spermatogenic arrest. The MORF4-related gene which lies on chromosome 15, *MRG15*, is a multifunctional chromatin organiser which is thought to play a role in histone acetylation and binds to methylated Histone H3 lysine 36 (H3K36) in introns of transcriptionally active cells (Peña et al., 2011; Luco et al., 2010). The gene *MRG15* is also implicated in homologous recombination DNA repair and alternative splicing (Iwamori et al., 2016; Hayakawa et al., 2010). Whilst pre-mRNA splicing is often regulated by RNA binding proteins, regulation via epigenetic factors such as histone modification can also occur. Spermatogenic arrest is observed when MRG15 is knocked out in mice, and one particular study notes that the mRNA of no less than 66 germ cell specific genes were no longer present in the mice (Iwamori et al., 2016). Furthermore, it was noted that introns typically removed in multiple mRNAs involved with sperm chromatin condensation in the

round spermatids were still present after MRG15 knockout. This suggests that depletion of the MRG15 protein, which usually co-localises with splicing factors within the round spermatids, potentially causes male infertility by disrupting the epigenetic regulation of pre-mRNA splicing via histone modification.

5.1.6 Using a FAS Minigene to Investigate the Effect of SNVs on the Splicing Functionality of RBM5

A study was previously published in 2008 in which the relationship between *RBM5* and the *FAS* (Mollinedo & Gajate, 2006) gene was investigated (Bonnal et al., 2008). *FAS*, also known as Fas Cell Surface Death Receptor, has been shown to play a significant role in the regulation of programmed cell death. *RBM5* is involved in the regulation of alternative splicing of apoptosis related genes, including the *FAS* receptor and when the levels of RBM5 were altered, a switch was seen between the levels of two isoforms of the *FAS* receptor gene products (Bonnal et al., 2008).

The *FAS* gene encodes two potential isoforms with antagonistic functions. When exon 6 of the gene is spliced, a membrane bound pro-apoptotic form of the *FAS* receptor is translated. In contrast, when *FAS* exon 6 is skipped the mRNA encodes a soluble apoptotic inhibitor protein (Bonnal et al., 2008). To perform these experiments a *FAS* minigene was created, expressing the genomic sequences between exons 5 and 7 of the gene, which also contained a mutation within the PY tract upstream of exon 6. This results in a naturally higher occurrence percentage of exon skipping (Figure 5.3). This mutation was included to allow for a more obvious observation of increased exon inclusion. The main result of the paper demonstrated that when levels of RBM5 are depleted, a decrease in the soluble antiapoptotic form of *FAS* receptor was seen. In contrast, RBM5 overexpression led to an increase in *FAS* exon 6 skipping (Bonnal et al., 2008).

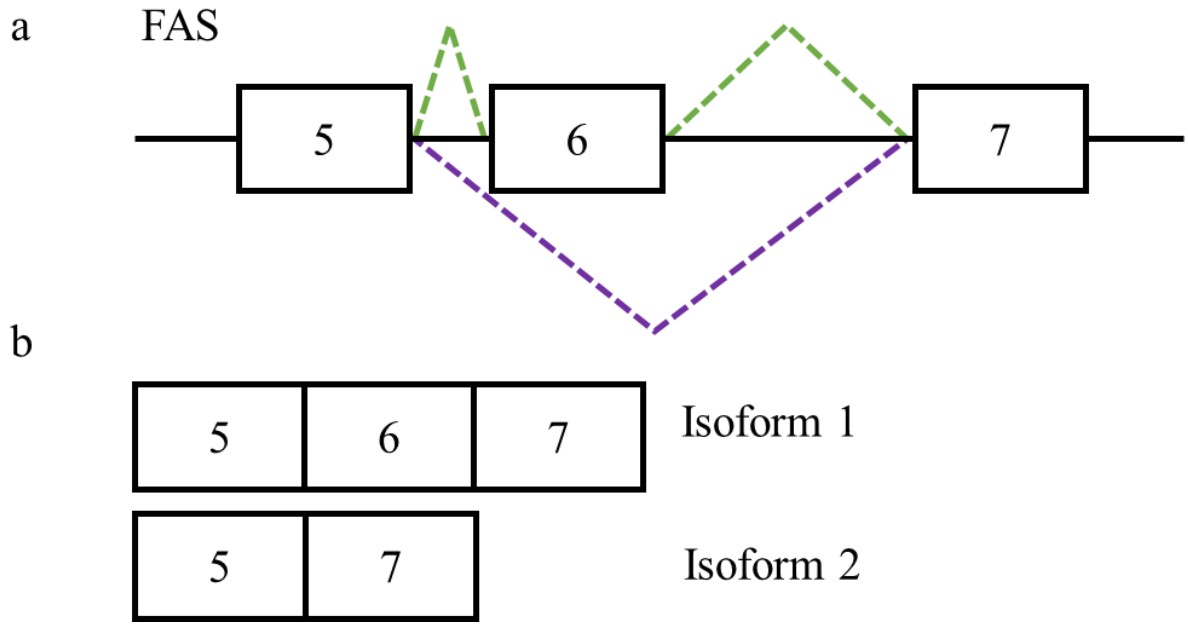


Figure 5.3: FAS Minigene and Products: *a)* The *FAS* minigene used in this Chapter's experiments and in the paper by (Bonnal et al., 2008) is represented here. The green dashed lines represent the exon junctions after splicing between exons 5 and 6 and exons 6 and 7 whereby the sixth exon is retained. The purple dashed annotation represents the exon junction of shorter isoform product with the exclusion of exon 6 of the gene.
b) The two splice products made from the minigene are shown, with the initial longer product and the shorter second product with a skipped exon.

In order to investigate the variants found in patients in Chapter 4 of this thesis, experiments were performed making use of the *FAS* minigene previously designed by and kindly supplied by Professor Juan Valcarcel Juarez and his group (Bonnal et al., 2008). My hypothesis was that if the mutations found in the gene *RBM5* in our infertile male patients are potentially causing infertility via changing the pre-mRNA splicing pathway, introducing these mutations into the RBM5 sequence, and co-expressing the mutated genes with the *FAS* minigene in a model cell system should cause a change in splice inclusion patterns between. To test this, five of the six RBM5 variant sequences and one wild-type sequence were ordered as synthesised DNA (G blocks), cloned into a splice reporter plasmid, and transformed into *E. coli*. Selected plasmids were then transfected into HEK293 cells alongside the *FAS* minigene before RNA and protein extraction. The level of splice inclusion was then measured as a Percentage Splice Inclusion (PSI), and then compared to the PSI made from HEK293 cells co-expressing wild type RBM5 protein to determine whether a significant difference was seen in the levels of exon inclusion.

5.2 Aims

1. To investigate the potential mechanism of the mRNA splicing pathway disruption as a cause of MI.
2. To design and implement a co-transfection assay utilising five mutant and one wildtype *RBM5* expression vector alongside a pre-existing *FAS* minigene in HEK293 cells.
3. To determine the effect of these *RBM5* mutations on the gene's capabilities to regulate splicing of the *FAS* minigene, analysing any resulting variation in *FAS* splice isoform ratios made between mutants and wild type.

5.3 Methods

The methods for work performed in this Chapter can be found in Chapter 2, Section 2.4. G-blocks representing five mutations, previously highlighted in Chapter 4 of this thesis, were ordered and the corresponding variants can be seen below in Table 5.1.

Insert Number	Patient Variant	RefSeq ID	HGVS
1	Chr3:50140556-A-G	c.524A>G	p.Tyr175Cys
2	Chr3:50150877-T-A	c.1517T>A	p.Val506Glu
3	Chr3:50154750-A-G	c.2260A>G	p.Met754Val
4	Chr3:50137446-C-T	c.371C>T	p.Pro124Leu
5	Chr3:50150955-C-T	c.1595C>T	p.Pro532Leu

Table 5.1: Information on the five variants identified in infertile men in the gene RBM5 and their corresponding insert number for the experiments performed within this chapter. Variant 1 was identified in the whole exome patient-parent trio data and is *de novo*, the remaining five variants were all identified in infertile singletons and thus the inheritance model is unknown.

The experiments performed utilising the FAS minigene were performed in HeLa cells (Bonnal et al., 2008). Human Embryonic Kidney cells were chosen as the most suitable for this experiment due to the robust nature of the cell line as well as its rapid growth rate and ease of transfection. Benefits of using HEK293 cells include their ability to stably express desired genes, their high efficiency at producing large amounts of recombinant proteins, and most importantly HEK293 are highly amenable to transfection and can be transfected using a variety of chemical and physical methods (Tan et al., 2021; Thomas & Smart, 2005). The nature of this experiment was to determine whether mutations in *RBM5* change the splice patterns of the *FAS* minigene. At this stage, a testis specific cell line was not essential and

HEK293 cells presented the most suitable option. HEK293 cells were utilised for all transfection stages in this experiment with the RNA and protein being isolated from these cells. Both *FAS* and *RBM5* show relatively consistent levels of RNA transcription in both HeLa and HEK293 cells with *FAS* showing low expression levels in both (4.8 nTPM in HeLa, 6.3 nTPM in HEK293) and *RBM5* showing high expression levels in both (87.2 nTPM in HeLa, 31.2 nTPM in HEK293).

5.4 Results

5.4.1 RNA Expression of FAS Isoforms in HEK293 Cells

To investigate the effect on splicing of the five previously mentioned *RBM5* mutations, RNA was extracted from HEK293 cells transfected with *RBM5* mutant plasmids and the *FAS* minigene plasmid. The RNA was then processed via a one-step RT-PCR to amplify the two separate splice isoforms transcribed from the *FAS* minigene. The larger 'Isoform 1' including exons 5,6 and 7 can be seen as the upper band (+ exon 6) in Figure 5.4, and the smaller 'Isoform 2' produced when exon 6 is spliced out is represented by the lower band (- exon 6). Cells were co-transfected with both the *RBM5* expression vector and the *FAS* minigene on three separate occasions as shown in Figure 5.5. These results were then quantified using the inbuilt Qiaxcel software. PSI values were calculated for each variant and compared to the Wild Type (WT) and are shown in Table 5.2.

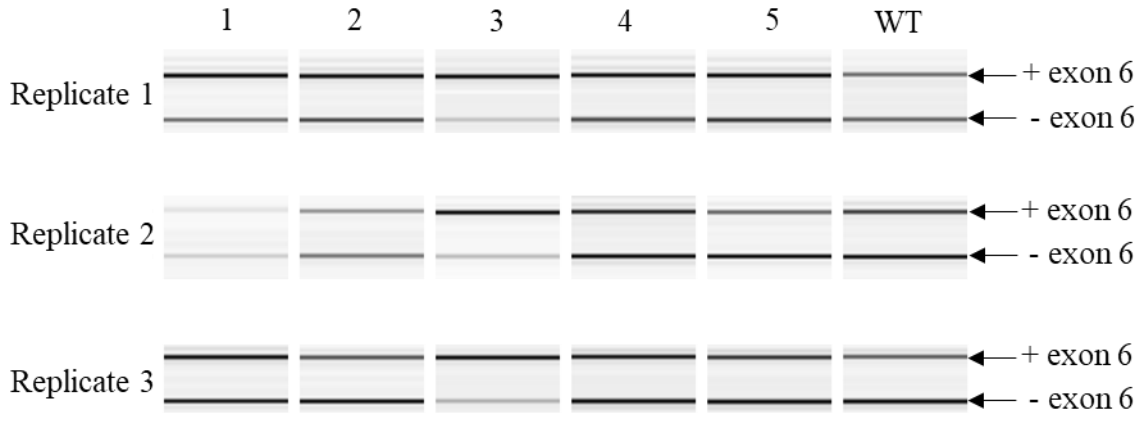


Figure 5.4: RT-PCR results showing RNA expression of Isoform 1 and 2 of FAS Minigene in HEK293 cells when transfected with various mutant RBM5 inserts. The upper band shown at around 180 bp represents the long isoform (Isoform 1) of FAS including exons 5, 6 and 7 of the minigene, the shorter isoform (Isoform 2) is represented by the band at around 115 bp. The stronger the band the greater the concentration of that isoform within the sample. The RT-PCR samples were run on the QIAxcel Advanced System (Qiagen) and concentration information for each isoform in each sample was also collected. The darker the band, the higher the concentration of the isoform.

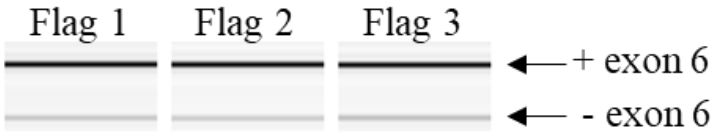


Figure 5.5: RT-PCR results showing RNA expression of Isoform 1 and 2 of FAS Minigene in HEK293 cells when transfected with empty flag vectors. Percentage splice inclusions for these three samples were 82.76 %, 83.05 % and 78.86 % respectively.

Variants	1	2	3	4	5	WT
Replicate 1	64.48	57.84	81.91	53.65	51.41	45.53
Replicate 2	35.71	39.66	78.60	45.38	36.45	42.89
Replicate 3	51.10	39.10	78.16	43.88	41.55	35.07
Standard Deviation	14.40	10.66	2.05	5.26	7.61	5.44
p-value	0.3788	0.4330	0.0037	0.0838	0.6854	-

Table 5.2: Percentage splice inclusion data of FAS exon 6 from RT-PCR results showing RNA expression of Isoform 1 and 2 of FAS Minigene in HEK293 cells when transfected with various mutant RBM5 inserts. The standard deviation and p-values were calculated using a two-tailed paired t-test comparing the data for each three replicates of a specific insert to the three data points for the Wild Type (WT) *RBM5* insert.

From Figure 5.4, it was clearly seen that across all three replicates of Variant 3, Isoform 1 was expressed at a higher concentration than Isoform 2. After quantification of the bands, it was found that, when the PSI values were compared to that of the Wild Type *RBM5* in a paired t-test, there was a statistically significant difference (p-value = 0.0037). This showed a significant shift towards splice inclusion in the cells expressing *RBM5* with a c.2260A>G variant existing in exon 24 of 25 of the gene compared to the level of splice inclusion occurring in cells expressing Wild Type *RBM5*. The three replicates of Variant 3 also were the most consistent (SD = 2.05 %). Interestingly the results for Variant 3 were very similar to the RT-PCR results

for the empty flag plasmids which were transfected into the same cells and contained no *RBM5* at all (Mean PSI = 81.56 %, Figure 5.5).

Whilst more difficult to visualise by eye, after quantification, Variant 4 showed consistently higher levels of splice inclusion of exon 6 across all replicates when compared to the WT replicates. When performing a paired t-test, however, it was found that although very close with a p-value of 0.0838, no statistical significance was found between Variant 4 and WT PSIs. The standard deviation for the Variant 4 RT-PCR highlighted slightly more variation between the replicates than seen in the Variant 3 replicates (SD = 5.26 %).

Variants 1, 2 and 5 were all found to show no statistically significant difference when PSI values were compared between these samples and the WT samples (p-value = 0.3788, 0.4330 and 0.6854 respectively). Interestingly, all three of these variants did however show increased levels of splice inclusion in replicates 1 and 3 when compared to the corresponding WT replicates. In replicate 2 for Variants 1, 2 and 5 there is a marked decrease in the PSI compared to WT replicate 2.

As well as a change when compared to the WT replicates, both Figure 5.4 and Table 5.2 highlight a level of variability across different repeats. The Variant 1 showed decreased splice inclusion in the replicate 2 with 35.71 % when compared to 64.48 % in replicate 1 and 51.10 % in replicate 3 (Standard deviation of 14.40 %). The Variant 2 repeats also showed varied results with Sets B and C showing lower levels of splice inclusion (39.66 % and 39.10 %) and Set A showing 57.84 % (Standard deviation of 10.66 %).

5.4.2 Protein Expression of RBM5 in HEK293 Cells

After the above analysis revealed a significant change in PSI of exon 6 in the cells co-transfected with *RBM5* Variant 3 and thus a greater concentration of splice isoform 1 of the *FAS* minigene, a Western Blot was performed. This was to identify any changes to the expression of RBM5 in this sample compared to the wild type which could explain the previously mentioned results. Figure 5.6 shows two bands per lane representing RBM5 protein expression. From this figure it was seen that the level of RBM5 protein expressed is significantly decreased in the three lanes loaded with the Variant 3 replicates. This is represented by the significantly fainter bands in these lanes. An error in loading was ruled out by monitoring equal expression of α-Tubulin which was detected as a loading control. It appears that whilst significantly depleted, very faint bands can be observed for RBM5 protein expression in all repeats of Variant 3 and thus protein expression is severely reduced but not completely absent.

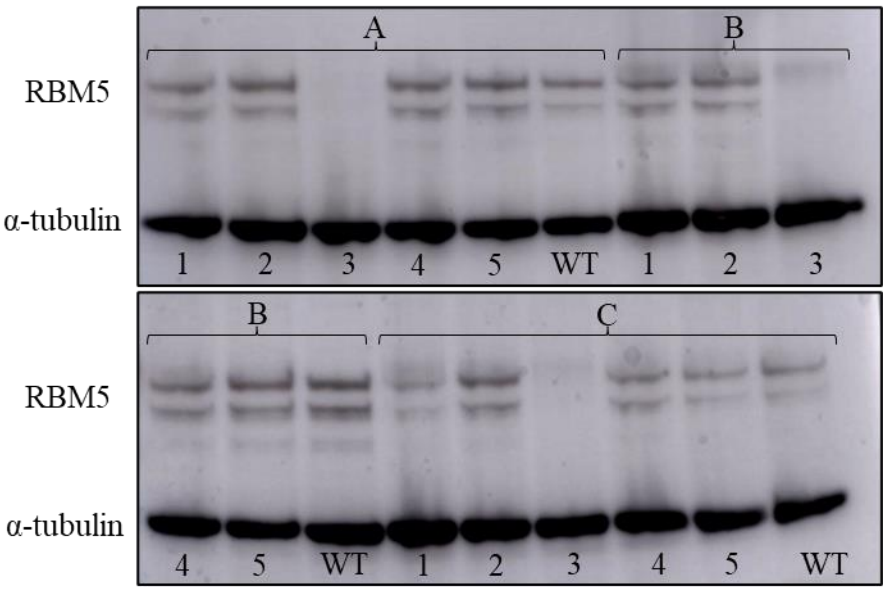


Figure 5.6: Western Blot showing protein expression of RBM5 in HEK293 cells co-transfected with mutated RBM5 and a FAS minigene. The RBM5 protein is represented by the dual bands on the western blot at a size of 92 kDa and α-tubulin was used as a loading control with a molecular weight of 50 kDA. Samples from replicate 1 are represented by "A", replicate 2 by "B" and replicate 3 by "C".

5.5 Discussion

The aim of the work in this chapter was to investigate the potential mechanism of the mRNA splicing pathway disruption as a cause of male infertility. After co-transfecting mutated *RBM5* inserts with a *FAS* minigene, one of the five *RBM5* variants (Variant 3) showed a statistically significant difference in FAS exon 6 splice inclusion when compared to the wild type RBM5 protein (p-value = 0.0037). One further variant (Variant 4) was close to achieving statistical significance with a p-value of 0.0838 and consistently higher PSI values in all three replicates compared to the WT. The remaining three variants showed no noteworthy results with variation across the replicates and high p-values.

RNA extracted from samples containing Variant 3, or c.2260A>G, showed a substantial increase in the concentration of the exon 6 inclusive isoform 1 of FAS (mean PSI of 79.56 %) compared to the WT (mean PSI of 41.16 %). Based on the PSI results in Table 5.2, WT RBM5 expression typically leads to exclusion of exon 6, however with the c.2260A>G almost 80 % of all FAS produced is exon 6 inclusive. Notably, the Variant 3 results closely resembled those achieved when transfecting the empty flag plasmid which itself had a mean PSI of 81.56 % (Figure 5.4, Table 5.2). It was confirmed, however, before transfection was performed that RBM5 variants had been correctly cloned into the p3XFLAG-CMV14 expression vector by

means of colony PCR and Sanger Sequencing (description in Chapter 2 Section 2.4.3). Western Blot analysis showed trace amounts of RBM5 protein expressed in these samples (Figure 5.6). These results suggest that the c.2260A>G variant had a significant effect on the translation of the *RBM5* gene into the protein itself leading to a dramatic under expression of RBM5. These results are consistent with previous literature in that reduced RBM5 protein levels lead to an increase in FAS exon 6 splicing inclusion. It has been previously reported that when levels of RBM5 are reduced, a greater percentage of the exon 6 inclusive isoform is produced, and overexpression of RBM5 leads to increased exon skipping. Thus, RBM5 causes skipping of *FAS* gene exon 6 (Bonnal et al., 2008).

The severity of the downregulation of the RBM5 protein, along with the large shift in splice inclusion of exon 6 of the *FAS* minigene leads to the suggestion that the variant itself causes protein instability. At a basic level, it can be seen that the position in which this mutation lies within RBM5 is highly conserved across over 99 species with a phyloP100 score of 9.025, indicating that changes made to the wild type allele at this position are much more likely to produce a damaging effect. The hypothesis that the introduction of the c.2260A>G variant at a highly conserved site effects the ability for the protein to remain stable would need to be further investigated. The p.Met754Val amino acid change seen here does not initially appear to be a severe one when looking at the properties of each amino acid with Methionine and Valine both being non-polar with hydrophobic side chains. However, Valine itself is a small aliphatic amino acid and Methionine is regarded as a long non-aliphatic amino acid. The SIFT score (Vaser et al., 2016), which predicts whether an amino acid substitution is likely to affect protein function, was the most severe value possible for this specific variant (SIFT = 0) and categorised the variant as pathogenic. Previous literature has shown disease states resulting from a single base substitution resulting in a Met-Val change, specifically in the AD disease Hyperkaliaemic periodic paralysis (Rojas et al., 1991).

The positioning of this mutation in the middle of the penultimate exon within the gene leads to speculation as to how a late-stage mutation can cause the severe under-expression of the protein seen in the Western Blot analysis. One such explanation for this is the formation of a degron at the position of the variant. A degron is a protein degradation signal which are often short linear motifs embedded within the protein sequence (Mészáros et al., 2017). The presence of a degron towards the terminal end of a protein has been shown to potentially contribute to Spinal Muscular Atrophy (Cho & Dreyfuss, 2010) . In this case, the majority of the time (ca. 80 %) *SMN2* produces an unstable transcript lacking the exon 7 of the gene and when this exon is missing two components of the transcript, the YG box and the C-terminal EMLA sequence,

form a degron (Seo et al., 2016). It is possible that a similar mechanism is occurring in the *RBM5* gene containing Variant 3 with a novel degradation signal being introduced, causing the destabilisation of the protein, and thus explaining the Western Blot analysis. If testis biopsy samples exist for the patient presenting with this RBM5 variant, it would be of great interest to determine the levels of RBM5 protein expression to confirm whether the downregulation is as severe in vivo. In the patient it would be expected that the expression of the RBM5 protein would be reduced however not as severe as seen in the cell culture due to the heterozygous nature of the variant within the patient.

Whilst four of the five mutations tested do not appear to have a significant effect on the splicing of *FAS*, the potential remains for these mutations to be causing the infertility observed in these patients. With the level of variation between the PSI values across samples 1,2 and 5 (SD = 14.40 %, 10.66 % and 7.61 % respectively) and the fact that Variant 4 almost showed significance (p-value = 0.0838), the first logical step would be to repeat these experiments with a greater level of stringency. The replicates (1 - 3) for this experiment were biological replicates, with each replicate performed on a separate passage of the same HEK293 cell line. A limitation of the initial experimental design was a lack of technical replicates for each biological replicate and the experiment could have benefited from the accuracy that these technical replicates would have provided. These would have been particularly important in the case of the second set of biological replicates where the data for Variants 1, 2 and 5 does not agree with that for Replicates 1 and 3. In Variants 1, 2 and 5 both replicates 1 and 3 show increased levels of splice inclusion between the variant and the WT, this however is not mirrored in Replicate 2 where for all 3 variants, a decrease is seen in splice inclusion when compared to the WT. The use of technical replicates would have allowed clarification as to whether the change in pattern seen in the second replicate set for these three variants was true or likely due to experimental error.

Another important component that this experiment would have benefitted from would have been the inclusion of a positive control to confirm whether the use of the *FAS* minigene as an *RBM5* splicing target was appropriate to reflect the expected male infertility phenotype resulting from damaging *RBM5* mutations. It is already known from work performed in mice that the introduction of a specific *Rbm5* point mutation (p.Arg263Pro) causes male infertility and effects the splicing capability of *Rbm5* for a number of testis specific proteins (O'Bryan et al., 2013). The addition of a g-block containing this specific mutation in this experiment would have been valuable, providing the ability to determine whether the use of the *FAS* minigene was appropriate for the human male infertility phenotype being investigated depending on any splicing changes observed.

Whilst the work in this chapter aimed to determine the effect of *RBM5* mutations identified in Chapter 4 on the gene's capabilities to regulate splicing of the *FAS* minigene, the rationale behind it is that mutations to *RBM5* could potentially be causing infertility in several infertile patients. There are multiple different methods by which this could be assessed within the scope of infertility and the process of spermatogenesis itself. Initially, if no significance was seen after repeating the experiments in the manner mentioned above for these variants, an alternative route to investigate the effect of these other mutations would be to utilise a different target of *RBM5* as a minigene. A target gene which plays a specific role in the process of spermatogenesis would be preferable, rather than *FAS* which has been shown to be involved specifically with programmed cell death (Scott et al., 2009). Such a potential target would be the *ASB1* gene which has been implicated in the process of spermatogenesis following studies with *Asb1* knockout mice which present with hypospermatogenesis (Kile et al., 2001). Both *ASB1* and *RBM5* show enhanced RNA single cell specificity to early spermatocytes and late spermatids respectively and are present throughout the process of spermatogenesis in all testis specific cell types in the Human Protein Atlas (Uhlén et al., 2015).

In 2013, it was shown that mice carrying a specific missense mutation (p.Arg263Pro) in the gene *RBM5* presented with aberrant splicing of the gene *Asb1* with a greater occurrence of exon skipping implying a disruption to the pre-mRNA splicing of this target due to RBM5 functional disruption (O'Bryan et al., 2013). This mutation lies within the RNA recognition motif (RRM) 2 domain of the gene, a domain untouched by the five variants found in the infertile patient cohort described in this chapter. Interestingly however, Variant 4 which was close to significant, lies within the RRM1 domain of the gene. With the presence of a known infertile mouse model already affected by *Rbm5* mutations, it would be possible and justified to perform in vivo studies within mouse models in future. This could be performed utilising the methodology previously described in the 2013 paper with the introduction of our own potentially causative mutations into the model. In vivo studies such as these are relatively time consuming and significantly more costly than working in in vitro models. Before moving on to studies like these, acquiring repeated successful results from in vitro studies would be advantageous. It would, be of interest to investigate the effect of the damaging c.2260A>G variant to observe whether in mice this does also produce a near total depletion of the *RBM5* protein.

Whilst studies have been performed in mice displaying disruption to fertility when *Rbm5* is mutated, these do not exist for human samples. The experiment that would be of most interest to back up the work in this chapter would be to assess *RBM5* expression levels, as well as other known key spermatogenic genes known to be regulated by RBM5, within testis biopsy samples

from the infertile patients with *RBM5* mutations. This would then require a comparative sample cohort of healthy testis biopsy tissue as well as biopsy material from other infertile males with no known RBM5 mutation. In the case of this experiment, the initial patient with a DNM in RBM5 did not have any testis biopsy material taken hence why this experiment was not conducted. It may be possible in future to revisit this idea if additional patients are identified with potential damaging DNMs in the gene *RBM5*, strengthening the evidence for a causative link between these mutations and the infertility phenotype.

The work performed in this chapter provides a start to functional studies into the effects of damaging mutations to the function of RBM5. After analysing the results, it is clear that the initial experiment had limitations and would need repeating with additional improvements to re-assess the validity of the results. If this work proved successful after improvement of the experimental design, the method could be applied to other the mutations found in genes involved in mRNA splicing. New targets for minigene creation would need to be identified for the genes *CDC5L*, *CWC27*, *HNRNPL* and *U2AF2*, however once this step has been completed, the experimental analysis would be as followed in this chapter.

Chapter 6: Discussion and Concluding Remarks

6.1 Introduction

Male infertility (MI) is a relatively common clinical condition with around 7 % of all males worldwide suffering from some form of infertility (Krausz & Antoni Riera-Escamilla, 2018). There are currently a large number of known causes for MI, both genetic and non-genetic (Kasak & Laan, 2021; Xue et al., 2012; Michalakis et al., 2013). The condition itself is extremely genetically heterogeneous which is unsurprising given the complexity of the process of spermatogenesis and the vast number of genes involved in its regulation (Neto et al., 2016). This complexity results in a wide variety of infertility phenotypes. Genetic factors are thought to play an important role in MI with the percentage of genetic causes increasing as the sperm output of the patient decreases (25 % in Azoospermia) (Krausz, 2011). Whilst some genetic causes, including Klinefelter syndrome, AZF region microdeletions and recessive *CFTR* mutations are well recorded and studied, a large percentage of all sporadic male infertility cases remain idiopathic. Recent studies within the field of MI highlight how between 40 % – 70 % of all MI cases presented to clinics are idiopathic (Punab et al., 2016; Houston et al., 2021; Kumar Mahat & Arora, 2016). With the large number of idiopathic cases seen in MI, it is logical to infer that there are numerous genes and mutations causing the infertility in these patients that have not yet been discovered. Conditions such as Intellectual Disability (ID), which show a similar reduction in reproductive fitness to MI, have found that DNMs play a large role in causing the disease. The introduction of NGS techniques such as whole exome sequencing (WES) and whole genome sequencing (WGS) have significantly advanced research in ID, with patient-parent trio NGS studies allowing the discovery of several disease genes (Gilissen et al., 2014; Vissers et al., 2016; Veltman & Brunner, 2012).

Investigating the *de novo* paradigm for MI has been the basis of my research as described in this thesis. Across the Chapters I have utilised different sequencing techniques in order to identify DNMs in potential novel candidate genes for MI. In Chapter 4 I successfully identified 145 rare protein coding DNMs in a cohort of 185 patient parent trios with 29 of these classified as likely causative of the infertility phenotype in these patients, and a further 50 remaining unclear with limited gene level evidence. After a systematic analysis of an additional 2,506 infertile singleton males, six additional males were identified with predicted pathogenic heterozygous missense mutations in the gene *RBM5*, in addition to the original patient with a pathogenic DNM in the same gene. Functional studies investigating the damaging effects of *RBM5* mutations on its function as pre-mRNA splicing factor showed promising results, indicating a potential pathway to further investigate in relation to autosomal dominant (AD) male infertility.

6.2 Next Generation Sequencing (NGS) as a Tool for Disease Gene Discovery in Male Infertility

6.2.1 NGS Use in Male Infertility Research

Over the last decade, NGS technologies have advanced considerably with a significant improvement in sequencing output and a drastic reduction in costs (Xavier et al., 2021). These technologies range from targeted panel sequencing of multiple genes to whole exome and whole genome sequencing. These NGS approaches have allowed for considerable amounts of data to be assessed in large cohorts of patients at one given time, in contrast to traditional technologies such as Sanger Sequencing and Karyotyping (Medeiros et al., 2022; van Dijk et al., 2014). It is no surprise that these technologies are being applied to study an increasing number of different diseases and disorders with the aim of identifying genetic causes to rapidly provide a genetic diagnostic in known disease genes or discovering new genetic causes and identifying novel candidate disease genes. For disorders such as ID, where the genetic heterogeneity and reduced reproductive fitness mirrors that of MI, a large number of genes have been identified using NGS techniques in both singleton and trio cohorts (Lelieveld et al., 2016; Kaplanis et al., 2020; Gilissen et al., 2014). These studies have proven to be successful with hundreds of genes now identified as causative ID genes, and high genetic diagnostic levels (Kaplanis et al., 2020; Vissers et al., 2016). A recent paper highlights how in the year 2019, 62 % of all genetic studies published on the subject of MI included some form of NGS based technique (Xavier et al., 2021). The implementation of NGS in MI research up to now however has mostly been successful in identifying new recessive disease genes for rarer, more well-defined forms of infertility such as Multiple Morphological Abnormalities of the Flagella (MMAF) (Lin Li et al., 2017; Zhu et al., 2018) and Hypogonadotropic Hypogonadism (Butz et al., 2021; Cangiano et al., 2021). Due to the large amount of heterogeneity leading to both Oligozoospermia and NOA giving rise to a wide variety of phenotypes, disease gene discovery for these forms of MI has not progressed as much.

6.2.2 Targeted Panel Sequencing; a Useful Clinical Diagnostic Tool but Inefficient for Disease Gene Discovery

The first analyses I performed using an NGS technique is reported in Chapter 3, where I designed and optimised a targeted smMIP panel to sequence a total of 54 genes which had been previously identified as containing DNMs. This smMIP panel was used to sequence 75 infertile males and their unaffected parents. After processing of the sequencing data and filtration of all initially identified DNMs, no rare DNMs were identified in any of the 54 targeted genes within

the cohort. There were likely a number of issues which contributed to the lack of results in this case. The first, and likely most important issue is the size of the gene pool as well as the genes targeted in the first place. It is well established that only one to two DNMs occur in the coding region of each person's DNA and with around 20,000 protein coding genes found in the human genome (Oud et al., 2017; Pertea et al., 2018). Taking this into account, it is not surprising that no DNMs were identified in such a small gene pool representing only 0.27 % of all protein coding genes. When compared to the previously mentioned study by Oud et al, where 1,112 infertile males were screened for all mutation types in 107 causal and candidate DNM genes, the number of patients assessed by using smMIPs in this thesis were drastically reduced (7 %) (Oud et al., 2017). In the Oud et al. study, only one heterozygous variant which may, or may not, have been *de novo* was identified in the gene *SYCP3*, a gene which has been previously identified as a monogenic cause of MI and only a further 16 patients received a genetic diagnosis (diagnostic yield = 1.5 %).

At an experimental level, despite two rounds of optimisation to ensure all smMIPs were performing within an acceptable range, only 87 % of all the smMIPs in the final run performed to an acceptable level. When looking at the number of smMIPs achieving 10 unique reads across the samples, 92 % fell within this range which is comparable to work published in a similar smMIP study performed on infertile males (Oud et al., 2017). These statistics would make the work seem successful. However, when looking in depth into individual genes, only six genes (11 %) within the panel had no smMIPs underperforming whereas a number of genes had regions which were covered by underperforming smMIPs across all samples, including the genes *APC2* (48 % of smMIPs underperforming), *CD81* (50 % of smMIPs underperforming) and *DGKZ* (28 % of smMIPs underperforming). This led to regions within these genes where the sequencing quality was not high enough to identify or call any potential DNMs. The only way to improve upon this would be to implement additional rounds of optimisation of the smMIPs within the pool, adding in supplementary smMIPs at 5x and 20x concentrations. This would hopefully help improve the coverage and performance of the underperforming smMIPs within the pool by reducing the amount of some smMIPs which were potentially overperforming and increasing the concentration of those which were drastically underperforming.

From this work it would appear that whilst the smMIP panel has great potential for identifying DNMs in a tailored group of genes, it should not be used in the early stages of disease gene discovery - but rather after a number of genes have been confidently linked to the disorder first. smMIP panels have been shown to be diagnostically useful in both colorectal cancers and *BRCA*

screening in clinical settings (Gallon et al., 2020; Neveling et al., 2017), but isolated MI has a long way to go before these panels are likely to be useful. Given the large genetic heterogeneity of MI and the lack of recurrently mutated genes identified, performing targeted sequencing in such a limited pool of genes in a relatively small cohort was unlikely to highlight any recurrent DNMs. At this early stage in MI candidate gene discovery, particularly for dominant causes of MI, unbiased, all-gene inclusive investigations are needed before we can move on to systematic targeted approaches.

Whilst isolated MI should not be targeted with this smMIP approach for the time being, rarer MI phenotypes such as MMAF have seen a dramatic rise in the number of associated disease genes in the last five years from WES discovery studies. One such study from 2020 consolidated a list of 18 novel MMAF linked genes (Touré et al., 2021). With these 18 genes accounting for between 30 % – 60 % of all MMAF cases across the discovery cohorts, it is possible that a smMIP panel could be applied in a clinical diagnostic setting, identifying variants in these genes in MMAF patients. As more genes arise from additional NGS studies in MI cohorts, additional MMAF genes could be added easily in order to perfect the panel providing a cost effective and efficient method for MMAF diagnostics.

6.2.3 Whole Exome Sequencing (WES) as an Unbiased Approach for Novel Male Infertility Gene Discovery

WES is now being used widely across the field of MI with studies occurring in large cohorts of infertile males across the globe. Research by groups such as the Genetics of Male Infertility Initiative (GEMINI) and Male Reproductive Genomics (MERGE) studies have identified a number of novel male infertility candidate genes utilising WES approaches in their cohorts of infertile males (Hardy et al., 2021; Nagirnaja et al., 2021; Salas-Huetos et al., 2021; Wyrwoll et al., 2020). In Chapter 4 of this thesis, I demonstrate that WES is a useful tool in identifying DNMs in a cohort of 185 idiopathic infertile patient-parent trios. All these patients had previously undergone multiple different clinical assessments to identify their potential cause of infertility including a physical examination, semen analysis, hormone level evaluation and genetic screening including Karyotyping and AZF deletion detection. The next logical step in investigating the patients who remained idiopathic was to look for any potentially pathogenic mutations which could be causing their infertility.

In total, 29 causative and 50 unclear rare protein coding DNMs were highlighted. It should be noted that none of the genes identified during this experiment was found containing a DNM in more than one patient and none of the genes containing these rare DNMs were previously

highlighted as causal MI genes. The lack of recurrence of mutations in these genes makes it impossible to classify any gene as a "true" MI gene. Identifying additional patients showing variation in these genes is the next step needed to start highlighting and determining true *de novo* male infertility genes. Whilst trio sequencing has been shown to provide a much higher diagnostic rate than singletons alone (Smedley et al., 2021), gathering a large cohort of not only infertile males but their parents too is logistically challenging, particularly for this disorder. Unlike with ID, where the presentation age of the disorder is likely as a child, males are unlikely to know they are infertile until they are much older which provides challenges in acquiring parental samples due to advanced parental age, lack of contact with parents or potentially the death of parents. Additionally, infertility is often regarded as a female issue and when a male factor is identified, this can be seen as shameful by many patients who then do not wish to share this information with parents or other family members. These factors mean that obtaining parental samples for MI is much more difficult than for other disorders, and this can be seen reflected in the lack of large-scale studies performed on infertile patient-parent trios so far. The work presented in Chapter 4 represents the largest published trio based NGS study to date regarding MI with only one smaller pilot study in 13 trios having been previously published (Hodžić et al., 2021). In order to reach the discovery levels seen in ID, the need for recruitment of larger cohorts is a necessity. In ID has been suggested that up to 350,000 trios need to be sequenced in order to have an 80 % chance of detecting all haplo-insufficient genes causing this disorder (Kaplanis et al., 2020). Whilst this number is a far cry from the modest 185 trios described in this thesis, I believe my work highlights the advantages of utilising WES for infertile patients in order to identify DNMs.

In an attempt to expand the discovery effort within the UK, collaborations have been set up between the Newcastle male infertility group and hospitals in both Manchester and Sheffield for both singleton and patient-parent trio recruitment as well as international collaborations with groups in India, Iran, and Germany. With the inclusion of these extra sites and cohorts as well as our continued collaboration with RadboudUMC, it is hoped we can build up a larger database of WES data to increase the power of our disease gene discovery, with the aim of identifying recurrently mutated genes in multiple patients. Additional international collaborations have also been set up via consortia such as the International Male Infertility Genomics Consortium (IMIGC) (http://www.imigc.org/). The IMIGC are currently working towards this aim of gathering large cohorts of infertile patients in order improve the statistical and bioinformatic analyses of genomic data as well as identifying novel candidate genes for MI. In Chapter 4, the presence of rare pathogenic mutations in genes containing rare protein coding DNMs was

investigated in WES data from 2,506 infertile males, originating from both the IMIGC and the Geisinger-Regeneron DiscovEHR collaboration (Dewey et al., 2016). Without the data provided through these collaborations, the discovery of additional patients with potentially pathogenic variants in the gene *RBM5* would not have been possible.

Widespread implementation of WES as a routine diagnostic test in global infertility clinics and health care services has the potential to not only identify recurrent mutations in known or implicated MI genes, but also continue the discovery of novel linked genes. With the nature of WES data, retrospective diagnoses are entirely possible as a growing list of genes are linked to MI, therefore, it is important to emphasise that both diagnostics and research go hand in hand.

6.2.4 Genome Sequencing for Future Advances in Genetic Studies

Chapters 3 and 4 of this thesis focus on two different NGS techniques, targeted panel sequencing and whole exome sequencing, both of which in this case only assessed the coding regions of DNA within the patients. I have previously established that within this coding region, a person has on average, one to two DNMs, however there are around 44 - 82 DNMs found in a person's entire genome (Acuna-Hidalgo et al., 2016). By excluding non-coding regions (approximately 98 % of the human genome) (The ENCODE Project Consortium, 2012), it is possible that a large number of causative DNMs are being missed. A 2022 study in 242 autism spectrum disorder patients and their families identified an enrichment for non-coding DNMs, causing disrupted chromatin interactions leading to altered expression of nearby genes involved in early neural development (Kim et al., 2022). This model could be applied to MI whereby mutations in non-coding regions alter the expression of one or a number of the thousands of genes involved in spermatogenesis. WGS not only allows for the investigation of non-coding regions but has also been shown to provide better overall sequencing coverage of exonic regions than WES (Meienberg et al., 2016). WGS is also particularly useful for the detection of structural variants (SV) and repeat expansions, with the continuous coverage of the genome allowing for a more reliable calling of SVs with greater confidence than WES data (Gilissen et al., 2014; Lelieveld et al., 2015; Rajan-Babu et al., 2021).

A recent study performed by Genomics England as a part of the 100,000 genomes project showed how the incorporation of WGS into diagnostic tests for patients with idiopathic rare diseases lead to an increase of diagnostic yield (Smedley et al., 2021). The study found that, after WGS analysis, the diagnostic yield for specific conditions such as ID, hearing disorders and vision disorders ranged from 40 % - 55 % and across all disorders with a likely monogenic

cause, 30 %. The results of this study led to 25 % of all those with a novel diagnosis receiving immediate clinical treatment or counselling. The large increase in diagnostic yield and actionable findings demonstrates the invaluable nature of implementing NGS sequencing as a routine test for those with idiopathic rare disorders. An important point to make is that whilst this thesis focuses specifically on DNMs in male infertility, introducing NGS into routine clinical assessments would not only allow the identification of these rare dominant causes of infertility, but also variants of all inheritance models which may provide further insight into different inheritance models for male infertility.

Currently, with dominant disease gene discovery still at its early stages for MI, the lower cost and lack of research into non-coding variation in MI means that WES is likely to remain the more common NGS technique used for disease gene identification at present . The significant increase in discovery potential of WGS should not be ignored and is likely to become used more in mainstream research as costs decrease. Related to this, a pilot study is currently ongoing to re-analyse all 185 trios assessed in Chapter 4 of this thesis utilising WGS in order to further causative gene discovery in MI cases for all inheritance models, including SVs and repeat expansions. From this data it will be possible to systematically compare the two techniques and better suggest which method is more appropriate and effective at identifying mutations of all inheritance patterns in males suffering from severe idiopathic MI based on the number of relevant findings as well as the costs associated with both technologies.

6.3 *De novo* Mutations as a Model for Dominant Disease in Male Infertility

The majority of research into MI and its associated genes has been focussed on inherited forms of the disorder, with autosomal recessive (AR) inheritance being the most widely researched monogenic inheritance pattern (Houston et al., 2021). According to Oud et al., around 66 % of gene-disease relationships highlighted in MI with at least a moderate level of evidence can be attributed to AR inheritance (n = 16), whilst only 17 % showed autosomal dominant (AD) inheritance (n = 4) (Oud et al., 2019). The discovery of many of these AR disease-gene relationships are due to research in consanguineous families, (Yıldırım et al., 2018; Ma et al., 2019) and whilst they may explain some familial forms of MI, they are likely to be far less common in the outbred population due to selection pressure. As previously stated, MI remains a relatively common condition despite it often leading to drastically reduced reproductive fitness. So clearly, investigation into other inheritance models may be the key to diagnosing a number of idiopathic cases. AD forms of MI are severely underrepresented, however, maternally inherited and *de novo* mutations are two strong inheritance models to investigate in

idiopathic MI cases. DNMs make an excellent model for investigation due to their ability to avoid typical evolutionary selection that would occur with inherited variation, thus making these mutations potentially more deleterious (Hodžić et al., 2021; Veltman & Brunner, 2012). In patients suffering from conditions such as Developmental Delay (DD), DNMs have been proved to be causative in around 42 % of all cases (Deciphering Developmental Disorders Study, 2017). Like DD, MI has a high mutational target size with over 2,000 genes involved in the process of spermatogenesis alone meaning that the likelihood of a DNM occurring in one of these genes is much higher than in a disease with a smaller number of causal genes (Veltman & Brunner, 2012). As highlighted earlier, chromosomal abnormalities underlying infertility such as Klinefelter's and Y chromosome microdeletions occur *de novo*, validating in part the role of this genetic model for MI.

Chapter 4 of this thesis provided vital evidence that *de novo* point mutations are likely to play a role in MI. A total of 192 rare DNMs were identified across the cohort (n = 185) giving a DNM rate of 1.04 per patient, in line with the one to two DNM determined to be present in the average person's coding region (Acuna-Hidalgo et al., 2016). At this stage, the DNMs had only been assessed relating to their rarity in the general population and whether they affected the amino acid structure of the given protein, therefore unbiased analyses were able to be performed to assess any patterns found in the genes in which the DNMs lay. This unbiased approach highlighted a significant enrichment for Loss of Function (LoF) DNMs occurring in genes which were more intolerant to LoF mutations when compared to a cohort of fertile controls (p value = 1.00×10^{-5}). The idea that LoF DNMs were found to occur more frequently in genes where they are not tolerated in infertile males and not in controls highlights an important pattern for a potential link between MI and DNMs. Interestingly, a paper published in 2022 investigated the inheritance patterns of 605 LoF intolerant genes (pLI = 1), 217 of which were associated with a mendelian disease (Fabre & Mancini, 2022). It was seen that of the 217 LoF intolerant genes with a known inheritance pattern, 69.1 % were AD, 12.4 % showed X linked inheritance followed by 12 % AR and the remaining 6.5 % had mixed inheritance. Whilst the authors determined no definitive preferential mode of inheritance being associated with highly constrained genes, it is clear from this data that a significant proportion of all known disease genes with high intolerance to LoF mutations showed an AD inheritance pattern.

After determining the likely pathogenicity of all 145 DNMs, 84 were found to be pathogenic at a variant level. From this a further enrichment was highlighted for likely pathogenic missense DNMs to be found in genes which are intolerant to missense mutations when compared to likely benign missense DNMs found in the patients (p value = 5.01×10^{-4}). This finding once again

indicated that, in these patients suffering from severe male infertility, more damaging mutations are occurring in genes which are less tolerant to these types of mutations. This furthermore highlights how, whilst there are over 20,000 genes expressed in the human genome, these infertile patients are showing clustering of DNMs in genes which are intolerant to these mutation types. A pathway analysis of the genes containing pathogenic DNM types showed an enrichment of interactions between these genes (p value = 2.35×10^{-2}), signifying common biological pathways in which the DNMs occurred. Two pathways in particular, the pre-mRNA splicing pathway and mitotic cell cycle, were shown to contain a number of the genes with predicted pathogenic DNMs. This finding was incredibly interesting, as both recurrently affected pathways identified are essential for the normal function of spermatogenesis within males (Gazvani et al., 2000; Song et al., 2020). High levels of alternative splicing and cellular division occur within the testis, especially during spermatogenesis, and identifying several likely pathogenic DNMs in genes within these pathways provides further evidence that DNMs likely play a role in male infertility (Naro et al., 2021).

For successful investigation into causative *de novo* or maternally inherited variants in MI in the future, there is a dire need for replication studies in large cohorts to assess the lists of current candidate genes. With 79 Potentially Causative and Unclear variants in as many genes identified in this cohort of only 185 trios, it is plausible that in larger cohorts considerably more disease-linked genes will be identified for the first time. The only way these can be confirmed as true infertility genes is via the identification of recurrently mutated genes and so, as mentioned in section 6.2.4, diagnostics and research must work in tandem to recruit and sequence larger cohorts of infertile males and their parents in particular for *de novo* gene investigation. Even with the increase of infertile males sequenced, there will likely always be a deficit in the number of parental samples needed for recurrent DNM identification in all potential dominant MI genes. In cases where parental samples are unavailable, it would be interesting to determine if there was a way to identify whether these variants occurred *de novo* or were inherited. The development and introduction of a tool which could analyse all available data on *de novo* mutations across a wide variety of disorders in order to identify patterns seen when DNMs are present and thus predict the likelihood of heterozygous variants being *de novo* in origin would be invaluable. A DNM likelihood score, much like those used to determine whether a variant is likely to be pathogenic or a gene is intolerant to certain mutation types, would mean the struggle for acquisition of parental samples would be less detrimental to the progress of investigations into *de novo* causes of infertility. In Chapter 4 of this thesis, an enrichment was found for LoF DNMs mutations in LoF intolerant genes, as well as potentially pathogenic missense DNMs in

missense intolerant genes. It would be interesting to see if this was found consistently in multiple patient-parent trio cohorts and, if so, whether genes associated with *de novo* mutations across multiple disease types show this increased level of intolerance in comparison to other inheritance models. If this was found to be the case, it could be easier to predict, to a degree of certainty, whether a potentially pathogenic mutation was likely to be *de novo* or not in the absence of parental samples.

6.4 RBM5; An Investigation into the Pre-mRNA Splicing Pathway as an Essential Male Fertility Pathway

The work presented in Chapter 4 provided no genes with recurrent mutations in the original 185 patients. Interestingly, however, after analysing a cohort of over 2,500 infertile singletons for heterozygous mutations in all 145 genes containing rare, protein altering DNMs, six additional patients were identified with likely pathogenic heterozygous missense mutations in the gene *RBM5*. In total, one likely causative missense DNM and six likely causative heterozygous missense mutations were identified in *RBM5*, a gene known to play a role in the pre-mRNA splicing pathway. *RBM5* was found to be one of the genes highly intolerant to missense mutations which further increased the interest in this gene. Whilst the additional mutations found in the replication cohort could not be determined as *de novo* or inherited, the fact that seven mutations were identified in idiopathic infertile males and non were present in a control cohort of infertile males (n = 11,587) made this gene the primary candidate for further investigation.

Previous studies into *RBM5* highlighted this gene as an essential regulator of haploid male germ cell pre-mRNA splicing and have shown that when specific point mutations were introduced into *Rbm5* in mice, a resulting azoospermic phenotype was observed (O'Bryan et al., 2013). It was found in this study that the *R263P* point mutation led to shifts in levels of *Rbm5* specific splice targets within the mouse testis, effecting multiple different pathways within the mice leading to their sterility. The evidence provided by this study, as well as the presence of seven different heterozygous mutations in this missense intolerant gene, lead to the functional work presented in Chapter 5 of this thesis. Using a *FAS* minigene assay, a gene known to be regulated by *RBM5*, I assessed the effect of five of the likely causative mutations - including both the DNM and four mutations identified in the replication cohort. The FAS minigene used in these experiments was provided by Professor Juan Valcarcel Juarez and the work performed in Chapter 5 was based around the findings of this paper where it is stated that increased levels of *RBM5* cause increased levels of exon 6 skipping and thus reduced levels lead to increased levels

of the exon 6 inclusive FAS isoform (Bonnal et al., 2008). If the likely causative DNMs had any effect on *RBM5* and its function, it would have been expected that a large increase in the levels of the exon 6 inclusive isoform would be seen in all cases. However, this was only seen for one variant of the five tested. The tests performed included co-transfection of plasmids containing either mutant *RBM5* sequence or the wild type (WT) into HEK293 cells, along with the *FAS* minigene. RNA and protein were extracted from the cells after a day and RT-PCR and Western Blotting were performed to assess the *FAS* RNA levels and the RBM5 protein expression. Variant 3 or c.2260A>G showed a statistically significant increase in splice inclusion when compared to the control wild type *RBM5,* with almost 80 % of the FAS produced including exon 6 (p-value = 0.0037). Interestingly, Western Blot analysis showed nearly zero expression of the RBM5 protein extracted from variant 3 transfected cells. Whilst this provided insight into the severely reduced levels of splicing occurring, due to the lack of *RBM5* expressed, it did not fit with the initial hypothesis that the variants introduced would affect the function of *RBM5*, rather that this particular variant appears to inhibit the expression of the gene altogether. One explanation for this could potentially be the introduction of a degradation signal, or degron, which destabilises the protein and thus leads to reduced expression (Mészáros et al., 2017). The fact this mutation occurs in a heterozygous state in the patient leads to the potential argument that the gene may be haploinsufficient, with one mutated copy being enough to cause an infertility phenotype. In order to determine the full effect of this mutation on infertile males, heterozygous mutations would need to be assessed. It has already been seen in mice that the specific point mutation that causes infertility, has no effect in heterozygous mice. This does not however mean that the gene is not haploinsufficient in humans (Elsea & Lucas, 2002).

A further variant showed close to significant levels of splice inclusion with a p value of 0.0838 with consistently higher PSI values than seen in the corresponding WT samples. The remaining three variants however showed no significance at all. When looking at the RT-PCR data of these samples, there was a lack of consistency across the samples with some showing increased levels of splicing compared to the WT and others, decreased with standard deviations of (14.4 %, 10.7 % and 7.6 % in across the three replicates for variants 1, 2 and 5. It is clear that the work performed in this Chapter would benefit significantly from repeat studies, to more accurately determine whether all five variants did have an effect on *RBM5*'s splicing abilities. Moreover, it is highly likely that this work will be repeated and continued in future within my research group.

The use of the *FAS* minigene in these experiments was chosen because was a well-known *RBM5* splicing target which had already undergone previous testing. If I were to investigate these

variants further, I believe it would be valuable to be able to test the effect of these variants on *RBM5*'s testis specific targets such as the gene *ASB1* which has been proven to show altered exon splicing levels when *RBM5* is mutated (O'Bryan et al., 2013). It may be the case that whilst these specific mutations do not disturb FAS, a cell surface death receptor, they may affect genes specifically involved in spermatogenesis. I would also recommend applying similar experimental design to studying the other four DNM containing pre-mRNA splicing genes; *CDCL5*, *CWC27*, *HNRNPL* and *U2AF2* as this could lead to a solid confirmation of disruption to the pre-mRNA splicing pathway leading to sporadic MI. An experiment I would have liked to perform in these patients would have been to assess the protein and RNA expression of *RBM5* and its splicing targets in vitro using testis biopsies from the infertile males presenting with *de novo* and heterozygous mutations. This however was not possible due to a lack of tissue samples from these males but would be of great benefit for further studies in these patients. The introduction of multi-omics has helped improve understanding of a number of different cancers, giving insights into the pathways which are altered during a disease state (Montagne et al., 2023; Xiao et al., 2022). The use of transcriptomics in particular could be of great use in idiopathic MI to identify changes that occur during the process of spermatogenesis in both healthy and infertile males (Hermann et al., 2018). Currently, variants in genes involved in the pre-MRNA splicing have only been identified as enriched in dominant MI. No such enrichment has been found in MI cases of different inheritance models. In recessive forms of MI, enrichments of mutations in other specific pathways essential for spermatogenesis to occur successfully have been identified. piRNAs are small non-coding RNAs essential for survival of the germ cell pool and are highly enriched within the testis (Nagirnaja et al., 2021). Biallelic mutations in a number of genes involved in the piRNA biogenesis pathway have recently been identified in infertile males suffering from NOA. (Wyrwoll et al., 2022; Nagirnaja et al., 2022)

With the significant amount of alternative splicing that takes place during spermatogenesis (Yeo et al., 2004) and previous research implicating disruption to mRNA splicing genes in MI (Iwamori et al., 2016; Hayakawa et al., 2010), I believe the investigations into the gene *RBM5* and the other four pre-mRNA splicing genes to be essential for a further understanding of the aetiology of MI.

6.5 Future Implications of Male Infertility Disease Gene Discovery

6.5.1 Implications for Assisted Reproductive Techniques

As more genetic causes of infertility are discovered, the need for genetic counselling in these patients will increase significantly. In dominant causes of disease, there is a 50 % chance of

offspring inheriting the damaging mutation, and so what was once a *de novo* variant appearing for the first time in the germline has the potential to become an inherited variant with the use of Assisted Reproductive Techniques (ART). ART bypass the infertility of a patient in all but the most severe cases, with only a singular sperm needed for ICSI-IVF to occur (Tournaye, 2012). It is reasonable to expect that, if *de novo* mutations do play a role in severe MI and more children who were born via ART reach reproductive age, that an increase in the number of MI cases will be seen. In fact, a 2019 study compared reproductive hormone levels and different semen parameters of young males conceived spontaneously versus those conceived by ICSI (Belva et al., 2019). It was found that, whilst no difference was seen in hormonal levels between the two groups, the ICSI conceived offspring presented with significantly decreased sperm concentration, sperm count and total motile sperm present. Whilst no correlation was identified between the semen concentration or motile sperm count of the young males compared to their fathers, this study still brings to light the worrying possibility of multigenerational infertility being introduced into the general population. As many of these genes being identified as potential *de novo* MI candidate genes are not investigated in the female population, their effect on daughters conceived by ART could potentially be detrimental to female fertility. Another possibility is that dominant MI variants inherited by daughters conceived through ART may remain "silent" with a 50 % chance then of passing on this damaging mutation to any future sons. These scenarios highlight how, if proper diagnosis and subsequent genetic counselling is not implemented, then it may be the case that the issues are simply passed onto future generations.

Understanding the disease mechanisms of causative variants and the genes in which they reside is of high importance for couples considering ART. In the UK in 2019, ART success rates ranged from 32 % in women under the age of 32 to only 5 % in women aged over 43 (Human Fertilisation and Embryology Authority (HFEA), 2021). Whilst ART works for some, for at least two thirds of all couples will be unsuccessful in achieving pregnancy. Being able to predict the likelihood of ART success based on the damaging effects of pathogenic variants identified in patients has the potential to save thousands of women from undergoing what is an incredible invasive procedure and stressful ordeal in cases where MI causing variants are disruptive not only to the process of spermatogenesis but also to fertilization or healthy development of an embryo. It could be expected that causative damaging mutations in genes involved in essential cell division or mitotic pathways may have catastrophic implications during the replication heavy process of foetal development.

6.5.2 Comorbidities Associated with Genetic Causes of Male Infertility

Understanding causative variation in infertile males cannot only provide increased diagnostic rates, but also help more fully interpret the overall health of infertile patients in general. Male infertility has been well established as a disorder associated with a number of comorbidities and an overall reduced life expectancy (Jensen et al., 2009; Eisenberg et al., 2015). The condition is associated with a number of different environmental risk factors such as smoking, obesity, exercise levels, and diet, as well as age, highlighting a pattern of poorer health in infertile males (Dupont et al., 2019; Ferramosca & Zara, 2022; Hassan & Killick, 2003; Sermondade et al., 2013; Sharma et al., 2016; Xie et al., 2022). A number of studies have communicated increased risks for serious conditions such as ischemic heart disease, diabetes, and cancer (Eisenberg et al., 2016; Hanson et al., 2016; Glazer et al., 2017; Laan et al., 2021). A 2022 study looking at WES data from a cohort of 836 NOA patients found that 1 in 28 NOA patients carried a medically actionable secondary finding after genetic investigations with the largest group identified with variants in cancer associated genes (Kasak et al., 2022). The early recognition of increased cancer risk as a secondary finding from WES could be invaluable to the patients allowing for early screening and potentially improved prognosis should cancer be found. The gene *RBM5*, in which seven different potentially pathogenic variants were identified, is a known tumour suppressor and reduction in RBM5 levels has been found in multiple types of lung and breast cancers (Yu et al., 2020; Liang et al., 2012; Jamsai et al., 2017). Whilst the implications of these heterozygous variants on the function of RBM5 are still not fully understood, the potential for comorbidities in these patients shines a light on the significant findings that could be found via WES in patients and the fact that utilising these NGS techniques can provide a wealth of invaluable information on a patient's overall health. In line with this, utilising a patients genotype and phenotype information to categorise them into specific subgroups could provide vital information into specific comorbidities associated with particular MI phenotypes caused by variation in specific gene families or pathways, allowing better clinical follow-ups and potentially treatment of these, and future patients.

6.5.3 Future investigations of candidate genes and potential therapies

The use of NGS for disease gene discovery in male infertility has so far shown promising results in determining the cause of MI in infertile patients, both in the work provided in this thesis as well as by others in the field (Houston et al., 2021). Whilst showing that DNMs are likely to play a role in male infertility is a milestone achievement, the natural progression of this research is in understanding the implications of these findings in relation to patient management and potential treatment. A first step that needs to be taken is the continuation of sequencing larger

patient-parent trio cohorts, utilising both WES and WGS to identify recurrently mutated genes in multiple patients to identify and validate a greater number of candidate MI genes. As this list of confidently linked candidate genes grows, functional studies can take place to confirm the role of these genes within the scope of male infertility. The discovery of further patients with mutations in pre-mRNA splicing genes, such as those found in the *RBM5* patients, would justify further experimentations like those performed in Chapter 5 where the effect of specific mutations can be observed on the splicing factor's known targets. This, however, is just one biological pathway in which dysregulation could lead to infertility. Additional investigations to progress infertility research and advance functional studies would be the collection of tissue samples from all infertility patients to allow transcriptional analyses on each patient. These analyses could be utilised to confirm newly discovered infertility-affected biological pathways in idiopathic patients, in addition to supporting the investigation into mechanisms of infertility in patients with identified causes for their infertility. A 2023 review highlights how the inclusion of multi-omic datasets could help advance understanding of the underlying mechanisms of male infertility and allow for development of novel treatment options (Wagner et al., 2023). The review discusses multiple studies which utilise transcriptomics, proteomics and metabolomics individually to identify genes, pathways and mechanisms which are all associated with impaired male infertility (He et al., 2009; Wang et al., 2009; Walters et al., 2020). The authors conclude that whilst integration of these techniques would require large datasets and greater international collaboration, this is likely necessary to obtain additional insight into the pathology of male infertilty.

As infertility research progresses, consolidation of a list of infertility genes should be possible and work can begin to establish any possible treatment options for these patients. As previously mentioned, for most patients, treatment will lie in genetic counselling and informing patients about their likelihood of successful ART based on their specific genetic variation. One alternative therapy which may become feasible in the future is gene therapy. This class of medical treatment involves the insertion, deletion or replacement of genetic material to alter a patient's genome in order to treat their disease (Nature 2023). The use of gene therapy to repair genetic defaults has been explored in infertile mice in multiple studies. In 2002 a study using lentiviral gene transfer was able to restore spermatogenesis in previously infertile *Sl/Sl*^{d} mutant mice (Ikawa et al., 2002). Despite the increase in spermatogenesis in these mice, ICSI was still required to generate normal offspring. A more recent study was published in 2021 in which gene therapy was used to correct genetic mutations in the gene *Tex11* in mice which lead to meiotic arrest. This study utilised the CRISPR-Cas9 method to correct defects in *Tex11* in

extracted Spermatogonial Stem Cells (SSCs) from infertile mice (Wang et al., 2021). These corrected SSCs were then re-implanted into the infertile mice and restored spermatogenesis. As with the 2002 study however, offspring were only obtained by utilising ART methods, in this case "Round Spermatid Injection" (ROSI). Whilst these studies look promising, as of 2022 there were only four approved gene therapy treatments in the United States. These therapies treat rare diseases such as Haemophilia B, Inherited Retinal Disease and Spinal Muscular Atrophy (SMA), all of which have a severe impact on a patient's quality of life (Heo, 2023; Maguire et al., 2021; Mahajan, 2019; Wong et al., 2023). Currently, these therapies cost millions of dollars per patient and therefore, whilst scientifically possible, the use of gene therapy to treat infertility is likely to remain a low priority when ART methods such as IVF and ICSI can be utilised in many cases (Wong et al., 2023). It could be possible in future that if sub-groups of male infertility patients are identified for whom ART is not viable, that gene therapy is seen as an alternative treatment. Infertile males often suffer with co-morbidities such as an increased risk of cancer due to their genotype. In these instances, gene therapy could be a viable route of investigation. Ideally, all patients would be offered this form of treatment but with the large associated costs and lack of current human research, corrective gene therapy has a long way to go before being introduced into the world of male infertility treatment.

6.6 Concluding Remarks

Male infertility is a complex and genetically heterogenic disorder which affects 1 in 6 couples worldwide. In 50 % of these cases a male factor is the cause. Despite the prevalence of this issue across the globe, very few genes have been identified as MI genes, and up to 70 % of all cases remain unsolved. The work in this thesis provides pivotal evidence for the role for *de novo* mutations (DNMs) in sporadic cases of severe male infertility and highlights the necessity for further research in larger cohorts to identify all MI linked genes. Implementation of NGS into routine clinical diagnostics in idiopathic males would provide a wealth of data with which should increase diagnostic rates and allow better genetic counselling for those undergoing ART. Until this occurs, global collaborations between research groups will continue to be essential to further the field of male infertility gene discovery.

Chapter 7: Appendices

Appendix A: Comparative assessment of the design and results of recent core publications using WES to investigate monogenic causes of male infertility.

Study Title and Reference	Year of Publication	Number of Cases	Patient Phenotypes	Key Findings of the Study
A genomics approach to male infertility (Alhathal et al., 2020)	2020	285 Singletons	NOA or SO	Variants were highlighted in 20 genes with previously established or implicated links to male infertility including *TEX11* and *SYCP2*. 33 novel candidate genes were also identified, all with biological links to male germ cell development, three of which were recurrently mutated within the cohort (*TERB1*, *PIWIL2*, and *ZSWIM7*).
Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility (Wyrwoll et al., 2020)	2020	58 Singletons	NOA, specifically MeiA	A homozygous frameshift variant c.676dup was identified in M1AP, encoding meiosis I associated protein, in three unrelated men. Three additional individuals carrying homozygous c.676dup and three carrying combinations of this and other likely causal variants in M1AP were identified across four additional infertile cohorts. The presence of a homozygous missense variant, c.1166C>T (p.Pro389Leu), was also identified in five men from a consanguineous Turkish family.
Genetic dissection of spermatogenic arrest through exome analysis: clinical implications for the management of azoospermic men (Krausz et al., 2020)	2020	147 Singletons	Maturation Arrest	Five novel genes were identified with strong evidence linking them to the MA phenotype (*ADAD2*, *TERB1*, *SHOC1*, *MSH4*, and *RAD21L1*). Four of these were validated in two independent MA cohorts. Nine additional patients carried pathogenic variants in seven previously reported genes (*TEX14*, *DMRT1*, *TEX11*, *SYCE1*, *MEIOB*, *MEI1*, and *STAG3*), upgrading the clinical significance of these genes.
Whole-exome sequencing of a large Chinese azoospermia and severe oligospermia cohort identifies novel infertility causative variants and genes (Chen et al., 2020)	2020	314 Singletons	NOA or SO	Six pathogenic/likely pathogenic variants and four variants of unknown significance were identified in genes known to cause NOA/SO (*DMC1*, *MEI1*, *TEX11*, *AR*, *HAUS7*, *TEX11*, *WNK3*). A further 20 novel NOA candidate genes affecting 25 patients were identified. Five of these genes

				(*BRDT*, *CHD5*, *MCM9*, *MLH3*, and *ZFX*) emerged as strong candidates, supported by evidence from both murine functional studies and human single-cell (sc)RNA-sequencing data.
De novo mutations in idiopathic male infertility—A pilot study (Hodžić et al., 2021)	2020	13 Trios	NOA	*De novo* mutations were identified in ten genes, of which five showed potential association with the azoospermia phenotype seen in patients (*SEMA5A*, *NEURL4*, *BRD2*, *CD1D*, and *CD63*). All candidate genes exhibited significant differential expression when comparing testis samples from individuals with severely impaired spermatogenesis to those with normal spermatogenesis. Rare potentially pathogenic mutations were also identified in previously implicated MI candidate genes *FKBPL*, *UPF2*, *CLCA4* and *NR0B1*.
Bi-allelic variants in DNA mismatch repair proteins MutS Homolog MSH4 and MSH5 cause infertility in both sexes (Wyrwoll et al., 2021)	2021	1305 Singletons	NOA or Crypto 90 with MeiA	Likely pathogenic homozygous variants in the gene *MSH5* were identified in six individual infertile males (two with MeiA and four with azoospermia). This was the first implication of mutations in this gene in male infertility. Biallelic variants were also highlighted in the gene *MSH4* in two males with MeiA phenotypes. γH2AX staining revealed an arrest in early prophase of meiosis I in individuals with pathogenic *MSH4* or *MSH5* variants. In vitro expression of the detected LoF *MSH5* variants showed MSH5 protein truncation in one case and a total loss of MSH5 in the other.
Exome sequencing reveals variants in known and novel candidate genes for severe sperm motility disorders (Oud et al., 2021)	2021	21 Singletons	Severe motility disorders	In 10 of 21 patients with severe sperm motility issues, pathogenic variants were found in known sperm assembly genes: *CFAP43* (n = 3), *CFAP44* (n = 2), *CFAP58* (n = 1), *QRICH2* (n = 2), *DNAH1* (n = 1) and *DNAH6* (n = 1) representing 48 % of all patients. Furthermore, six novel human candidate sperm motility genes were identified *DNAH12*, *DRC1*, *MDC1*, *PACRG*, *SSPL2C* and *TPTE2*.
Diverse monogenic subforms of human spermatogenic failure (Nagirnaja et al., 2022)	2022	924 Singletons	NOA	WES of 924 singletons lead to the identification of 221 potentially NOA associated genes. A further investigation in an additional replication cohort lead to the compilation of a higher confidence list of 21 genes associated with NOA. All

				these 21 genes were observed in multiple NOA cases and had nominally significant association with NOA in a comparison with 11,587 fertile controls. These genes however did not reach exome wide significance. In the initial 221 identified genes an enrichment was seen for previously unidentified LoF mutations on the X and Y chromosomes ($p = 0.018$).
Large-scale analyses of the X chromosome in 2,354 infertile men discover recurrently affected genes associated with spermatogenic failure (Riera-Escamilla et al., 2022)	2022	2354 singletons	NOA or Crypto	Focussing on the X chromosome, 55 recurrently mutated genes were identified as novel candidate genes for azoospermia/cryptozoospermia. Of these, 21 were strongly associated with the patient phenotype and 34 showed moderate association. The gene *RBBP7* was seen to be most frequently affected with nine NOA in the cohort presenting with mutations. Functional studies supported the role of this gene in germ stem cell maintenance.
DDX3Y is likely the key spermatogenic factor in the AZFa region that contributes to human non-obstructive azoospermia (Dicke et al., 2023)	2023	1655 singletons	NOA or Crypto	In the initial cohort of 1655 singletons, three males carried different LoF variants in *DDX3Y*. In a WES replication cohort (GEMINI), an additional LoF variant was identified. Significantly, none of the four variants were present in the gnomAD database. TESE was performed for three of the variant carriers and providing a testicular phenotype of SCO for all three cases, a phenotype associated typically with carriers of complete AZFa deletions. The study provided evidence and suggested that the gene *DDX3Y* is the key gene leading to spermatogenic failure observed in men with complete AZFa deletions.

N.B. Crypto: Cryptozoospermia , MeiA: Meiotic Arrest, NOA: Non-Obstructive Azoospermia, SO: Severe Oligozoospermia

Appendix B: Table of all DNMs identified in 185 patient-parent trios in NIJ / NCL cohorts of infertile men.

Patient ID	Origin	Semen Analysis	Gene	Chromosome coordinates (GRCh37)	Conclusion Variant / Gene	Conclusion Patient
Proband_005	Netherlands	Azoospermia	*CDK5RAP2*	chr9:123215805-G-A	Possibly causative	Candidate *de novo* point mutation (*CDK5RAP2*)
Proband_006	Netherlands	Azoospermia	*ATP1A1*	chr1:116930014-CT-C	Unclear	Multiple candidate genes
			TLN2	chr15:63029134-G-A	Unclear	
			HUWE1	chrX:53589090-TTCCTCC-T	Possibly causative	
Proband_008	Netherlands	Azoospermia	*ABCC10*	chr6:43417749-C-T	Not causative	No candidates
			CP	chr3:148927135-C-T	Unlikely causative	
Proband_010	Netherlands	Azoospermia	*FUS*	chr16:31196402-TGGCGGCGGC-T	Not causative	No candidates
			LTBP1	chr2:33246090-C-G	Unlikely causative	
Proband_012	Netherlands	Extreme oligozoospermia	*RP1L1*	chr8:10480174-C-T	Not causative	No candidates
Proband_013	Netherlands	Azoospermia	*ERG*	chr21:39755563-G-A	Unclear	Candidate *de novo* point mutation (*ERG*)
Proband_017	Netherlands	Azoospermia	*CDC5L*	chr6:44413480-G-A	Possibly causative	Candidate *de novo* point mutation (*CDC5L*)
Proband_019	Netherlands	Azoospermia	*ABLIM1*	chr10:116205100-G-A	Possibly causative	Candidate *de novo* point mutation (*ABLIM1*)
Proband_020	Netherlands	Azoospermia	*CCDC126*	chr7:23682709-C-T	Not causative	Candidate *de novo* point mutation (*RASEF*)
			RASEF	chr9:85607885-C-T	Unclear	
Proband_022	Netherlands	Azoospermia	*APC2*	chr19:1453118-G-A	Unclear	Candidate *de novo* point mutation (*APC2*)

Proband_025	Netherlands	Azoospermia	*NEO1*	chr15:73575428-G-A	Not causative	No candidates
Proband_028	Netherlands	Extreme oligozoospermia	*OR5P3*	chr11:7846930-A-T	Unclear	Candidate *de novo* point mutation *(OR5P3*)
Proband_030	Netherlands	Extreme oligozoospermia	*SIKE1*	chr1:115323119-T-G	Unclear	Multiple candidate genes
			TRAF7	chr16:2218149-C-T	Possibly Causative	
Proband_033	Netherlands	Azoospermia	*KRT33B*	chr17:39521752-G-C	Unlikely causative	Candidate *de novo* point mutation (*ATP8A1*)
			SENP7	chr3:101212750-G-C	Unlikely causative	
			ATP8A1	chr4:42626573-T-C	Unclear	
Proband_038	Netherlands	Azoospermia	*NOC3L*	chr10:96100055-T-G	Unlikely causative	No candidates
Proband_039	Netherlands	Azoospermia	*NXT2*	chrX:108779109:108785919	Unlikely causative	No candidates
Proband_041	Netherlands	Azoospermia	*ASIC5*	chr4:156763436-C-T	Unclear	Candidate *de novo* point mutation (*ASIC5*)
Proband_042	Netherlands	Azoospermia	*PLCL1*	chr2:198966024-C-T	Unclear	Multiple candidate genes
			DNAJC2	chr7:102957321-T-A	Possibly causative	
			AK3	chr9:4722547-T-C	Unlikely causative	
Proband_043	Netherlands	Azoospermia	*IL33*	chr9:6251221-G-A	Unlikely causative	No candidates
Proband_044	Netherlands	Azoospermia	*PRDM16*	chr1:3328833-A-T	Unclear	Multiple candidate genes
			PPP1R7	chr2:242099831-CAATAA-C	Possibly Causative	
Proband_045	Netherlands	Azoospermia	*EVC*	chr4:5743515-C-T	Unclear	

			BHMT	chr5:78426892-C-G	Unlikely causative	Candidate *de novo* LoF mutation (*EVC*)
Proband_048	Netherlands	Extreme oligozoospermia	*MCM6*	chr2:136624195-T-C	Possibly causative	Candidate *de novo* point mutation (*MCM6*)
Proband_049	Netherlands	Azoospermia	*ATP8B4*	chr15:50168651-T-C	Unlikely causative	No candidates
			ZNF577	chr19:52376320-T-C	Unlikely causative	
			NUP210	chr3:13359251-T-C	Unlikely causative	
Proband_050	Netherlands	Extreme oligozoospermia	*HIST1H1D*	chr6:26234816-C-G	Unlikely causative	No candidates
Proband_051	Netherlands	Extreme oligozoospermia	*FNDC8*	chr17:33448840-G-A	Unlikely causative	No candidates
Proband_052	Netherlands	Severe oligozoospermia	*SOGA1*	chr20:35438426-G-A	Unclear	Candidate *de novo* point mutation (*SOGA1*)
Proband_053	Netherlands	Extreme oligozoospermia	*CD81*	chr11:2417877-T-C	Possibly causative	Candidate *de novo* point mutation (*CD81*)
Proband_055	Netherlands	Azoospermia	*OSBPL3*	chr7:24874131-G-A	Unclear	Candidate *de novo* point mutation (*OSBPL3*)
Proband_057	Netherlands	Extreme oligozoospermia	*ABCB9*	chr12:123430670-C-T	Unlikely causative	No candidates
Proband_060	Netherlands	Azoospermia	*IL12RB2*	chr1:67861761-C-G	Unlikely causative	Candidate *de novo* point mutation (*TOPAZ1*)
			TOPAZ1	chr3:44286601-A-G	Possibly causative	

Proband_061	Netherlands	Azoospermia	*SEC14L1*	chr17:75208114-A-G	Not causative	No candidates
Proband_062	Netherlands	Extreme oligozoospermia	*FOXF2*	chr6:1391182-A-G	Unclear	Candidate *de novo* point mutation (*FOXF2*)
Proband_063	Netherlands	Extreme oligozoospermia	*CNOT4*	chr7:135048789-T-C	Unlikely causative	No candidates
Proband_064	Netherlands	Severe oligozoospermia	*C9orf50*	chr9:132374702-G-A	Possibly causative	Candidate *de novo* point mutation (*C9orf50*)
Proband_066	Sint Maarten (Caribbean)	Extreme oligozoospermia	*HIPK3*	chr11:32975325:33631588	Unclear	Multiple novel candidate genes
			QSER1			
			DEPDC7			
			TCP11L1			
			CSTF3			
			KIAA1549L			
			TDRD10	chr1:154493890-G-A	Unlikely causative	
			CWC27	chr5:64077814-C-G	Possibly causative	
Proband_073	Netherlands	Extreme oligozoospermia	*INO80*	chr15:41372056-G-A	Possibly Causative	Candidate *de novo* point mutation (*INO80*)
Proband_074	Netherlands	Severe oligoasthenozoospermia	*ZNF709*	chr19:12575362-A-AA	Unclear	Multiple novel candidate genes
			EMILIN1	chr2:27305208-G-A	Unclear	
			WDR17	chr4:177067235-G-A	Unclear	
			ZNF311	chr6:28963503-G-A	Unlikely causative	
Proband_076	Netherlands	Azoospermia	*STARD10*	chr11:72466763-GAGA-G	Possibly Causative	Multiple novel candidate genes

			GREB1L	chr18:19019514-AAGGGC-A	Possibly Causative	
Proband_077	Netherlands	Extreme oligozoospermia	*MSH5*	chr6:31721100-CAT-C	Unlikely causative	No candidates
Proband_079	Netherlands	Azoospermia	*ILVBL*	chr19:15226717-C-T	Unclear	Candidate *de novo* point mutation (*ILVBL*)
Proband_080	Netherlands	Extreme oligozoospermia	*HOXA1*	chr7:27134363-T-G	Unclear	Candidate *de novo* point mutation (*HOXA1*)
Proband_081	Netherlands	Azoospermia	*ZFHX4*	chr8:77763486-CT-C	Unclear	Candidate *de novo* LoF mutation (*ZFHX4*)
Proband_083	Netherlands	Azoospermia	*F13B*	chr1:197026291-G-T	Unlikely causative	Candidate *de novo* point mutation (*HNRNPL*)
			HNRNPL	chr19:39329152-C-T	Possibly causative	
Proband_085	Netherlands	Severe oligoasthenozoospermia	*NLRP10*	chr11:7981967-C-T	Unlikely causative	No candidates
Proband_087	Netherlands	Severe oligoasthenozoospermia	*LEO1*	chr15:52252183-A-C	Possibly causative	Candidate *de novo* point mutation (*LEO1*)
Proband_088	Netherlands	Extreme oligozoospermia	*GDAP1L1*	chr20:42893167-C-T	Not causative	No candidates
Proband_089	Netherlands	Azoospermia	*COLGALT2*	chr1:184006280-A-C	Unlikely causative	Candidate *de novo* point mutation (*SSH2*)
			SSH2	chr17:27963222-G-A	Unclear	
Proband_090	Unknown	Extreme oligozoospermia	*SETX*	chr9:135205489-C-T	Unlikely causative	No candidates

Proband_095	Netherlands	Azoospermia	*MPRIP*	chr17:17062191-C-T	Unclear	Candidate *de novo* point mutation (*MPRIP*)
			SORCS2	chr4:7728558-G-A	Unlikely causative	
Proband_097	Netherlands	Azoospermia	*TENM2*	chr5:167642269-GC-G	Unclear	Candidate *de novo* LoF mutation (*TENM2*)
Proband_101	Netherlands	Extreme oligozoospermia	*CHST12*	chr7:2472611-C-T	Unlikely causative	No candidates
Proband_102	Netherlands	Extreme oligozoospermia	*HR*	chr8:21973239-C-T	Unlikely causative	Candidate *de novo* point mutation (*SMC2*)
			SMC2	chr9:106885442-T-C	Possibly causative	
Proband_106	Netherlands	Azoospermia	*TACC2*	chr10:123976284-C-T	Unlikely causative	No candidates
Proband_108	Netherlands	Severe oligoasthenozoospermia	*CYP4F12*	chr19:15794526-G-A	Unlikely causative	Candidate *de novo* point mutation (*RBM5*)
			RBM5	chr3:50140556-A-G	Possibly Causative	
			TAF9	chr5:68660785-GTCA-G	Unlikely causative	
Proband_115	Netherlands	Severe oligoasthenozoospermia	*CDC42BPG*	chr11:64603286-G-A	Not causative	Candidate *de novo* point mutation (*RPA1*)
			RPA1	chr17:1756424-T-C	Possibly causative	
Proband_116	Netherlands	Severe oligoasthenozoospermia	*FBXO5*	chr6:153293449-A-ATCAC	Possibly causative	Candidate *de novo* LoF mutation (*FBXO5*)
Proband_117	Netherlands	Azoospermia	*FLNC*	chr7:128477754-C-T	Unclear	Candidate *de novo* point mutation (*FLNC*)
Proband_118	Netherlands	Azoospermia	*CDCA8*	chr1:38166170-C-T	Not causative	No candidates

Proband_119	Netherlands	Severe oligoasthenozoospermia	*ZCCHC2*	chr18:60242391-A-G	Unlikely causative	No candidates
Proband_121	Netherlands	Azoospermia	*AMPD2*	chr1:110168336-A-T	Unclear	Candidate *de novo* point mutation (*AMPD2*)
			HIVEP1	chr6:12163643C-T	Not causative	
			SPEF2	chr5:35792492-C-G	Unlikely causative	
Proband_122	Netherlands	Azoospermia	*SIGLEC10*	chr19:51919175-C-T	Not causative	No candidates
Proband_124	Netherlands	Severe oligoasthenozoospermia	*PPP1R3A*	chr7:113518521-A-C	Unlikely causative	No candidates
Proband_125	Netherlands	Azoospermia	*CHST4*	chr16:71571634-G-A	Not causative	Candidate *de novo* point mutation (*STXBP2*)
			STXBP2	chr19:7711198-C-T	Unclear	
Proband_126	Netherlands	Severe oligoasthenozoospermia	*ABCF3*	chr3:183907504-C-T	Not causative	No candidates
Proband_127	Netherlands	Extreme oligozoospermia	*TMEM62*	chr15:43441280-C-G	Unclear	Candidate *de novo* point mutation (*U2AF2*)
			U2AF2	chr19:56170622-T-TCGGAGC	Possibly causative	
Proband_128	Netherlands	Azoospermia	*HELZ2*	chr20:62195532-G-A	Unclear	Candidate *de novo* point mutation (*HELZ2*)
Proband_129	Netherlands	Azoospermia	*FIZ1*	chr19:56104069-ATCT-A	Unclear	Candidate *de novo* point mutation (*FIZ1*)
Proband_130	Netherlands	Azoospermia	*MAVS*	chr20:3846631-C-T	Unlikely causative	Candidate *de novo* point mutation (*TMPPE*)
			TMPPE	chr3:33134784-T-C	Unclear	
Proband_132	Netherlands	Azoospermia	*USH2A*	chr1:216595434-C-A	Unlikely causative	No candidates
Proband_133	Netherlands	Azoospermia	*EMP1*	chr12:13366446-G-C	Unlikely causative	No candidates

Proband_134	Netherlands	Extreme oligozoospermia	*ERI3*	chr1:44687249-T-C	Unlikely causative	No candidates
Proband_135	Netherlands	Severe oligozoospermia	*POPDC3*	chr6:105609709-C-T	Unclear	Candidate *de novo* point mutation (*POPDC3*)
Proband_136	Netherlands	Severe oligoasthenozoospermia	*PLEKHA1*	chr10:124189195-C-G	Unlikely causative	No candidates
			BTAF1	chr10:93753563-T-C	Unlikely causative	
Proband_137	Netherlands	Azoospermia	*C12orf49*	chr12:117155674-G-A	Unclear	Candidate *de novo* point mutation (*C12orf49*)
Proband_138	Netherlands	Severe oligozoospermia	*GRIP1*	chr12:66849967-G-A	Unlikely causative	No candidates
Proband_139	Netherlands	Extreme oligozoospermia	*RNF223*	chr1:1007489-C-T	Unlikely causative	Multiple novel candidate genes
			ZNF469	chr16:88494628-T-TC	Unclear	
			MAP3K3	chr17:61759150-C-T	Possibly Causative	
			C17orf74	chr17:7330308-G-A	Unlikely causative	
			TMPRSS11B	chr4:69107421-T-C	Unclear	
Proband_142	Netherlands	Azoospermia	*GPR75-ASB3*	chr2:53921057-G-A	Unclear	Candidate *de novo* point mutation (*GPR75-ASB3*)
Proband_144	Unknown	Azoospermia	*ODF1*	chr8:103563960-C-G	Possibly causative	Candidate *de novo* point mutation (*ODF1*)
Proband_145	Netherlands	Extreme oligozoospermia	*EXOSC10*	chr1:11136965-A-AT	Unclear	Candidate *de novo* point mutation (*EXOSC10*)

Proband_146	Unknown	Severe oligoasthenozoospermia	*OTOA*	chr16:21728238-C-T	Not causative	No candidates
Proband_148	Netherlands	Azoospermia	*CRHR1*	chr17:43907477-G-A	Unclear	Multiple novel candidate genes
			HTT	chr4:3213834-C-CG	Possibly Causative	
Proband_149	Netherlands	Azoospermia	*SNED1*	chr2:242012773-A-G	Unlikely causative	Candidate *de novo* LoF mutation (*PCDHB1*)
			PCDHB1	chr5:140431405-T-TCCGGG	Unclear	
Proband_150	Netherlands	Azoospermia	*SPECC1L*	chr22:24761453-G-A	Possibly Causative	Candidate *de novo* point mutation (*SPECC1L*)
Proband_153	Netherlands	Azoospermia	*IQSEC1*	chr3:13008948-A-T	Unclear	Candidate *de novo* point mutation (*IQSEC1*)
Proband_154	Netherlands	Azoospermia	*ARHGAP33*	chr19:36276183-G-A	Unclear	Candidate *de novo* point mutation (*ARHGAP33*)
			C10orf107	chr10:63445916-A-G	Not causative	
Proband_156	Netherlands	Azoospermia	*LRRN2*	chr1:204587973-G-A	Unclear	Candidate *de novo* point mutation (*LRRN2*)
			SRCIN1	chr17:36714613-G-A	Unlikely causative	
Proband_157	Netherlands	Azoospermia	*REN*	chr1:204124193-G-A	Unclear	Multiple novel candidate genes
			SIPA1L3	chr19:38643580-G-A	Unclear	
Proband_158	Netherlands	Azoospermia	*TP53TG5*	chr20:44003729-T-A	Possibly Causative	Candidate *de novo* point mutation (*TP53TG5*)
			DHX36	chr3:153994607-C-T	Unlikely causative	
			YEATS2	chr3:183524758-C-A	Unlikely causative	
Proband_160	Netherlands	Azoospermia	*SDF4*	chr1:1158720-C-T	Possibly Causative	Multiple novel candidate genes
			ITSN2	chr2:24522905-C-T	Unclear	

Proband_165	Netherlands	Extreme oligozoospermia	*MYOF*	chr10:95168662-C-T	Not causative	Candidate *de novo* LoF mutation (*RASAL2*)
			RASAL2	chr1:178435121-G-T	Possibly Causative	
Proband_166	Netherlands	Azoospermia	*C6orf25*	chr6:31691437-G-A	Unclear	Candidate *de novo* point mutation (*C6orf25*)
Proband_168	Netherlands	Extreme oligozoospermia	*KLC1*	chr14:104129206-TC-T	Unclear	Candidate *de novo* LoF mutation (*KLC1*)
Proband_170	Netherlands	Azoospermia	*PRPF4B*	chr6:4049340-C-T	Unclear	Candidate *de novo* point mutation (*PRPF4B*)
Proband_173	United Kingdom	Azoospermia	*ZNF629*	chr16:30793127-G-C	Unclear	Candidate *de novo* point mutation (*ZNF629*)
			CXXC11	chr2:242812032-G-A	Unlikely causative	
			IRAK2	chr3:10255045-A-T	Unlikely causative	
Proband_178	United Kingdom	Severe Oligozoospermia	*MICU1*	chr10:74322772-C-T	Unlikely causative	No candidates
			GHRHR	chr7:31016058-C-T	Unlikely causative	
Proband_179	United Kingdom	Azoospermia	*CELSR2*	chr1:109807146-T-C	Unclear	Candidate *de novo* point mutation (*CELSR2*)
			BASP1	chr5:17275915-C-T	Unlikely causative	

Chapter 8: Bibliography

Achermann, A.P.P., Pereira, T.A. & Esteves, S.C. (2021) 'Microdissection testicular sperm extraction (micro-TESE) in men with infertility due to nonobstructive azoospermia: summary of current literature', *International Urology and Nephrology*, 53(11), pp. 2193–2210.

Acuna-Hidalgo, R., Veltman, J.A. & Hoischen, A. (2016) 'New insights into the generation and role of de novo mutations in health and disease', *Genome Biology*, 17(1), p. 241.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. & Sunyaev, S.R. (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods*, 7(4), pp. 248–249.

Agarwal, A., Mulgund, A., Hamada, A. & Chyatte, M.R. (2015) 'A unique view on male infertility around the globe', *Reproductive Biology and Endocrinology*, 13(1), p. 37.

Aggarwal, B., Evans, A.L., Ryan, H. & Martins da Silva, S.J. (2021) 'IVF or ICSI for fertility preservation?', *Reproduction and Fertility*, 2(1), pp. L1–L3.

Ajuh, P. (2000) 'Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry', *The EMBO Journal*, 19(23), pp. 6569–6581.

Alhathal, N., Maddirevula, S., Coskun, S., Alali, H., Assoum, M., Morris, T., Deek, H.A., Hamed, S.A., Alsuhaibani, S., Mirdawi, A., Ewida, N., Al-Qahtani, M., Ibrahim, N., Abdulwahab, F., Altaweel, W., Dasouki, M.J., Assiri, A., Qabbaj, W. & Alkuraya, F.S. (2020) 'A genomics approach to male infertility', *Genetics in Medicine*, 22(12), pp. 1967–1975.

Alksere, B., Berzina, D., Dudorova, A., Conka, U., Andersone, S., Pimane, E., Krasucka, S., Blumberga, A., Dzalbs, A., Grinfelde, I., Vedmedovska, N., Fodina, V. & Erenpreiss, J. (2019) 'Case of Inherited Partial AZFa Deletion without Impact on Male Fertility', *Case Reports in Genetics*, 2019pp. 1–5.

de Almeida, R.A. & O'Keefe, R.T. (2015) 'The NineTeen Complex (NTC) and NTC-associated proteins as targets for spliceosomal ATPase action during pre-mRNA splicing', *RNA Biology*, 12(2), pp. 109–114.

Almesned, R.K., Alsuhaibani, S.S., Alali, H.J., Qubbaj, W.A. & Al Hathal, N.K. (2020) 'Male Infertility in Robertsonian Translocation: A Case Report', *American Journal of Case Reports*, 21.

Amiri-Yekta, A., Coutton, C., Kherraf, Z.-E., Karaouzène, T., Le Tanno, P., Sanati, M.H., Sabbaghian, M., Almadani, N., Sadighi Gilani, M.A., Hosseini, S.H., Bahrami, S., Daneshipour, A., Bini, M., Arnoult, C., Colombo, R., Gourabi, H. & Ray, P.F. (2016) 'Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new *DNAH1* mutations', *Human Reproduction*, 31(12), pp. 2872–2880.

Anna, A. & Monika, G. (2018) 'Splicing mutations in human genetic disorders: examples, detection, and confirmation', *Journal of Applied Genetics*, 59(3), pp. 253–268.

Arafat, M., Har-Vardi, I., Harlev, A., Levitas, E., Zeadna, A., Abofoul-Azab, M., Dyomin, V., Sheffield, V.C., Lunenfeld, E., Huleihel, M. & Parvari, R. (2017) 'Mutation in TDRD9 causes non-obstructive azoospermia in infertile men', *Journal of Medical Genetics*, 54(9), pp. 633–639.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. & DePristo, M.A. (2013) 'From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline', *Current Protocols in Bioinformatics*, 43(1), .

Baillet, A., Le Bouffant, R., Volff, J.N., Luangpraseuth, A., Poumerol, E., Thépot, D., Pailhoux, E., Livera, G., Cotinot, C. & Mandon-Pépin, B. (2011) 'TOPAZ1, a Novel Germ Cell-Specific Expressed Gene Conserved during Evolution across Vertebrates', *PLoS ONE*, 6(11), p. e26950.

Balasubramanian, R. & Crowley, W.F. (2017) 'Reproductive endocrine phenotypes relating to *CHD7* mutations in humans', *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 175(4), pp. 507–515.

Bamshad, M.J., Nickerson, D.A. & Chong, J.X. (2019) 'Mendelian Gene Discovery: Fast and Furious with No End in Sight', *The American Journal of Human Genetics*, 105(3), pp. 448–455.

Bardoni, B., Zuffardi, O., Guioli, S., Ballabio, A., Simi, P., Cavalli, P., Grimoldi, M.G., Fraccaro, M. & Camerino, G. (1991) 'A deletion map of the human Yq11 region: Implications for the evolution of the Y chromosome and tentative mapping of a locus involved in spermatogenesis', *Genomics*, 11(2), pp. 443–451.

Bashamboo, A., Ferraz-de-Souza, B., Lourenço, D., Lin, L., Sebire, N.J., Montjean, D., Bignon-Topalovic, J., Mandelbaum, J., Siffroi, J.-P., Christin-Maitre, S., Radhakrishna, U., Rouba, H., Ravel, C., Seeler, J., Achermann, J.C. & McElreavey, K. (2010) 'Human Male Infertility Associated with Mutations in NR5A1 Encoding Steroidogenic Factor 1', *The American Journal of Human Genetics*, 87(4), pp. 505–512.

Bekers, E.M., Eijkelenboom, A., Rombout, P., van Zwam, P., Mol, S., Ruijter, E., Scheijen, B. & Flucke, U. (2019) 'Identification of novel GNAS mutations in intramuscular myxoma using next-generation sequencing with single-molecule tagged molecular inversion probes', *Diagnostic Pathology*, 14(1), p. 15.

Belva, F., Bonduelle, M., Roelants, M., Michielsen, D., Van Steirteghem, A., Verheyen, G. & Tournaye, H. (2016) 'Semen quality of young adult ICSI offspring: the first results', *Human Reproduction*, 31(12), pp. 2811–2820.

Belva, F., Bonduelle, M. & Tournaye, H. (2019) 'Endocrine and reproductive profile of boys and young adults conceived after ICSI', *Current Opinion in Obstetrics & Gynecology*, 31(3), pp. 163–169.

Ben Khelifa, M., Coutton, C., Zouari, R., Karaouzène, T., Rendu, J., Bidart, M., Yassine, S., Pierre, V., Delaroche, J., Hennebicq, S., Grunwald, D., Escalier, D., Pernet-Gallay, K., Jouk, P.-S., Thierry-Mieg, N., Touré, A., Arnoult, C. & Ray, P.F. (2014) 'Mutations in DNAH1, which Encodes an Inner Arm Heavy Chain Dynein, Lead to Male Infertility from Multiple Morphological Abnormalities of the Sperm Flagella', *The American Journal of Human Genetics*, 94(1), pp. 95–104.

Berglund, J.A., Abovich, N. & Rosbash, M. (1998) 'A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition', *Genes & Development*, 12(6), pp. 858–867.

Bertrand, R.E., Wang, J., Li, Y., Cheng, X., Wang, K., Stoilov, P. & Chen, R. (2022) 'Cwc27, associated with retinal degeneration, functions as a splicing factor *in vivo*', *Human Molecular Genetics*, 31(8), pp. 1278–1292.

Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H. & Lührmann, R. (2008) 'Isolation of an active step I spliceosome and composition of its RNP core', *Nature*, 452(7189), pp. 846–850.

Bettio, D., Venci, A. & Levi Setti, P.E. (2008) 'Chromosomal Abnormalities in Miscarriages after Different Assisted Reproduction Procedures', *Placenta*, 29pp. 126–128.

Bieth, E., Hamdi, S.M. & Mieusset, R. (2021) 'Genetics of the congenital absence of the vas deferens', *Human Genetics*, 140(1), pp. 59–76.

Bliss, S.P., Navratil, A.M., Xie, J. & Roberson, M.S. (2010) 'GnRH signaling, the gonadotrope and endocrine control of fertility', *Frontiers in Neuroendocrinology*, 31(3), pp. 322–340.

De Boeck, K. (2020) 'Cystic fibrosis in the year 2020: A disease with a new face', *Acta Paediatrica*, 109(5), pp. 893–899.

Bonduelle, M., Wilikens, A., Buysse, A., Van Assche, E., Devroey, P., Van Steirteghem, A.C. & Liebaers, I. (1998) 'A follow-up study of children born after intracytoplasmic sperm injection (ICSI) with epididymal and testicular spermatozoa and after replacement of cryopreserved embryos obtained after ICSI', *Human Reproduction*, 13(suppl 1), pp. 196–207.

Bonnal, S., Martínez, C., Förch, P., Bachi, A., Wilm, M. & Valcárcel, J. (2008) 'RBM5/Luca-15/H37 Regulates Fas Alternative Splice Site Pairing after Exon Definition', *Molecular Cell*, 32(1), pp. 81–95.

Bose, R., Sheng, K., Moawad, A.R., Manku, G., O'Flaherty, C., Taketo, T., Culty, M., Fok, K.L. & Wing, S.S. (2017) 'Ubiquitin Ligase Huwe1 Modulates Spermatogenesis by Regulating Spermatogonial Differentiation and Entry into Meiosis', *Scientific Reports*, 7(1), p. 17759.

Boursier, A., Dumont, A., Boitrelle, F., Prasivoravong, J., Lefebvre-Khalil, V., Robin, G. & Barbotin, A. (2022) 'Necrozoospermia: The tree that hides the forest', *Andrology*, 10(4), pp. 642–659.

Boycott, K.M., Vanstone, M.R., Bulman, D.E. & MacKenzie, A.E. (2013) 'Rare-disease genetics in the era of next-generation sequencing: discovery to translation', *Nature Reviews Genetics*, 14(10), pp. 681–691.

Boyle, E.A., O'Roak, B.J., Martin, B.K., Kumar, A. & Shendure, J. (2014) 'MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing', *Bioinformatics*, 30(18), pp. 2670–2672.

Brahem, S., Jellad, S., Ibala, S., Saad, A. & Mehdi, M. (2012) 'DNA fragmentation status in patients with necrozoospermia', *Systems Biology in Reproductive Medicine*, 58(6), pp. 319–323.

Brea-Fernández, A.J., Cabanas, P., Dacruz-Álvarez, D., Caamaño, P., Limeres, J. & Loidi, L. (2019) 'Expanding the clinical and molecular spectrum of the CWC27-related spliceosomopathy', *Journal of Human Genetics*, 64(11), pp. 1133–1136.

Brophy, P.D., Rasmussen, M., Parida, M., Bonde, G., Darbro, B.W., Hong, X., Clarke, J.C., Peterson, K.A., Denegre, J., Schneider, M., Sussman, C.R., Sunde, L., Lildballe, D.L., Hertz, J.M., Cornell, R.A., Murray, S.A. & Manak, J.R. (2017) 'A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans', *Genetics*, 207(1), pp. 215–228.

Brown, T.R., Lubahn, D.B., Wilson, E.M., Joseph, D.R., French, F.S. & Migeon, C.J. (1988) 'Deletion of the steroid-binding domain of the human androgen receptor gene in one family with complete androgen insensitivity syndrome: evidence for further genetic heterogeneity in this syndrome.', *Proceedings of the National Academy of Sciences*, 85(21), pp. 8151–8155.

Brue, T. & Delemer, B. (2007) 'Diagnosis and management of hyperprolactinemia: expert consensus – French Society of Endocrinology', *Annales d'Endocrinologie*, 68(1), pp. 58–64.

Busetto, V., Barbosa, I., Basquin, J., Marquenet, É., Hocq, R., Hennion, M., Paternina, J.A., Namane, A., Conti, E., Bensaude, O. & Le Hir, H. (2020) 'Structural and functional insights into CWC27/CWC22 heterodimer linking the exon junction complex to spliceosomes', *Nucleic Acids Research*, 48(10), pp. 5670–5683.

Butz, H., Nyírő, G., Kurucz, P.A., Likó, I. & Patócs, A. (2021) 'Molecular genetic diagnostics of hypogonadotropic hypogonadism: from panel design towards result interpretation in clinical practice', *Human Genetics*, 140(1), pp. 113–134.

Camats, N., Pandey, A. V., Fernández-Cancio, M., Andaluz, P., Janner, M., Torán, N., Moreno, F., Bereket, A., Akcay, T., García-García, E., Muñoz, M.T., Gracia, R., Nistal, M., Castaño, L., Mullis, P.E., Carrascosa, A., Audí, L. & Flück, C.E. (2012) 'Ten Novel Mutations in the NR5A1 Gene Cause Disordered Sex Development in 46,XY and Ovarian Insufficiency in 46,XX Individuals', *The Journal of Clinical Endocrinology & Metabolism*, 97(7), pp. E1294–E1306.

Cangiano, B., Swee, D.S., Quinton, R. & Bonomi, M. (2021) 'Genetics of congenital hypogonadotropic hypogonadism: peculiarities and phenotype of an oligogenic disease', *Human Genetics*, 140(1), pp. 77–111.

Cannarella, R., Condorelli, R.A., Duca, Y., La Vignera, S. & Calogero, A.E. (2019) 'New insights into the genetics of spermatogenic failure: a review of the literature', *Human Genetics*, 138(2), pp. 125–140.

Cao, H., Wu, J., Lam, S., Duan, R., Newnham, C., Molday, R.S., Graziotto, J.J., Pierce, E.A. & Hu, J. (2011) 'Temporal and Tissue Specific Regulation of RP-Associated Splicing Factor Genes PRPF3, PRPF31 and PRPC8—Implications in the Pathogenesis of RP', *PLoS ONE*, 6(1), p. e15860.

Capron, C., Januel, L., Vieville, G., Jaillard, S., Kuentz, P., Salaun, G., Nadeau, G., Clement, P., Brechard, M.P., Herve, B., Dupont, J.M., Gruchy, N., Chambon, P., Abdelhedi, F., Dahlen, E., Vago, P., Harbuz, R., Plotton, I., Coutton, C., et al. (2022) 'Evidence for high

breakpoint variability in 46, XX, SRY-positive testicular disorder and frequent ARSE deletion that may be associated with short stature', *Andrology*, 10(8), pp. 1625–1631.

Casbon, J.A., Osborne, R.J., Brenner, S. & Lichtenstein, C.P. (2011) 'A method for counting PCR template molecules with application to next-generation sequencing', *Nucleic Acids Research*, 39(12), pp. e81–e81.

de Castro, F., Seal, R. & Maggi, R. (2017) 'ANOS1: a unified nomenclature for Kallmann syndrome 1 gene (KAL1) and anosmin-1', *Briefings in Functional Genomics*, 16(4), pp. 205–210.

Chapelle, A., Hortling, H., Niemi, M. & Wennström, J. (1964) 'XX Sex Chromosomes in a Human Male', *Acta Medica Scandinavica*, 175pp. 25–38.

Check Hayden, E. (2014) 'Technology: The $1,000 genome', *Nature*, 507(7492), pp. 294–295.

Chen, M., Yao, C., Qin, Y., Cui, X., Li, P., Ji, Z., Lin, L., Wu, H., Zhou, Z., Gui, Y., Li, Z. & Gao, F. (2022) 'Mutations of MSH5 in nonobstructive azoospermia (NOA) and rescued via in vivo gene editing', *Signal Transduction and Targeted Therapy*, 7(1), p. 1.

Chen, S., Wang, G., Zheng, X., Ge, S., Dai, Y., Ping, P., Chen, X., Liu, G., Zhang, J., Yang, Y., Zhang, X., Zhong, A., Zhu, Y., Chu, Q., Huang, Y., Zhang, Y., Shen, C., Yuan, Y., Yuan, Q., et al. (2020) 'Whole-exome sequencing of a large Chinese azoospermia and severe oligospermia cohort identifies novel infertility causative variants and genes', *Human Molecular Genetics*, 29(14), pp. 2451–2459.

Cho, S. & Dreyfuss, G. (2010) 'A degron created by *SMN2* exon 7 skipping is a principal contributor to spinal muscular atrophy severity', *Genes & Development*, 24(5), pp. 438–442.

Choy, J.T. & Amory, J.K. (2020) 'Nonsurgical Management of Oligozoospermia', *The Journal of Clinical Endocrinology & Metabolism*, 105(12), pp. e4194–e4207.

Cioppi, F., Rosta, V. & Krausz, C. (2021) 'Genetics of Azoospermia', *International Journal of Molecular Sciences*, 22(6), p. 3264.

Clavijo, R.I. & Hsiao, W. (2018) 'Update on male reproductive endocrinology', *Translational Andrology and Urology*, 7(S3), pp. S367–S372.

Colaco, S. & Modi, D. (2018) 'Genetics of the human Y chromosome and its association with male infertility', *Reproductive Biology and Endocrinology*, 16(1), p. 14.

Constine, L.S., Woolf, P.D., Cann, D., Mick, G., McCormick, K., Raubertas, R.F. & Rubin, P. (1993) 'Hypothalamic-Pituitary Dysfunction after Radiation for Brain Tumors', *New England Journal of Medicine*, 328(2), pp. 87–94.

Coutton, C., Escoffier, J., Martinez, G., Arnoult, C. & Ray, P.F. (2015) 'Teratozoospermia: spotlight on the main genetic actors in the human', *Human Reproduction Update*, 21(4), pp. 455–485.

Coutton, C., Vargas, A.S., Amiri-Yekta, A., Kherraf, Z.-E., Ben Mustapha, S.F., Le Tanno, P., Wambergue-Legrand, C., Karaouzène, T., Martinez, G., Crouzy, S., Daneshipour, A., Hosseini, S.H., Mitchell, V., Halouani, L., Marrakchi, O., Makni, M., Latrous, H., Kharouf, M., Deleuze, J.-F., et al. (2018) 'Mutations in CFAP43 and CFAP44 cause male

infertility and flagellum defects in Trypanosoma and human', *Nature Communications*, 9(1), p. 686.

Culard, J.-F., Desgeorges, M., Costa, P., Laussel, M., Razakatzara, G., Navratil, H., Demaille, J. & Claustres, M. (1994) 'Analysis of the whole CFTR coding regions and splice junctions in azoospermic men with congenital bilateral aplasia of epididymis or vas deferens', *Human Genetics*, 93(4), .

Dávalos, V., Súarez-López, L., Castaño, J., Messent, A., Abasolo, I., Fernandez, Y., Guerra-Moreno, A., Espín, E., Armengol, M., Musulen, E., Ariza, A., Sayós, J., Arango, D. & Schwartz, S. (2012) 'Human SMC2 Protein, a Core Subunit of Human Condensin Complex, Is a Novel Transcriptional Target of the WNT Signaling Pathway and a New Therapeutic Target', *Journal of Biological Chemistry*, 287(52), pp. 43472–43481.

Deák, G. & Cook, A.G. (2022) 'Missense Variants Reveal Functional Insights Into the Human ARID Family of Gene Regulators', *Journal of Molecular Biology*, 434(9), p. 167529.

Deciphering Developmental Disorders Study (2017) 'Prevalence and architecture of de novo mutations in developmental disorders', *Nature*, 542(7642), pp. 433–438.

Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C. V., Staples, J., Gonzaga-Jauregui, C., Metpally, R., Pendergrass, S.A., Giovanni, M.A., Kirchner, H.L., Balasubramanian, S., Abul-Husn, N.S., Hartzel, D.N., Lavage, D.R., Kost, K.A., et al. (2016) 'Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study', *Science*, 354(6319), .

Dicke, A.-K., Pilatz, A., Wyrwoll, M.J., Punab, M., Ruckert, C., Nagirnaja, L., Aston, K.I., Conrad, D.F., Di Persio, S., Neuhaus, N., Fietz, D., Laan, M., Stallmeyer, B. & Tüttelmann, F. (2023) 'DDX3Y is likely the key spermatogenic factor in the AZFa region that contributes to human non-obstructive azoospermia', *Communications Biology*, 6(1), p. 350.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014) 'Ten years of next-generation sequencing technology', *Trends in Genetics*, 30(9), pp. 418–426.

Dodé, C., Levilliers, J., Dupont, J.-M., De Paepe, A., Le Dû, N., Soussi-Yanicostas, N., Coimbra, R.S., Delmaghani, S., Compain-Nouaille, S., Baverel, F., Pêcheux, C., Le Tessier, D., Cruaud, C., Delpech, M., Speleman, F., Vermeulen, S., Amalfitano, A., Bachelot, Y., Bouchard, P., et al. (2003) 'Loss-of-function mutations in FGFR1 cause autosomal dominant Kallmann syndrome', *Nature Genetics*, 33(4), pp. 463–465.

Dumur, V., Gervais, R., Rigot, J.-M., Lafitte, J.-J., Manouvrier, S., Biserte, J., Mazeman, E. & Roussel, P. (1990) 'Abnormal distribution of CF ΔF508 allele in azoospermic men with congenital aplasia of epididymis and vas deferens', *The Lancet*, 336(8713), p. 512.

Dupont, C., Faure, C., Daoud, F., Gautier, B., Czernichow, S. & Lévy, R. (2019) 'Metabolic syndrome and smoking are independent risk factors of male idiopathic infertility', *Basic and Clinical Andrology*, 29(1), p. 9.

Eisenberg, M.L., Li, S., Behr, B., Pera, R.R. & Cullen, M.R. (2015) 'Relationship between semen production and medical comorbidity', *Fertility and Sterility*, 103(1), pp. 66–71.

Eisenberg, M.L., Li, S., Cullen, M.R. & Baker, L.C. (2016) ‘Increased risk of incident chronic medical conditions in infertile men: analysis of United States claims data’, *Fertility and Sterility*, 105(3), pp. 629–636.

Elsea, S.H. & Lucas, R.E. (2002) ‘The Mousetrap: What We Can Learn When the Mouse Model Does Not Mimic the Human Disease’, *ILAR Journal*, 43(2), pp. 66–79.

Esteves, S.C., Roque, M., Bedoschi, G., Haahr, T. & Humaidan, P. (2018) ‘Intracytoplasmic sperm injection for male infertility and consequences for offspring’, *Nature Reviews Urology*, 15(9), pp. 535–562.

Fabbri-Scallet, H., Sousa, L.M., Maciel-Guerra, A.T., Guerra-Júnior, G. & Mello, M.P. (2020) ‘Mutation update for the *NR5A1* gene involved in DSD and infertility’, *Human Mutation*, 41(1), pp. 58–68.

Fabre, A. & Mancini, J. (2022) ‘No preferential mode of inheritance for highly constrained genes’, *Intractable & Rare Diseases Research*, 11(1), p. 2022.01011.

Famini, P., Maya, M.M. & Melmed, S. (2011) ‘Pituitary Magnetic Resonance Imaging for Sellar and Parasellar Masses: Ten-Year Experience in 2598 Patients’, *The Journal of Clinical Endocrinology & Metabolism*, 96(6), pp. 1633–1641.

Ferlin, A., Vinanzi, C., Garolla, A., Selice, R., Zuccarello, D., Cazzadore, C. & Foresta, C. (2006) ‘Male infertility and androgen receptor gene mutations: clinical features and identification of seven novel mutations’, *Clinical Endocrinology*, 65(5), pp. 606–610.

Ferramosca, A. & Zara, V. (2022) ‘Diet and Male Fertility: The Impact of Nutrients and Antioxidants on Sperm Energetic Metabolism’, *International Journal of Molecular Sciences*, 23(5), p. 2542.

Fesahat, F., Henkel, R. & Agarwal, A. (2020) ‘Globozoospermia syndrome: An update’, *Andrologia*, 52(2), .

Fica, S.M. & Nagai, K. (2017) ‘Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine’, *Nature Structural & Molecular Biology*, 24(10), pp. 791–799.

Flannigan, R. & Schlegel, P.N. (2017) ‘Genetic diagnostics of male infertility in clinical practice’, *Best Practice & Research Clinical Obstetrics & Gynaecology*, 44pp. 26–37.

França, L.R., Hess, R.A., Dufour, J.M., Hofmann, M.C. & Griswold, M.D. (2016) ‘The Sertoli cell: one hundred fifty years of beauty and plasticity’, *Andrology*, 4(2), pp. 189–212.

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G., Slagboom, P.E., Boomsma, D.I., Ye, K., Guryev, V., Arndt, P.F., Kloosterman, W.P., de Bakker, P.I.W. & Sunyaev, S.R. (2015) ‘Genome-wide patterns and properties of de novo mutations in humans’, *Nature Genetics*, 47(7), pp. 822–826.

Franco, B., Guioli, S., Pragliola, A., Incerti, B., Bardoni, B., Tonlorenzi, R., Carrozzo, R., Maestrini, E., Pieretti, M., Taillon-Miller, P., Brown, C.J., Willard, H.F., Lawrence, C., Persico, M.G., Camerino, G. & Ballabio, A. (1991) ‘A gene deleted in Kallmann’s

syndrome shares homology with neural cell adhesion and axonal path-finding molecules', *Nature*, 353(6344), pp. 529–536.

Fridman, H., Yntema, H.G., Mägi, R., Andreson, R., Metspalu, A., Mezzavilla, M., Tyler-Smith, C., Xue, Y., Carmi, S., Levy-Lahad, E., Gilissen, C. & Brunner, H.G. (2021) 'The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects', *The American Journal of Human Genetics*, 108(4), pp. 608–619.

Fritzen, D., Kuechler, A., Grimmel, M., Becker, J., Peters, S., Sturm, M., Hundertmark, H., Schmidt, A., Kreiß, M., Strom, T.M., Wieczorek, D., Haack, T.B., Beck-Wödl, S., Cremer, K. & Engels, H. (2018) 'De novo FBXO11 mutations are associated with intellectual disability and behavioural anomalies', *Human Genetics*, 137(5), pp. 401–411.

Gallon, R., Sheth, H., Hayes, C., Redford, L., Alhilal, G., O'Brien, O., Spiewak, H., Waltham, A., McAnulty, C., Izuogu, O.G., Arends, M.J., Oniscu, A., Alonso, A.M., Laguna, S.M., Borthwick, G.M., Santibanez-Koref, M., Jackson, M.S. & Burn, J. (2020) 'Sequencing-based microsatellite instability testing using as few as six markers for high-throughput clinical diagnostics', *Human Mutation*, 41(1), pp. 332–341.

Gao, J., Zhang, H., Jiang, X., Ali, A., Zhao, D., Bao, J., Jiang, L., Iqbal, F., Shi, Q. & Zhang, Y. (2022) 'FertilityOnline: A Straightforward Pipeline for Functional Gene Annotation and Disease Mutation Discovery', *Genomics, Proteomics & Bioinformatics*, 20(3), pp. 455–465.

Garibotto, G., Esposito, P., Picciotto, D. & Verzola, D. (2021) 'Testosterone Disorders and Male Hypogonadism in Kidney Disease', *Seminars in Nephrology*, 41(2), pp. 114–125.

Gazvani, M.R., Wilson, E.D.A., Richmond, D.H., Howard, P.J., Kingsland, C.R. & Lewis-Jones, D.I. (2000) 'Role of mitotic control in spermatogenesis', *Fertility and Sterility*, 74(2), pp. 251–256.

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H. & Muller, J. (2018) 'AnnotSV: an integrated tool for structural variations annotation', *Bioinformatics*, 34(20), pp. 3572–3574.

Georgieva, L., Rees, E., Moran, J.L., Chambert, K.D., Milanova, V., Craddock, N., Purcell, S., Sklar, P., McCarroll, S., Holmans, P., O'Donovan, M.C., Owen, M.J. & Kirov, G. (2014) 'De novo CNVs in bipolar affective disorder and schizophrenia', *Human Molecular Genetics*, 23(24), pp. 6677–6683.

Gershoni, M., Hauser, R., Barda, S., Lehavi, O., Arama, E., Pietrokovski, S. & Kleiman, S.E. (2019) 'A new MEIOB mutation is a recurrent cause for azoospermia and testicular meiotic arrest', *Human Reproduction*, 34(4), pp. 666–671.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H.G., de Vries, B.B.A., Kleefstra, T., Brunner, H.G., et al. (2014) 'Genome sequencing identifies major causes of severe intellectual disability', *Nature*, 511(7509), pp. 344–347.

Glasser, E., Maji, D., Biancon, G., Puthenpeedikakkal, A.M.K., Cavender, C.E., Tebaldi, T., Jenkins, J.L., Mathews, D.H., Halene, S. & Kielkopf, C.L. (2022) 'Pre-mRNA splicing

factor U2AF2 recognizes distinct conformations of nucleotide variants at the center of the pre-mRNA splice site signal', *Nucleic Acids Research*, 50(9), pp. 5299–5312.

Glazer, C.H., Bonde, J.P., Giwercman, A., Vassard, D., Pinborg, A., Schmidt, L. & Vaclavik Bräuner, E. (2017) 'Risk of diabetes according to male factor infertility: a register-based cohort study', *Human Reproduction*, 32(7), pp. 1474–1481.

Goldmann, J.M., Wong, W.S.W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A.B., Glusman, G., Vissers, L.E.L.M., Hoischen, A., Roach, J.C., Vockley, J.G., Veltman, J.A., Solomon, B.D., Gilissen, C. & Niederhuber, J.E. (2016) 'Parent-of-origin-specific signatures of de novo mutations', *Nature Genetics*, 48(8), pp. 935–939.

Gonçalves, Catarina I, Aragüés, J.M., Bastos, M., Barros, L., Vicente, N., Carvalho, D. & Lemos, M.C. (2017) 'GNRHR biallelic and digenic mutations in patients with normosmic congenital hypogonadotropic hypogonadism', *Endocrine Connections*, 6(6), pp. 360–366.

Gonçalves, C.I., Fonseca, F., Borges, T., Cunha, F. & Lemos, M.C. (2017) 'Expanding the genetic spectrum of *ANOS1* mutations in patients with congenital hypogonadotropic hypogonadism', *Human Reproduction*,

Gottlieb, B., Beitel, L.K., Nadarajah, A., Paliouras, M. & Trifiro, M. (2012) 'The androgen receptor gene mutations database: 2012 update', *Human Mutation*, 33(5), pp. 887–894.

Gottlieb, B. & Trifiro, M.A. (1993) *Androgen Insensitivity Syndrome*.

Grinspon, R.P., Bergadá, I. & Rey, R.A. (2020) 'Male Hypogonadism and Disorders of Sex Development', *Frontiers in Endocrinology*, 11.

Grote, M., Wolf, E., Will, C.L., Lemm, I., Agafonov, D.E., Schomburg, A., Fischle, W., Urlaub, H. & Lührmann, R. (2010) 'Molecular Architecture of the Human Prp19/CDC5L Complex', *Molecular and Cellular Biology*, 30(9), pp. 2105–2119.

Grozdanov, P.N., Li, J., Yu, P., Yan, W. & MacDonald, C.C. (2018) '*Cstf2t* Regulates expression of histones and histone-like proteins in male germ cells', *Andrology*, 6(4), pp. 605–615.

Gu, J., Chen, Z., Chen, X. & Wang, Z. (2020) 'Heterogeneous nuclear ribonucleoprotein (hnRNPL) in cancer', *Clinica Chimica Acta*, 507pp. 286–294.

Gu, Y., Xiang, B., Zhu, L., Ma, X., Chen, X. & Cai, T. (2020) 'Three intellectual disability-associated de novo mutations in MECP2 identified by trio-WES analysis', *BMC Medical Genetics*, 21(1), p. 99.

Gulía, C., Baldassarra, S., Zangari, A., Briganti, V., Gigli, S., Gaffi, M., Signore, F., Vallone, C., Nucciotti, R., Costantini, F.M., Pizzuti, A., Bernardo, S., Porrello, A. & Piergentili, R. (2018) 'Androgen insensitivity syndrome.', *European review for medical and pharmacological sciences*, 22(12), pp. 3873–3887.

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., Wang, W., Li, R., Yan, J., Zhi, X., Zhang, Y., Jin, H., Zhang, W., Hou, Y., Zhu, P., et al. (2015) 'The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells', *Cell*, 161(6), pp. 1437–1452.

Guo, J., Grow, E.J., Mlcochova, H., Maher, G.J., Lindskog, C., Nie, X., Guo, Y., Takei, Y., Yun, J., Cai, L., Kim, R., Carrell, D.T., Goriely, A., Hotaling, J.M. & Cairns, B.R. (2018) 'The adult human testis transcriptional cell atlas', *Cell Research*, 28(12), pp. 1141–1157.

Guo, T., Zhao, Shidou, Zhao, Shigang, Chen, M., Li, G., Jiao, X., Wang, Z., Zhao, Y., Qin, Y., Gao, F. & Chen, Z.-J. (2017) 'Mutations in MSH5 in primary ovarian insufficiency', *Human Molecular Genetics*, 26(8), pp. 1452–1457.

Hamdan, F.F., Myers, C.T., Cossette, P., Lemay, P., Spiegelman, D., Laporte, A.D., Nassif, C., Diallo, O., Monlong, J., Cadieux-Dion, M., Dobrzeniecka, S., Meloche, C., Retterer, K., Cho, M.T., Rosenfeld, J.A., Bi, W., Massicotte, C., Miguet, M., Brunga, L., et al. (2017) 'High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies', *The American Journal of Human Genetics*, 101(5), pp. 664–685.

Hanson, H.A., Anderson, R.E., Aston, K.I., Carrell, D.T., Smith, K.R. & Hotaling, J.M. (2016) 'Subfertility increases risk of testicular cancer: evidence from population-based semen samples', *Fertility and Sterility*, 105(2), pp. 322-328.e1.

Hardy, J.J., Wyrwoll, M.J., Mcfadden, W., Malcher, A., Rotte, N., Pollock, N.C., Munyoki, S., Veroli, M. V., Houston, B.J., Xavier, M.J., Kasak, L., Punab, M., Laan, M., Kliesch, S., Schlegel, P., Jaffe, T., Hwang, K., Vukina, J., Brieño-Enríquez, M.A., et al. (2021) 'Variants in GCNA, X-linked germ-cell genome integrity gene, identified in men with primary spermatogenic failure', *Human Genetics*, 140(8), pp. 1169–1182.

Hassan, M.A.M. & Killick, S.R. (2003) 'Effect of male age on fertility: evidence for the decline in male fertility with increasing age', *Fertility and Sterility*, 79pp. 1520–1527.

Hayakawa, T., Zhang, F., Hayakawa, N., Ohtani, Y., Shinmyozu, K., Nakayama, J. & Andreassen, P.R. (2010) 'MRG15 binds directly to PALB2 and stimulates homology-directed repair of chromosomal breaks', *Journal of Cell Science*, 123(7), pp. 1124–1130.

He, Z., Kokkinaki, M., Pant, D., Gallicano, G.I. & Dym, M. (2009) 'Small RNA molecules in the regulation of spermatogenesis', *REPRODUCTION*, 137(6), pp. 901–911.

Hensel, N., Kubinski, S. & Claus, P. (2020) 'The Need for SMN-Independent Treatments of Spinal Muscular Atrophy (SMA) to Complement SMN-Enhancing Drugs', *Frontiers in Neurology*, 11.

Heo, Y.-A. (2023) 'Etranacogene Dezaparvovec: First Approval', *Drugs*, 83(4), pp. 347–352.

Hermann, B.P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K.N., Chen, I.-C., Gildersleeve, H., Lehle, J.D., Mayo, M., Westernströer, B., Law, N.C., Oatley, M.J., Velte, E.K., Niedenberger, B.A., Fritze, D., Silber, S., Geyer, C.B., Oatley, J.M. & McCarrey, J.R. (2018) 'The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids', *Cell Reports*, 25(6), pp. 1650-1667.e8.

Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. (2010) 'Parallel, tag-directed assembly of locally derived short sequence reads', *Nature Methods*, 7(2), pp. 119–122.

Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J. & Shendure, J. (2013) 'Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation', *Genome Research*, 23(5), pp. 843–854.

Hodžić, A., Maver, A., Plaseska-Karanfilska, D., Ristanović, M., Noveski, P., Zorn, B., Terzic, M., Kunej, T. & Peterlin, B. (2021) 'De novo mutations in idiopathic male infertility—A pilot study', *Andrology*, 9(1), pp. 212–220.

Hohl, A., Mazzuco, T.L., Coral, M.H.C., Schwarzbold, M. & Walz, R. (2009) 'Hypogonadism after traumatic brain injury', *Arquivos Brasileiros de Endocrinologia & Metabologia*, 53(8), pp. 908–914.

Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H., Gorham, J., Jin, S.C., Deanfield, J., Giardini, A., Porter, G.A., Kim, R., Bilguvar, K., López-Giráldez, F., Tikhonova, I., Mane, S., et al. (2015) 'De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies', *Science*, 350(6265), pp. 1262–1266.

Houston, B.J., Riera-Escamilla, A., Wyrwoll, M.J., Salas-Huetos, A., Xavier, M.J., Nagirnaja, L., Friedrich, C., Conrad, D.F., Aston, K.I., Krausz, C., Tüttelmann, F., O'Bryan, M.K., Veltman, J.A. & Oud, M.S. (2021) 'A systematic review of the validated monogenic causes of human male infertility: 2020 update and a discussion of emerging gene–disease relationships', *Human Reproduction Update*, 28(1), pp. 15–29.

Hughes, I.A., Davies, J.D., Bunch, T.I., Pasterski, V., Mastroyannopoulou, K. & MacDougall, J. (2012) 'Androgen insensitivity syndrome', *The Lancet*, 380(9851), pp. 1419–1428.

Hui, J., Stangl, K., Lane, W.S. & Bindereif, A. (2003) 'HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats', *Nature Structural Biology*, 10(1), pp. 33–37.

Human Fertilisation and Embryology Authority (HFEA) (2021) *Fertility treatment 2019: trends and figures*.

Ichioka, K., Yoshikawa, T., Kimura, H. & Saito, R. (2024) 'Additional mutation in *PROKR2* and phenotypic differences in a Kallmann syndrome/normosmic congenital hypogonadotropic hypogonadism family carrying *FGFR1* missense mutation', *BMJ Case Reports*, 17(1), p. e258042.

Ikawa, M., Tergaonkar, V., Ogura, A., Ogonuki, N., Inoue, K. & Verma, I.M. (2002) 'Restoration of spermatogenesis by lentiviral gene transfer: Offspring from infertile mice', *Proceedings of the National Academy of Sciences*, 99(11), pp. 7524–7529.

Illumina (2022) *Illumina Unveils Revolutionary NovaSeq X Series to Rapidly Accelerate Genomic Discoveries and Improve Human Health*. [Online] [online]. Available from: https://emea.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=8d04df3f-d9c1-4c85-8177-6ea604627ccd (Accessed 25 October 2022).

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., Smith, J.D., Paeper, B., Nickerson, D.A., Dea, J., Dong, S., Gonzalez, L.E., Mandell, J.D., Mane, S.M., Murtha, M.T., et al. (2014) 'The contribution of de novo coding mutations to autism spectrum disorder', *Nature*, 515(7526), pp. 216–221.

Iwamori, N., Tominaga, K., Sato, T., Riehle, K., Iwamori, T., Ohkawa, Y., Coarfa, C., Ono, E. & Matzuk, M.M. (2016) 'MRG15 is required for pre-mRNA splicing and spermatogenesis', *Proceedings of the National Academy of Sciences*, 113(37), .

Jacobs, P.A. & Strong, J.A. (1959) 'A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism', *Nature*, 183(4657), pp. 302–303.

Jamsai, D., Watkins, D.N., O'Connor, A.E., Merriner, D.J., Gursoy, S., Bird, A.D., Kumar, B., Miller, A., Cole, T.J., Jenkins, B.J. & O'Bryan, M.K. (2017) 'In vivo evidence that RBM5 is a tumour suppressor in the lung', *Scientific Reports*, 7(1), p. 16323.

Jayasena, C.N. & Sironen, A. (2021) 'Diagnostics and Management of Male Infertility in Primary Ciliary Dyskinesia', *Diagnostics*, 11(9), p. 1550.

Jedidi, I., Ouchari, M. & Yin, Q. (2019) 'Sex chromosomes-linked single-gene disorders involved in human infertility', *European Journal of Medical Genetics*, 62(9), p. 103560.

Jensen, T.K., Jacobsen, R., Christensen, K., Nielsen, N.C. & Bostofte, E. (2009) 'Good Semen Quality and Life Expectancy: A Cohort Study of 43,277 Men', *American Journal of Epidemiology*, 170(5), pp. 559–565.

Jiang, C., Zhang, X., Zhang, H., Guo, J., Zhang, C., Li, J. & Yang, Y. (2021) 'Novel bi-allelic mutations in DNAH1 cause multiple morphological abnormalities of the sperm flagella resulting in male infertility', *Translational Andrology and Urology*, 10(4), pp. 1656–1664.

Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., Hung, W.-C., Haider, S., Zhang, J., Knight, J., Bjornson, R.D., Castaldi, C., Tikhonoa, I.R., Bilguvar, K., Mane, S.M., et al. (2017) 'Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands', *Nature Genetics*, 49(11), pp. 1593–1601.

De Jonge, C. & Barratt, C.L.R. (2019) 'The present crisis in male reproductive health: an urgent need for a political, social, and research roadmap', *Andrology*, 7(6), pp. 762–768.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., Ward, L.D., Arnadottir, G.A., Helgason, E.A., Helgason, H., Gylfason, A., Jonasdottir, Adalbjorg, Jonasdottir, Aslaug, Rafnar, T., Frigge, M., et al. (2017) 'Parental influence on human germline de novo mutations in 1,548 trios from Iceland', *Nature*, 549(7673), pp. 519–522.

Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., Short, P.J., Torene, R.I., de Boer, E., Danecek, P., Gardner, E.J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., et al. (2020) 'Evidence for 28 genetic disorders discovered by combining healthcare and research data', *Nature*, 586(7831), pp. 757–762.

Kaprara, A. & Huhtaniemi, I.T. (2018) 'The hypothalamus-pituitary-gonad axis: Tales of mice and men', *Metabolism*, 86pp. 3–17.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G.,

Kosmicki, J.A., et al. (2020) ‘The mutational constraint spectrum quantified from variation in 141,456 humans’, *Nature*, 581(7809), pp. 434–443.

Kasak, L. & Laan, M. (2021) ‘Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives’, *Human Genetics*, 140(1), pp. 135–154.

Kasak, L., Lillepea, K., Nagirnaja, L., Aston, K.I., Schlegel, P.N., Gonçalves, J., Carvalho, F., Moreno-Mendoza, D., Almstrup, K., Eisenberg, M.L., Jarvi, K.A., O’Bryan, M.K., Lopes, A.M., Conrad, D.F., Conrad, D.F., Nagirnaja, L., Aston, K.I., Carrell, D.T., Hotaling, J.M., et al. (2022) ‘Actionable secondary findings following exome sequencing of 836 non-obstructive azoospermia cases and their value in patient management’, *Human Reproduction*, 37(7), pp. 1652–1663.

Kasman, A.M., Del Giudice, F. & Eisenberg, M.L. (2020) ‘New insights to guide patient care: the bidirectional relationship between male infertility and male health’, *Fertility and Sterility*, 113(3), pp. 469–477.

Katoh-Fukui, Y., Igarashi, M., Nagasaki, K., Horikawa, R., Nagai, T., Tsuchiya, T., Suzuki, E., Miyado, M., Hata, K., Nakabayashi, K., Hayashi, K., Matsubara, Y., Baba, T., Morohashi, K., Igarashi, A., Ogata, T., Takada, S. & Fukami, M. (2015) ‘Testicular dysgenesis/regression without campomelic dysplasia in patients carrying missense mutations and upstream deletion of *<scp>SOX</scp> 9* ’, *Molecular Genetics & Genomic Medicine*, 3(6), pp. 550–557.

Kent, O.A. & MacMillan, A.M. (2002) ‘Early organization of pre-mRNA during spliceosome assembly’, *Nature Structural Biology*,

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, and D. (2002) ‘The Human Genome Browser at UCSC’, *Genome Research*, 12(6), pp. 996–1006.

Kerem, B.-S., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. & Tsui, L.-C. (1989) ‘Identification of the Cystic Fibrosis Gene: Genetic Analysis’, *Science*, 245(4922), pp. 1073–1080.

Khan, M., Cornelis, S.S., Khan, M.I., Elmelik, D., Manders, E., Bakker, S., Derks, R., Neveling, K., Vorst, M., Gilissen, C., Meunier, I., Defoort, S., Puech, B., Devos, A., Schulz, H.L., Stöhr, H., Grassmann, F., Weber, B.H.F., Dhaenens, C., et al. (2019) ‘Cost-effective molecular inversion probe-based *ABCA4* sequencing reveals deep-intronic variants in Stargardt disease’, *Human Mutation*, 40(10), pp. 1749–1759.

Kherraf, Z.-E., Cazin, C., Bouker, A., Fourati Ben Mustapha, S., Hennebicq, S., Septier, A., Coutton, C., Raymond, L., Nouchy, M., Thierry-Mieg, N., Zouari, R., Arnoult, C. & Ray, P.F. (2022) ‘Whole-exome sequencing improves the diagnosis and care of men with non-obstructive azoospermia’, *The American Journal of Human Genetics*, 109(3), pp. 508–517.

Kibbe, W.A. (2007) ‘OligoCalc: an online oligonucleotide properties calculator’, *Nucleic Acids Research*, 35(Web Server), pp. W43–W46.

Kim, I. Bin, Lee, T., Lee, J., Kim, J., Lee, S., Koh, I.G., Kim, J.H., An, J.-Y., Lee, H., Kim, W.K., Ju, Y.S., Cho, Y., Yu, S.J., Kim, S.A., Oh, M., Han, D.W., Kim, E., Choi, J.K., Yoo,

H.J., et al. (2022) ‘Non-coding de novo mutations in chromatin interactions are implicated in autism spectrum disorder’, *Molecular Psychiatry*, 27(11), pp. 4680–4694.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. (2011) ‘Detection and quantification of rare mutations with massively parallel sequencing’, *Proceedings of the National Academy of Sciences*, 108(23), pp. 9530–9535.

Klinefelter, H.F., Reifenstein, E.C. & Albright, F. (1942) ‘Syndrome Characterized by Gynecomastia, Aspermatogenesis without A-Leydigism, and Increased Excretion of Follicle-Stimulating Hormone1’, *The Journal of Clinical Endocrinology & Metabolism*, 2(11), pp. 615–627.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, Aslaug, Jonasdottir, Adalbjorg, Wong, W.S.W., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.Th., et al. (2012) ‘Rate of de novo mutations and the importance of father’s age to disease risk’, *Nature*, 488(7412), pp. 471–475.

Koscinski, I., Wittemer, C., Lefebvre-Khalil, V., Marcelli, F., Defossez, A. & Rigot, J.M. (2007) ‘Optimal management of extreme oligozoospermia by an appropriate cryopreservation programme’, *Human Reproduction*, 22(10), pp. 2679–2684.

Krausz, C. (2011) ‘Male infertility: Pathogenesis and clinical diagnosis’, *Best Practice & Research Clinical Endocrinology & Metabolism*, 25(2), pp. 271–285.

Krausz, C., Degl’Innocenti, S., Nuti, F., Morelli, A., Felici, F., Sansone, M., Varriale, G. & Forti, G. (2006) ‘Natural transmission of USP9Y gene mutations: a new perspective on the role of AZFa genes in male fertility’, *Human Molecular Genetics*, 15(18), pp. 2673–2681.

Krausz, C. & Riera-Escamilla, A (2018) ‘Genetics of male infertility’, *Nat. Rev. Urol.*, 15.

Krausz, C. & Riera-Escamilla, Antoni (2018) ‘Genetics of male infertility’, *Nature Reviews Urology*, 15(6), pp. 369–384.

Krausz, C., Riera-Escamilla, A., Moreno-Mendoza, D., Holleman, K., Cioppi, F., Algaba, F., Pybus, M., Friedrich, C., Wyrwoll, M.J., Casamonti, E., Pietroforte, S., Nagirnaja, L., Lopes, A.M., Kliesch, S., Pilatz, A., Carrell, D.T., Conrad, D.F., Ars, E., Ruiz-Castañé, E., et al. (2020) ‘Genetic dissection of spermatogenic arrest through exome analysis: clinical implications for the management of azoospermic men’, *Genetics in Medicine*, 22(12), pp. 1956–1966.

Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.-X., Leal, S.M., Bernier, R. & Eichler, E.E. (2015) ‘Excess of rare, inherited truncating mutations in autism’, *Nature Genetics*, 47(6), pp. 582–588.

Kumar Mahat, R. & Arora, M. (2016) ‘Risk Factors and Causes of Male Infertility-A Review’, *Biochemistry & Analytical Biochemistry*, 5(2), .

Künzle, R., Mueller, M.D., Hänggi, W., Birkhäuser, M.H., Drescher, H. & Bersinger, N.A. (2003) ‘Semen quality of male smokers and nonsmokers in infertile couples’, *Fertility and Sterility*, 79(2), pp. 287–291.

Laan, M. (2019) ‘Systematic review of the monogenetic causes of male infertility: the first step towards diagnostic gene panels in the andrology clinic’, *Human Reproduction*, 34(5), pp. 783–785.

Laan, M., Grigorova, M. & Huhtaniemi, I.T. (2012) ‘Pharmacogenetics of follicle-stimulating hormone action’, *Current Opinion in Endocrinology, Diabetes & Obesity*, 19(3), pp. 220–227.

Laan, M., Kasak, L. & Punab, M. (2021) ‘Translational aspects of novel findings in genetics of male infertility—status quo 2021’, *British Medical Bulletin*, 140(1), pp. 5–22.

de Lange, I.M., Koudijs, M.J., van ‘t Slot, R., Sonsma, A.C.M., Mulder, F., Carbo, E.C., van Kempen, M.J.A., Nijman, I.J., Ernst, R.F., Savelberg, S.M.C., Knoers, N.V.A.M., Brilstra, E.H. & Koeleman, B.P.C. (2019) ‘Assessment of parental mosaicism in *SCN1A* -related epilepsy by single-molecule molecular inversion probes and next-generation sequencing’, *Journal of Medical Genetics*, 56(2), pp. 75–80.

Legouis, R., Hardelin, J.-P., Levilliers, J., Claverie, J.-M., Compain, S., Wunderle, V., Millasseau, P., Le Paslier, D., Cohen, D. & Caterina, D. (1991) ‘The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules’, *Cell*, 67(2), pp. 423–435.

Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Tukiainen, T., Birnbaum, D.P., Kosmicki, J.A., Duncan, L.E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., et al. (2016) ‘Analysis of protein-coding genetic variation in 60,706 humans’, *Nature*, 536(7616), pp. 285–291.

Lelieveld, S.H., Reijnders, M.R.F., Pfundt, R., Yntema, H.G., Kamsteeg, E.-J., de Vries, P., de Vries, B.B.A., Willemsen, M.H., Kleefstra, T., Löhner, K., Vreeburg, M., Stevens, S.J.C., van der Burgt, I., Bongers, E.M.H.F., Stegmann, A.P.A., Rump, P., Rinne, T., Nelen, M.R., Veltman, J.A., et al. (2016) ‘Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability’, *Nature Neuroscience*, 19(9), pp. 1194–1196.

Lelieveld, S.H., Spielmann, M., Mundlos, S., Veltman, J.A. & Gilissen, C. (2015) ‘Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions’, *Human Mutation*, 36(8), pp. 815–822.

Leslie, S.W., Mejias, S.G. & Ramphul, K. (2023) *Sertoli Cell-Only Syndrome*.

Leslie, S.W., Sajjad, H. & Siref, L.E. (2023) *Varicocele*.

Li, H. & Durbin, R. (2010) ‘Fast and accurate long-read alignment with Burrows–Wheeler transform’, *Bioinformatics*, 26(5), pp. 589–595.

Li, Jingping, Guo, W., Li, F., He, J., Yu, Q., Wu, X., Li, Jianming & Mao, X. (2012) ‘HnRNPL as a key factor in spermatogenesis: Lesson from functional proteomic studies of azoospermia patients with sertoli cell only syndrome’, *Journal of Proteomics*, 75(10), pp. 2879–2891.

Li, Li, Dong, J., Yan, L., Yong, J., Liu, X., Hu, Y., Fan, X., Wu, X., Guo, H., Wang, X., Zhu, X., Li, R., Yan, J., Wei, Y., Zhao, Y., Wang, W., Ren, Y., Yuan, P., Yan, Z., et al. (2017) ‘Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions’, *Cell Stem Cell*, 20(6), pp. 858-873.e4.

Li, Lin, Sha, Y., Wang, X., Li, P., Wang, J., Kee, K. & Wang, B. (2017) 'Whole-exome sequencing identified a homozygous BRDT mutation in a patient with acephalic spermatozoa', *Oncotarget*, 8(12), pp. 19914–19922.

Liang, H., Zhang, J., Shao, C., Zhao, L., Xu, W., Sutherland, L.C. & Wang, K. (2012) 'Differential Expression of RBM5, EGFR and KRAS mRNA and protein in non-small cell lung cancer tissues', *Journal of Experimental & Clinical Cancer Research*, 31(1), p. 36.

Liao, P., Satten, G.A. & Hu, Y.-J. (2017) 'PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies', *Genetic Epidemiology*, 41(5), pp. 375–387.

Lim, E.T., Uddin, M., De Rubeis, S., Chan, Y., Kamumbu, A.S., Zhang, X., D'Gama, A.M., Kim, S.N., Hill, R.S., Goldberg, A.P., Poultney, C., Minshew, N.J., Kushima, I., Aleksic, B., Ozaki, N., Parellada, M., Arango, C., Penzol, M.J., Carracedo, A., et al. (2017) 'Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder', *Nature Neuroscience*, 20(9), pp. 1217–1224.

Lin, L. & Achermann, J.C. (2008) 'Steroidogenic Factor-1 <i>(SF-1, Ad4BP, NR5A1)</i> and Disorders of Testis Development', *Sexual Development*, 2(4–5), pp. 200–209.

Lin, Y.-L., Shivji, M.K.K., Chen, C., Kolodner, R., Wood, R.D. & Dutta, A. (1998) 'The Evolutionarily Conserved Zinc Finger Motif in the Largest Subunit of Human Replication Protein A Is Required for DNA Replication and Mismatch Repair but Not for Nucleotide Excision Repair', *Journal of Biological Chemistry*, 273(3), pp. 1453–1461.

Lopes, A.M., Aston, K.I., Thompson, E., Carvalho, F., Gonçalves, J., Huang, N., Matthiesen, R., Noordam, M.J., Quintela, I., Ramu, A., Seabra, C., Wilfert, A.B., Dai, J., Downie, J.M., Fernandes, S., Guo, X., Sha, J., Amorim, A., Barros, A., et al. (2013) 'Human Spermatogenic Failure Purges Deleterious Mutation Load from the Autosomes and Both Sex Chromosomes, including the Gene DMRT1', *PLoS Genetics*, 9(3), p. e1003349.

Luangpraseuth-Prosper, A., Lesueur, E., Jouneau, L., Pailhoux, E., Cotinot, C. & Mandon-Pépin, B. (2015) 'TOPAZ1, a germ cell specific factor, is essential for male meiotic progression', *Developmental Biology*, 406(2), pp. 158–171.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M. & Misteli, T. (2010) 'Regulation of Alternative Splicing by Histone Modifications', *Science*, 327(5968), pp. 996–1000.

Luddi, A., Margollicci, M., Gambera, L., Serafini, F., Cioni, M., De Leo, V., Balestri, P. & Piomboni, P. (2009) 'Spermatogenesis in a Man with Complete Deletion of *USP9Y*', *New England Journal of Medicine*, 360(9), pp. 881–885.

M. Al-Ali, B., Gutschi, T., Pummer, K., Zigeuner, R., Brookman-May, S., Wieland, W.F., Fritsche, H.M. & Aziz, A. (2014) 'Body mass index has no impact on sperm quality but on reproductive hormones levels', *Andrologia*, 46(2), pp. 106–111.

Ma, Y., Xie, N., Xie, D., Sun, L., Li, S., Li, P., Li, Y., Li, J., Dong, Z. & Xie, X. (2019) 'A novel homozygous FBXO43 mutation associated with male infertility and

teratozoospermia in a consanguineous Chinese family', *Fertility and Sterility*, 111(5), pp. 909-917.e1.

Maggi, R., Cariboni, A.M., Marelli, M.M., Moretti, R.M., Andrè, V., Marzagalli, M. & Limonta, P. (2016) 'GnRH and GnRH receptors in the pathophysiology of the human female reproductive system', *Human Reproduction Update*, 22(3), pp. 358–381.

Maguire, A.M., Russell, S., Chung, D.C., Yu, Z.-F., Tillman, A., Drack, A. V., Simonelli, F., Leroy, B.P., Reape, K.Z., High, K.A. & Bennett, J. (2021) 'Durability of Voretigene Neparvovec for Biallelic RPE65-Mediated Inherited Retinal Disease', *Ophthalmology*, 128(10), pp. 1460–1468.

Mahajan, R. (2019) 'Onasemnogene abeparvovec for spinal muscular atrophy: The costlier drug ever', *International Journal of Applied and Basic Medical Research*, 9(3), p. 127.

Maji, D., Glasser, E., Henderson, S., Galardi, J., Pulvino, M.J., Jenkins, J.L. & Kielkopf, C.L. (2020) 'Representative cancer-associated U2AF2 mutations alter RNA interactions and splicing', *Journal of Biological Chemistry*, 295(50), pp. 17148–17157.

Mann, J.M., Wei, C. & Chen, C. (2022) 'How genetic defects in piRNA trimming contribute to male infertility', *Andrology*,

Mao, J.-F., Liu, Z.-X., Nie, M., Wang, X., Xu, H.-L., Huang, B.-K., Zheng, J.-J., Min, L., Kaiser, U. & Wu, X.-Y. (2017) 'Pulsatile gonadotropin-releasing hormone therapy is associated with earlier spermatogenesis compared to combined gonadotropin therapy in patients with congenital hypogonadotropic hypogonadism', *Asian Journal of Andrology*, 19(6), p. 680.

Maréchal, A., Li, J.-M., Ji, X.Y., Wu, C.-S., Yazinski, S.A., Nguyen, H.D., Liu, S., Jiménez, A.E., Jin, J. & Zou, L. (2014) 'PRP19 Transforms into a Sensor of RPA-ssDNA after DNA Damage and Drives ATR Activation via a Ubiquitin-Mediated Circuitry', *Molecular Cell*, 53(2), pp. 235–246.

Martin-Merida, I., Avila-Fernandez, A., Del Pozo-Valero, M., Blanco-Kelly, F., Zurita, O., Perez-Carro, R., Aguilera-Garcia, D., Riveiro-Alvarez, R., Arteche, A., Trujillo-Tiebas, M.J., Tahsin-Swafiri, S., Rodriguez-Pinilla, E., Lorda-Sanchez, I., Garcia-Sandoval, B., Corton, M. & Ayuso, C. (2019) 'Genomic Landscape of Sporadic Retinitis Pigmentosa', *Ophthalmology*, 126(8), pp. 1181–1188.

Masarani, M., Wazait, H. & Dinneen, M. (2006) 'Mumps orchitis', *Journal of the Royal Society of Medicine*, 99(11), pp. 573–575.

McCarthy, S.E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., Mistry, M., Pavlidis, P., Solomon, R., Ghiban, E., Antoniou, E., Kelleher, E., O'Brien, C., Donohoe, G., Gill, M., Morris, D.W., McCombie, W.R. & Corvin, A. (2014) 'De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability', *Molecular Psychiatry*, 19(6), pp. 652–658.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. (2010) 'The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, 20(9), pp. 1297–1303.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. & Cunningham, F. (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), p. 122.

Medeiros, J.J.F., Capo-Chichi, J.-M., Shlush, L.I., Dick, J.E., Arruda, A., Minden, M.D. & Abelson, S. (2022) 'SmMIP-tools: a computational toolset for processing and analysis of single-molecule molecular inversion probes-derived data', *Bioinformatics*, 38(8), pp. 2088–2095.

Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. (2016) 'Clinical sequencing: is WGS the better WES?', *Human Genetics*, 135(3), pp. 359–362.

Mészáros, B., Kumar, M., Gibson, T.J., Uyar, B. & Dosztányi, Z. (2017) 'Degrons in cancer', *Science Signaling*, 10(470), .

Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, Á., Koren, A., Gore, A., Kang, S., Lin, G.N., Estabillo, J., Gadomski, T., et al. (2012) 'Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation', *Cell*, 151(7), pp. 1431–1442.

Michalakis, K., Mintziori, G., Kaprara, A., Tarlatzis, B.C. & Goulis, D.G. (2013) 'The complex interaction between obesity, metabolic syndrome and reproductive axis: A narrative review', *Metabolism*, 62(4), pp. 457–478.

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M. & Vagner, S. (2006) 'An interaction between U2AF 65 and CF Im links the splicing and 3′ end processing machineries', *The EMBO Journal*, 25(20), pp. 4854–4864.

Miyamoto, T., Hasuike, S., Yogev, L., Maduro, M.R., Ishikawa, M., Westphal, H. & Lamb, D.J. (2003) 'Azoospermia in patients heterozygous for a mutation in SYCP3', *The Lancet*, 362(9397), pp. 1714–1719.

MOLLINEDO, F. & GAJATE, C. (2006) 'Fas/CD95 death receptor and lipid rafts: New targets for apoptosis-directed cancer therapy', *Drug Resistance Updates*, 9(1–2), pp. 51–73.

Montagne, J.M., Jaffee, E.M. & Fertig, E.J. (2023) 'Multiomics Empowers Predictive Pancreatic Cancer Immunotherapy', *The Journal of Immunology*, 210(7), pp. 859–868.

Mordes, D., Luo, X., Kar, A., Kuo, D., Xu, L., Fushimi, K., Yu, G., Sternberg, P. & Wu, J.Y. (2006) 'Pre-mRNA splicing and retinitis pigmentosa.', *Molecular vision*, 12pp. 1259–71.

Morris, G.C., Lloyd-Evans, E. & Cahill, D.J. (2021) 'Induction of spermatogenesis in men with hypogonadotropic hypogonadism', *Journal of Assisted Reproduction and Genetics*, 38(4), pp. 803–807.

Mu, R., Wang, Y.-B., Wu, M., Yang, Y., Song, W., Li, T., Zhang, W.-N., Tan, B., Li, A.-L., Wang, N., Xia, Q., Gong, W.-L., Wang, C.-G., Zhou, T., Guo, N., Sang, Z.-H. & Li, H.-Y. (2014) 'Depletion of pre-mRNA splicing factor Cdc5L inhibits mitotic progression and triggers mitotic catastrophe', *Cell Death & Disease*, 5(3), pp. e1151–e1151.

Müller, K., Oh, K.-W., Nordin, A., Panthi, S., Kim, S.H., Nordin, F., Freischmidt, A., Ludolph, A.C., Ki, C.S., Forsberg, K., Weishaupt, J., Kim, Y.-E. & Andersen, P.M.

(2022) ‘De novo mutations in *SOD1* are a cause of ALS’, *Journal of Neurology, Neurosurgery & Psychiatry*, 93(2), pp. 201–206.

Muñoz, X., Mata, A., Bassas, L. & Larriba, S. (2015) ‘Altered miRNA Signature of Developing Germ-cells in Infertile Patients Relates to the Severity of Spermatogenic Failure and Persists in Spermatozoa’, *Scientific Reports*, 5(1), p. 17991.

Nagirnaja, L., Lopes, A.M., Charng, W.-L., Miller, B., Stakaitis, R., Golubickaite, I., Stendahl, A., Luan, T., Friedrich, C., Mahyari, E., Fadial, E., Kasak, L., Vigh-Conrad, K., Oud, M.S., Xavier, M.J., Cheers, S.R., James, E.R., Guo, J., Jenkins, T.G., et al. (2022) ‘Diverse monogenic subforms of human spermatogenic failure’, *Nature Communications*, 13(1), p. 7953.

Nagirnaja, L., Mørup, N., Nielsen, J.E., Stakaitis, R., Golubickaite, I., Oud, M.S., Winge, S.B., Carvalho, F., Aston, K.I., Khani, F., van der Heijden, G.W., Marques, C.J., Skakkebaek, N.E., Rajpert–De Meyts, E., Schlegel, P.N., Jørgensen, N., Veltman, J.A., Lopes, A.M., Conrad, D.F., et al. (2021) ‘Variant *PNLDC1* , Defective piRNA Processing, and Azoospermia’, *New England Journal of Medicine*, 385(8), pp. 707–719.

Naro, C., Cesari, E. & Sette, C. (2021) ‘Splicing regulation in brain and testis: common themes for highly specialized organs’, *Cell Cycle*, 20(5–6), pp. 480–489.

Neale, B.M., Kou, Y., Liu, L., Ma’ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E.L., Campbell, N.G., Geller, E.T., Valladares, O., Schafer, C., Liu, H., et al. (2012) ‘Patterns and rates of exonic de novo mutations in autism spectrum disorders’, *Nature*, 485(7397), pp. 242–245.

Neto, F.T.L., Bach, P.V., Najari, B.B., Li, P.S. & Goldstein, M. (2016) ‘Spermatogenesis in humans and its affecting factors’, *Seminars in Cell & Developmental Biology*, 59pp. 10–26.

Neveling, K., Mensenkamp, A.R., Derks, R., Kwint, M., Ouchene, H., Steehouwer, M., van Lier, B., Bosgoed, E., Rikken, A., Tychon, M., Zafeiropoulou, D., Castelein, S., Hehir-Kwa, J., Tjwan Thung, D., Hofste, T., Lelieveld, S.H., Bertens, S.M.M., Adan, I.B.J.F., Eijkelenboom, A., et al. (2017) ‘BRCA Testing by Single-Molecule Molecular Inversion Probes’, *Clinical Chemistry*, 63(2), pp. 503–512.

Newman, L., Chopra, J., Dossett, C., Shepherd, E., Bercusson, A., Carroll, M., Walker, W., Lucas, J.S. & Cheong, Y. (2023) ‘The impact of primary ciliary dyskinesia on female and male fertility: a narrative review’, *Human Reproduction Update*, 29(3), pp. 347–367.

Nickkholgh, B., Korver, C.M., van Daalen, S.K.M., van Pelt, A.M.M. & Repping, S. (2015) ‘AZFc deletions do not affect the function of human spermatogonia in vitro’, *Molecular Human Reproduction*, 21(7), pp. 553–562.

Nilsen, T.W. & Graveley, B.R. (2010) ‘Expansion of the eukaryotic proteome by alternative splicing’, *Nature*, 463(7280), pp. 457–463.

Niu, Z., Jin, W., Zhang, L. & Li, X. (2012) ‘Tumor suppressor RBM5 directly interacts with the DExD/H-box protein DHX15 and stimulates its helicase activity’, *FEBS Letters*, 586(7), pp. 977–983.

Norman, A.W. & Henry, H.L. (2015) ‘Androgens’, in *Hormones*. [Online]. Elsevier. pp. 255–273.

O’Bryan, M.K., Clark, B.J., McLaughlin, E.A., D’Sylva, R.J., O’Donnell, L., Wilce, J.A., Sutherland, J., O’Connor, A.E., Whittle, B., Goodnow, C.C., Ormandy, C.J. & Jamsai, D. (2013) ‘RBM5 Is a Male Germ Cell Splicing Factor and Is Required for Spermatid Differentiation and Male Fertility’, *PLoS Genetics*, 9(7), p. e1003628.

O’Donnell, L., Smith, L.B. & Rebourcet, D. (2022) ‘Sertoli cells as key drivers of testis function’, *Seminars in Cell & Developmental Biology*, 121pp. 2–9.

Oh, J.J., Razfar, A., Delgado, I., Reed, R.A., Malkina, A., Boctor, B. & Slamon, D.J. (2006) ‘3p21.3 Tumor Suppressor Gene *H37/Luca15/RBM5* Inhibits Growth of Human Lung Cancer Cells through Cell Cycle Arrest and Apoptosis’, *Cancer Research*, 66(7), pp. 3419–3427.

Okutman, O., Muller, J., Baert, Y., Serdarogullari, M., Gultomruk, M., Piton, A., Rombaut, C., Benkhalifa, M., Teletin, M., Skory, V., Bakircioglu, E., Goossens, E., Bahceci, M. & Viville, S. (2015) ‘Exome sequencing reveals a nonsense mutation in TEX15 causing spermatogenic failure in a Turkish family’, *Human Molecular Genetics*, 24(19), pp. 5581–5588.

Olesen, I.A., Andersson, A.-M., Aksglaede, L., Skakkebaek, N.E., Rajpert–de Meyts, E., Joergensen, N. & Juul, A. (2017) ‘Clinical, genetic, biochemical, and testicular biopsy findings among 1,213 men evaluated for infertility’, *Fertility and Sterility*, 107(1), pp. 74-82.e7.

O’Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., Bernier, R., Shendure, J. & Eichler, E.E. (2014) ‘Recurrent de novo mutations implicate novel genes underlying simplex autism risk’, *Nature Communications*, 5(1), p. 5595.

O’Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J.B., Turner, E.H., Levy, R., O’Day, D.R., Krumm, N., Coe, B.P., Martin, B.K., Borenstein, E., et al. (2012) ‘Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders’, *Science*, 338(6114), pp. 1619–1622.

O’Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., Turner, E.H., Stanaway, I.B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J.M., Borenstein, E., Rieder, M.J., et al. (2012) ‘Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations’, *Nature*, 485(7397), pp. 246–250.

Oud, M.S., Houston, B.J., Volozonoka, L., Mastrorosa, F.K., Holt, G.S., Alobaidi, B.K.S., deVries, P.F., Astuti, G., Ramos, L., Mclachlan, R.I., O’Bryan, M.K., Veltman, J.A., Chemes, H.E. & Sheth, H. (2021) ‘Exome sequencing reveals variants in known and novel candidate genes for severe sperm motility disorders’, *Human Reproduction*, 36(9), pp. 2597–2611.

Oud, M.S., Ramos, L., O’Bryan, M.K., McLachlan, R.I., Okutman, Ö., Viville, S., Vries, P.F., Smeets, D.F.C.M., Lugtenberg, D., Hehir-Kwa, J.Y., Gilissen, C., de Vorst, M., Vissers, L.E.L.M., Hoischen, A., Meijerink, A.M., Fleischer, K., Veltman, J.A. & Noordam, M.J.

(2017) ‘Validation and application of a novel integrated genetic screening method to a cohort of 1,112 men with idiopathic azoospermia or severe oligozoospermia’, *Human Mutation*, 38(11), pp. 1592–1605.

Oud, M.S., Smits, R.M., Smith, H.E., Mastrorosa, F.K., Holt, G.S., Houston, B.J., de Vries, P.F., Alobaidi, B.K.S., Batty, L.E., Ismail, H., Greenwood, J., Sheth, H., Mikulasova, A., Astuti, G.D.N., Gilissen, C., McEleny, K., Turner, H., Coxhead, J., Cockell, S., et al. (2022) ‘A de novo paradigm for male infertility’, *Nature Communications*, 13(1), p. 154.

Oud, M.S., Volozonoka, L., Smits, R.M., Vissers, L.E.L.M., Ramos, L. & Veltman, J.A. (2019) ‘A systematic review and standardized clinical validity assessment of male infertility genes’, *Human Reproduction*, 34(5), pp. 932–941.

Oved, J.H., Babushok, D. V., Lambert, M.P., Wolfset, N., Kowalska, M.A., Poncz, M., Karczewski, K.J. & Olson, T.S. (2020) ‘Human mutational constraint as a tool to understand biology of rare and emerging bone marrow failure syndromes’, *Blood Advances*, 4(20), pp. 5232–5245.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. (2008) ‘Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing’, *Nature Genetics*, 40(12), pp. 1413–1415.

Patat, O., Pagin, A., Siegfried, A., Mitchell, V., Chassaing, N., Faguer, S., Monteil, L., Gaston, V., Bujan, L., Courtade-Saïdi, M., Marcelli, F., Lalau, G., Rigot, J.-M., Mieusset, R. & Bieth, E. (2016) ‘Truncating Mutations in the Adhesion G Protein-Coupled Receptor G2 Gene ADGRG2 Cause an X-Linked Congenital Bilateral Absence of Vas Deferens’, *The American Journal of Human Genetics*, 99(2), pp. 437–442.

Pathak, P., Chandrashekar, A., Hakky, T. & Pastuszak, A. (2016) ‘Varicocele management in the era of in vitro fertilization/intracytoplasmic sperm injection’, *Asian Journal of Andrology*, 18(3), p. 343.

Pedersen, B.S. & Quinlan, A.R. (2017) ‘Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy’, *The American Journal of Human Genetics*, 100(3), pp. 406–413.

Pellizzoni, L., Kataoka, N., Charroux, B. & Dreyfuss, G. (1998) ‘A Novel Function for SMN, the Spinal Muscular Atrophy Disease Gene Product, in Pre-mRNA Splicing’, *Cell*, 95(5), pp. 615–624.

Peña, A.N., Tominaga, K. & Pereira-Smith, O.M. (2011) ‘MRG15 activates the cdc2 promoter via histone acetylation in human cells’, *Experimental Cell Research*, 317(11), pp. 1534–1540.

Persily, J.B., Vijay, V. & Najari, B.B. (2021) ‘How do we counsel men with obstructive azoospermia due to CF mutations?-a review of treatment options and outcomes.’, *Translational andrology and urology*, 10(3), pp. 1467–1478.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K., Pandey, A. & Salzberg, S.L. (2018) ‘CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise’, *Genome Biology*, 19(1), p. 208.

Pierik, F.H., Vreeburg, J.T.M., Stijnen, T., de Jong, F.H. & Weber, R.F.A. (1998) 'Serum Inhibin B as a Marker of Spermatogenesis', *The Journal of Clinical Endocrinology & Metabolism*, 83(9), pp. 3110–3114.

Pogoda, M., Hilke, F.-J., Lohmann, E., Sturm, M., Lenz, F., Matthes, J., Muyas, F., Ossowski, S., Hoischen, A., Faust, U., Sepahi, I., Casadei, N., Poths, S., Riess, O., Schroeder, C. & Grundmann, K. (2019) 'Single Molecule Molecular Inversion Probes for High Throughput Germline Screenings in Dystonia', *Frontiers in Neurology*, 10.

Poplin, R., Ruano-Rubio, V., DePristo, M., Fennell, T., Carneiro, M., Van der Auwera, G., Kling, D., Gauthier, L., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M., Neale, B., MacArthur, D., et al. (2017) *Scaling accurate genetic variant discovery to tens of thousands of samples*.

Prokop, J.W., May, T., Strong, K., Bilinovich, S.M., Bupp, C., Rajasekaran, S., Worthey, E.A. & Lazar, J. (2018) 'Genome sequencing in the clinic: the past, present, and future of genomic medicine', *Physiological Genomics*, 50(8), pp. 563–579.

Punab, M., Poolamets, O., Paju, P., Vihljajev, V., Pomm, K., Ladva, R., Korrovits, P. & Laan, M. (2016) 'Causes of male infertility: a 9-year prospective monocentre study on 1737 patients with reduced total sperm counts', *Human Reproduction*,

R Core Team (2020) *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. [Online]

Rahnema, C.D., Lipshultz, L.I., Crosnoe, L.E., Kovac, J.R. & Kim, E.D. (2014) 'Anabolic steroid–induced hypogonadism: diagnosis and treatment', *Fertility and Sterility*, 101(5), pp. 1271–1279.

Rajan-Babu, I.-S., Peng, J.J., Chiu, R., Birch, P., Couse, M., Guimond, C., Lehman, A., Mwenifumbo, J., van Karnebeek, C., Friedman, J., Adam, S., Souich, C. Du, Elliott, A., Lehman, A., Mwenifumbo, J., Nelson, T., van Karnebeek, C., Friedman, J., Li, C., et al. (2021) 'Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions', *Genome Medicine*, 13(1), p. 126.

Ramlau-Hansen, C.H., Thulstrup, A.M., Aggerholm, A.S., Jensen, M.S., Toft, G. & Bonde, J.P. (2007) 'Is smoking a risk factor for decreased semen quality? A cross-sectional analysis', *Human Reproduction*, 22(1), pp. 188–196.

Reijo, R., Lee, T.-Y., Salo, P., Alagappan, R., Brown, L.G., Rosenberg, M., Rozen, S., Jaffe, T., Straus, D., Hovatta, O., de la Chapelle, A., Silber, S. & Page, D.C. (1995) 'Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA–binding protein gene', *Nature Genetics*, 10(4), pp. 383–393.

Ren, P., Lu, L., Cai, S., Chen, J., Lin, W. & Han, F. (2021) 'Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease', *Frontiers in Immunology*, 12.

Reynolds, N. & Cooke, H.J. (2005) 'Role of the DAZ genes in male fertility', *Reproductive BioMedicine Online*, 10(1), pp. 72–80.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. & Rehm, H.L. (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the

American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genetics in Medicine*, 17(5), pp. 405–424.

Riera-Escamilla, A., Vockel, M., Nagirnaja, L., Xavier, M.J., Carbonell, A., Moreno-Mendoza, D., Pybus, M., Farnetani, G., Rosta, V., Cioppi, F., Friedrich, C., Oud, M.S., van der Heijden, G.W., Soave, A., Diemer, T., Ars, E., Sánchez-Curbelo, J., Kliesch, S., O'Bryan, M.K., et al. (2022) 'Large-scale analyses of the X chromosome in 2,354 infertile men discover recurrently affected genes associated with spermatogenic failure', *The American Journal of Human Genetics*, 109(8), pp. 1458–1471.

Riordan, J.R., Rommens, J.M., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.-L., Drumm, M.L., Iannuzzi, M.C., Collins, F.S. & Tsui, L.-C. (1989) 'Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA', *Science*, 245(4922), pp. 1066–1073.

Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L.B., Hood, L. & Galas, D.J. (2010) 'Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing', *Science*, 328(5978), pp. 636–639.

Rocca, M.S., Minervini, G., Vinanzi, C., Bottacin, A., Lia, F., Foresta, C., Pennuto, M. & Ferlin, A. (2023) 'Mutational Screening of Androgen Receptor Gene in 8224 Men of Infertile Couples', *The Journal of Clinical Endocrinology & Metabolism*, 108(5), pp. 1181–1191.

Rocca, M.S., Msaki, A., Ghezzi, M., Cosci, I., Pilichou, K., Celeghin, R., Foresta, C. & Ferlin, A. (2020) 'Development of a novel next-generation sequencing panel for diagnosis of quantitative spermatogenic impairment', *Journal of Assisted Reproduction and Genetics*, 37(4), pp. 753–762.

Rojas, C. V., Wang, J., Schwartz, L.S., Hoffman, E.P., Powell, B.R. & Brown Jr, R.H. (1991) 'A Met-to-Val mutation in the skeletal muscle Na+ channel α-subunit in hyperkalaemic periodic paralysis', *Nature*, 354(6352), pp. 387–389.

Röpke, A. & Tüttelmann, F. (2017) 'MECHANISMS IN ENDOCRINOLOGY: Aberrations of the X chromosome as cause of male infertility', *European Journal of Endocrinology*, 177(5), pp. R249–R259.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Fu, S.-C., Aleksic, B., Biscaldi, M., Bolton, P.F., Brownfeld, J.M., Cai, J., et al. (2014) 'Synaptic, transcriptional and chromatin genes disrupted in autism', *Nature*, 515(7526), pp. 209–215.

Růžičková, Š. & Staněk, D. (2017) 'Mutations in spliceosomal proteins and retina degeneration', *RNA Biology*, 14(5), pp. 544–552.

Salas-Huetos, A., Tüttelmann, F., Wyrwoll, M.J., Kliesch, S., Lopes, A.M., Goncalves, J., Boyden, S.E., Wöste, M., Hotaling, J.M., Nagirnaja, L., Conrad, D.F., Carrell, D.T. & Aston, K.I. (2021) 'Disruption of human meiotic telomere complex genes TERB1, TERB2 and MAJIN in men with non-obstructive azoospermia', *Human Genetics*, 140(1), pp. 217–227.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., Wall, D.P., MacArthur, D.G., Gabriel, S.B., DePristo, M., Purcell, S.M., Palotie, A., Boerwinkle, E., Buxbaum, J.D., Cook, E.H., et al. (2014) 'A framework for the interpretation of de novo mutation in human disease', *Nature Genetics*, 46(9), pp. 944–950.

Sang, Q., Ray, P.F. & Wang, L. (2023) 'Understanding the genetics of human infertility', *Science*, 380(6641), pp. 158–163.

Sarfati, J., Dodé, C. & Young, J. (2010) *Kallmann Syndrome Caused by Mutations in the <i>PROK2</i> and <i>PROKR2</i> Genes: Pathophysiology and Genotype-Phenotype Correlations*, in [Online]. pp. 121–132.

Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M.S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., et al. (2020) 'Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism', *Cell*, 180(3), pp. 568-584.e23.

Scanlan, M.J., Gordan, J.D., Williamson, B., Stockert, E., Bander, N.H., Jongeneel, V., Gure, A.O., Jager, D., Jager, E., Knuth, A., Chen, Y.-T. & Old, L.J. (1999) 'Antigens recognized by autologous antibody in patients with renal-cell carcinoma', *International Journal of Cancer*, 83(4), pp. 456–464.

Schilit, S.L.P., Menon, S., Friedrich, C., Kammin, T., Wilch, E., Hanscom, C., Jiang, S., Kliesch, S., Talkowski, M.E., Tüttelmann, F., MacQueen, A.J. & Morton, C.C. (2020) 'SYCP2 Translocation-Mediated Dysregulation and Frameshift Variants Cause Human Male Infertility', *The American Journal of Human Genetics*, 106(1), pp. 41–57.

Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. (2010) 'MutationTaster evaluates disease-causing potential of sequence alterations', *Nature Methods*, 7(8), pp. 575–576.

Scott, F.L., Stec, B., Pop, C., Dobaczewska, M.K., Lee, J.J., Monosov, E., Robinson, H., Salvesen, G.S., Schwarzenbacher, R. & Riedl, S.J. (2009) 'The Fas–FADD death domain complex structure unravels signalling by receptor clustering', *Nature*, 457(7232), pp. 1019–1022.

Seabra, C.M., Quental, S., Neto, A.P., Carvalho, F., Gonçalves, J., Oliveira, J.P., Fernandes, S., Sousa, M., Barros, A., Amorim, A. & Lopes, A.M. (2014) 'A novel Alu-mediated microdeletion at 11p13 removes WT1 in a patient with cryptorchidism and azoospermia', *Reproductive BioMedicine Online*, 29(3), pp. 388–391.

Seo, J., Singh, N.N., Ottesen, E.W., Lee, B.M. & Singh, R.N. (2016) 'A novel human-specific splice isoform alters the critical C-terminus of Survival Motor Neuron protein', *Scientific Reports*, 6(1), p. 30778.

Sermondade, N., Faure, C., Fezeu, L., Shayeb, A.G., Bonde, J.P., Jensen, T.K., Van Wely, M., Cao, J., Martini, A.C., Eskandar, M., Chavarro, J.E., Koloszar, S., Twigt, J.M., Ramlau-Hansen, C.H., Borges, E., Lotti, F., Steegers-Theunissen, R.P.M., Zorn, B., Polotsky, A.J., et al. (2013) 'BMI in relation to sperm count: an updated systematic review and collaborative meta-analysis', *Human Reproduction Update*, 19(3), pp. 221–231.

Shalet, S.M. (2009) ‘Normal testicular function and spermatogenesis’, *Pediatric Blood & Cancer*, 53(2), pp. 285–288.

Sharma, A., Minhas, S., Dhillo, W.S. & Jayasena, C.N. (2021) ‘Male infertility due to testicular disorders’, *The Journal of Clinical Endocrinology & Metabolism*, 106(2), pp. e442–e459.

Sharma, R. & Agarwal, A. (2011) ‘Spermatogenesis: An Overview’, in *Sperm Chromatin*. [Online]. New York, NY: Springer New York. pp. 19–44.

Sharma, R., Harlev, A., Agarwal, A. & Esteves, S.C. (2016) ‘Cigarette Smoking and Semen Quality: A New Meta-analysis Examining the Effect of the 2010 World Health Organization Laboratory Methods for the Examination of Human Semen’, *European Urology*, 70(4), pp. 635–645.

Shen, P., Wang, W., Krishnakumar, S., Palm, C., Chi, A.-K., Enns, G.M., Davis, R.W., Speed, T.P., Mindrinos, M.N. & Scharfe, C. (2011) ‘High-quality DNA sequence capture of 524 disease candidate genes’, *Proceedings of the National Academy of Sciences*, 108(16), pp. 6549–6554.

Sherry, S.T., Ward, M. & Sirotkin, K. (1999) ‘dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.’, *Genome research*, 9(8), pp. 677–9.

Shi, Y. (2017) ‘Mechanistic insights into precursor messenger RNA splicing by the spliceosome’, *Nature Reviews Molecular Cell Biology*, 18(11), pp. 655–670.

Singh, V., Agrawal, N.K., Verma, R. & Singh, K. (2017) ‘HPG Axis: The Central Regulator of Spermatogenesis and Male Fertility’, in *Male Infertility: Understanding, Causes and Treatment*. [Online]. Singapore: Springer Singapore. pp. 25–36.

Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., Vandrovcova, J., Chan, G., Williams, H.J., Ratnaike, T., Wei, W., Stirrups, K., Ibanez, K., Moutsianas, L., Wielscher, M., Need, A., et al. (2021) ‘100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report’, *New England Journal of Medicine*, 385(20), pp. 1868–1880.

Smits, R.M., Xavier, M.J., Oud, M.S., Astuti, G.D.N., Meijerink, A.M., de Vries, P.F., Holt, G.S., Alobaidi, B.K.S., Batty, L.E., Khazeeva, G., Sablauskas, K., Vissers, L.E.L.M., Gilissen, C., Fleischer, K., Braat, D.D.M., Ramos, L. & Veltman, J.A. (2022) ‘De novo mutations in children born after medical assisted reproduction’, *Human Reproduction*, 37(6), pp. 1360–1369.

Song, H., Wang, L., Chen, D. & Li, F. (2020) ‘The Function of Pre-mRNA Alternative Splicing in Mammal Spermatogenesis’, *International Journal of Biological Sciences*, 16(1), pp. 38–48.

de Sousa Filho, E.P., Christofolini, D.M., Barbosa, C.P., Glina, S. & Bianco, B. (2018) ‘Y chromosome microdeletions and varicocele as aetiological factors of male infertility: A cross-sectional study’, *Andrologia*, 50(3), p. e12938.

Spaziani, M., Tarantino, C., Tahani, N., Gianfrilli, D., Sbardella, E., Lenzi, A. & Radicioni, A.F. (2021) ‘Hypothalamo-Pituitary axis and puberty’, *Molecular and Cellular Endocrinology*, 520p. 111094.

Steptoe, P.C. & Edwards, R.G. (1978) 'Birth after the reimplantation of a human embryo', *The Lancet*, 312(8085), p. 366.

Sun, C., Skaletsky, H., Birren, B., Devon, K., Tang, Z., Silber, S., Oates, R. & Page, D.C. (1999) 'An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y', *Nature Genetics*, 23(4), pp. 429–432.

Sunderam, S., Kissin, D.M., Crawford, S.B., Folger, S.G., Boulet, S.L., Warner, L. & Barfield, W.D. (2018) 'Assisted Reproductive Technology Surveillance — United States, 2015', *MMWR. Surveillance Summaries*, 67(3), pp. 1–28.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J. & von Mering, C. (2017) 'The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible', *Nucleic Acids Research*, 45(D1), pp. D362–D368.

Tan, E., Chin, C.S.H., Lim, Z.F.S. & Ng, S.K. (2021) 'HEK293 Cell Line as a Platform to Produce Recombinant Proteins and Viral Vectors', *Frontiers in Bioengineering and Biotechnology*, 9.

Tanackovic, G., Ransijn, A., Thibault, P., Abou Elela, S., Klinck, R., Berson, E.L., Chabot, B. & Rivolta, C. (2011) 'PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa', *Human Molecular Genetics*, 20(11), pp. 2116–2130.

The ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74.

Thomas, P. & Smart, T.G. (2005) 'HEK293 cell line: A vehicle for the expression of recombinant proteins', *Journal of Pharmacological and Toxicological Methods*, 51(3), pp. 187–200.

De Tomasi, L., David, P., Humbert, C., Silbermann, F., Arrondel, C., Tores, F., Fouquet, S., Desgrange, A., Niel, O., Bole-Feysot, C., Nitschké, P., Roume, J., Cordier, M.-P., Pietrement, C., Isidor, B., Khau Van Kien, P., Gonzales, M., Saint-Frison, M.-H., Martinovic, J., et al. (2017) 'Mutations in GREB1L Cause Bilateral Kidney Agenesis in Humans and Mice', *The American Journal of Human Genetics*, 101(5), pp. 803–814.

Touré, A., Martinez, G., Kherraf, Z.-E., Cazin, C., Beurois, J., Arnoult, C., Ray, P.F. & Coutton, C. (2021) 'The genetic architecture of morphological abnormalities of the sperm tail', *Human Genetics*, 140(1), pp. 21–42.

Tournaye, H. (2012) 'Male factor infertility and ART', *Asian Journal of Andrology*, 14(1), pp. 103–108.

Tournaye, H., Krausz, C. & Oates, R.D. (2017) 'Novel concepts in the aetiology of male reproductive impairment', *The Lancet Diabetes & Endocrinology*, 5(7), pp. 544–553.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*, 28(5), pp. 511–515.

Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. (2009) 'Massively parallel exon capture and library-free resequencing across 16 genomes', *Nature Methods*, 6(5), pp. 315–316.

Tüttelmann, F., Ruckert, C. & Röpke, A. (2018) 'Disorders of spermatogenesis', *Medizinische Genetik*, 30(1), pp. 12–20.

Tüttelmann, F., Werny, F., Cooper, T.G., Kliesch, S., Simoni, M. & Nieschlag, E. (2011) 'Clinical experience with azoospermia: aetiology and chances for spermatozoa detection upon biopsy', *International Journal of Andrology*, 34(4pt1), pp. 291–298.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A.-K., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., et al. (2015) 'Tissue-based map of the human proteome', *Science*, 347(6220), .

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. & Rozen, S.G. (2012) 'Primer3—new capabilities and interfaces', *Nucleic Acids Research*, 40(15), pp. e115–e115.

De Vas, M.G., Boulet, F., Joshi, S.S., Garstang, M.G., Khan, T.N., Atla, G., Parry, D., Moore, D., Cebola, I., Zhang, S., Cui, W., Lampe, A.K., Lam, W.W., Ferrer, J., Pradeepa, M.M. & Atanur, S.S. (2023) 'Regulatory de novo mutations underlying intellectual disability', *Life Science Alliance*, 6(5), p. e202201843.

Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M. & Ng, P.C. (2016) 'SIFT missense predictions for genomes', *Nature Protocols*, 11(1), pp. 1–9.

Veltman, J.A. & Brunner, H.G. (2012) 'De novo mutations in human genetic disease', *Nature Reviews Genetics*, 13(8), pp. 565–575.

Vissers, L.E.L.M., Gilissen, C. & Veltman, J.A. (2016) 'Genetic studies in intellectual disability and related disorders', *Nature Reviews Genetics*, 17(1), pp. 9–18.

Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., van Bon, B.W.M., Hoischen, A., de Vries, B.B.A., Brunner, H.G. & Veltman, J.A. (2010) 'A de novo paradigm for mental retardation', *Nature Genetics*, 42(12), pp. 1109–1112.

Vockel, M., Riera-Escamilla, A., Tüttelmann, F. & Krausz, C. (2021) 'The X chromosome and male infertility', *Human Genetics*, 140(1), pp. 203–215.

Vogt, P., Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, Köhn F M, Schill W B, Farah S, Ramos C, Hartmann M, Hartschuh W, Meschede D, Behre H M, Castel A, Nieschlag E, Weidner W, Gröne H J, Jung A, et al. (1996) 'Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11', *Human Molecular Genetics*, 5(7), pp. 933–943.

Vorona, E., Zitzmann, M., Gromoll, J., Schüring, A.N. & Nieschlag, E. (2007) 'Clinical, Endocrinological, and Epigenetic Features of the 46,XX Male Syndrome, Compared with 47,XXY Klinefelter Patients', *The Journal of Clinical Endocrinology & Metabolism*, 92(9), pp. 3458–3465.

Wagner, A.O., Turk, A. & Kunej, T. (2023) ‘Towards a Multi-Omics of Male Infertility’, *The World Journal of Men's Health*, 41(2), p. 272.

Walters, J.L.H., Gadella, B.M., Sutherland, J.M., Nixon, B. & Bromfield, E.G. (2020) ‘Male Infertility: Shining a Light on Lipids and Lipid-Modulating Enzymes in the Male Germline’, *Journal of Clinical Medicine*, 9(2), p. 327.

Wang, J.-M., Li, Z.-F., Yang, W.-X. & Tan, F.-Q. (2022) ‘Follicle-stimulating hormone signaling in Sertoli cells: a licence to the early stages of spermatogenesis’, *Reproductive Biology and Endocrinology*, 20(1), p. 97.

Wang, Jun, Wang, Jian, Zhang, H.-R., Shi, H.-J., Ma, D., Zhao, H.-X., Lin, B. & Li, R.-S. (2009) ‘Proteomic analysis of seminal plasma from asthenozoospermia patients reveals proteins that affect oxidative stress responses and semen quality’, *Asian Journal of Andrology*, 11(4), pp. 484–491.

Wang, K., Gao, Y., Wang, C., Liang, M., Liao, Y. & Hu, K. (2022) ‘Role of Oxidative Stress in Varicocele’, *Frontiers in Genetics*, 13.

Wang, M., Liu, X., Chang, G., Chen, Yidong, An, G., Yan, L., Gao, S., Xu, Y., Cui, Y., Dong, J., Chen, Yuhan, Fan, X., Hu, Y., Song, K., Zhu, X., Gao, Y., Yao, Z., Bian, S., Hou, Y., et al. (2018) ‘Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis’, *Cell Stem Cell*, 23(4), pp. 599-614.e4.

Wang, Y.-H., Yan, M., Zhang, X., Liu, X.-Y., Ding, Y.-F., Lai, C.-P., Tong, M.-H. & Li, J.-S. (2021) ‘Rescue of male infertility through correcting a genetic mutation causing meiotic arrest in spermatogonial stem cells’, *Asian Journal of Andrology*, 23(6), p. 590.

Wolf, E., Kastner, B., Deckert, J., Merz, C., Stark, H. & Lührmann, R. (2009) ‘Exon, intron and splice site locations in the spliceosomal B complex’, *The EMBO Journal*, 28(15), pp. 2283–2292.

Wong, C.H., Li, D., Wang, N., Gruber, J., Lo, A.W. & Conti, R.M. (2023) ‘The estimated annual financial impact of gene therapy in the United States’, *Gene Therapy*, 30(10–11), pp. 761–773.

World Health Organization (2019) *International Classification of Diseases, Eleventh Revision (ICD-11)*,

World Health Organization (2010) *WHO laboratory manual for the examination and processing of human semen*. 5th edition.

World Health Organization (2021) *WHO Laboratory Manual for the Examination and Processing of Human Semen. 6th ed*.

Wyrwoll, M.J., Gaasbeek, C.M., Golubickaite, I., Stakaitis, R., Oud, M.S., Nagirnaja, L., Dion, C., Sindi, E.B., Leitch, H.G., Jayasena, C.N., Sironen, A., Dicke, A.-K., Rotte, N., Stallmeyer, B., Kliesch, S., Grangeiro, C.H.P., Araujo, T.F., Lasko, P., D’Hauwers, K., et al. (2022) ‘The piRNA-pathway factor FKBP6 is essential for spermatogenesis but dispensable for control of meiotic LINE-1 expression in humans’, *The American Journal of Human Genetics*, 109(10), pp. 1850–1866.

Wyrwoll, M.J., Temel, Ş.G., Nagirnaja, L., Oud, M.S., Lopes, A.M., van der Heijden, G.W., Heald, J.S., Rotte, N., Wistuba, J., Wöste, M., Ledig, S., Krenz, H., Smits, R.M.,

Carvalho, F., Gonçalves, J., Fietz, D., Türkgenç, B., Ergören, M.C., Çetinkaya, M., et al. (2020) 'Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility', *The American Journal of Human Genetics*, 107(2), pp. 342–351.

Wyrwoll, M.J., van Walree, E.S., Hamer, G., Rotte, N., Motazacker, M.M., Meijers-Heijboer, H., Alders, M., Meißner, A., Kaminsky, E., Wöste, M., Krallmann, C., Kliesch, S., Hunt, T.J., Clark, A.T., Silber, S., Stallmeyer, B., Friedrich, C., van Pelt, A.M.M., Mathijssen, I.B., et al. (2021) 'Bi-allelic variants in DNA mismatch repair proteins MutS Homolog *MSH4* and *MSH5* cause infertility in both sexes', *Human Reproduction*, 37(1), pp. 178–189.

Xavier, M.J., Salas-Huetos, A., Oud, M.S., Aston, K.I. & Veltman, J.A. (2021) 'Disease gene discovery in male infertility: past, present and future', *Human Genetics*, 140(1), pp. 7–19.

Xiao, Y., Bi, M., Guo, H. & Li, M. (2022) 'Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis', *eBioMedicine*, 79p. 104001.

Xie, F., You, Y., Guan, C., Gu, Y., Yao, F. & Xu, J. (2022) 'Association between physical activity and infertility: a comprehensive systematic review and meta-analysis', *Journal of Translational Medicine*, 20(1), p. 237.

Xue, J., Yang, J., Yan, J., Jiang, X., He, L.-Y., Wu, T. & Guo, J. (2012) 'Abnormalities of the testes and semen parameters in clinical varicocele.', *Nan fang yi ke da xue xue bao = Journal of Southern Medical University*, 32(4), pp. 439–42.

Yan, Y., Yang, X., Liu, Y., Shen, Y., Tu, W., Dong, Q., Yang, D., Ma, Y. & Yang, Y. (2017) 'Copy number variation of functional RBMY1 is associated with sperm motility: an azoospermia factor-linked candidate for asthenozoospermia', *Human Reproduction*, 32(7), pp. 1521–1531.

Yang, B., Wang, J., Zhang, W., Pan, H., Li, T., Liu, B., Li, H. & Wang, B. (2017) 'Pathogenic role of *ADGRG2* in CBAVD patients replicated in Chinese population', *Andrology*, 5(5), pp. 954–957.

Yang, C., Georgiou, M., Atkinson, R., Collin, J., Al-Aama, J., Nagaraja-Grellscheid, S., Johnson, C., Ali, R., Armstrong, L., Mozaffari-Jovin, S. & Lako, M. (2021) 'Pre-mRNA Processing Factors and Retinitis Pigmentosa: RNA Splicing and Beyond', *Frontiers in Cell and Developmental Biology*, 9.

Yang, F., Silber, S., Leu, N.A., Oates, R.D., Marszalek, J.D., Skaletsky, H., Brown, L.G., Rozen, S., Page, D.C. & Wang, P.J. (2015) ' *<scp>TEX</scp> 11* is mutated in infertile men with azoospermia and regulates genome-wide recombination rates in mouse', *EMBO Molecular Medicine*, 7(9), pp. 1198–1210.

Yatsenko, A.N., Georgiadis, A.P., Röpke, A., Berman, A.J., Jaffe, T., Olszewska, M., Westernströer, B., Sanfilippo, J., Kurpisz, M., Rajkovic, A., Yatsenko, S.A., Kliesch, S., Schlatt, S. & Tüttelmann, F. (2015) 'X-Linked *TEX11* Mutations, Meiotic Arrest, and Azoospermia in Infertile Men', *New England Journal of Medicine*, 372(22), pp. 2097–2107.

Yatsenko, A.N., Roy, A., Chen, R., Ma, L., Murthy, L.J., Yan, W., Lamb, D.J. & Matzuk, M.M. (2006) 'Non-invasive genetic diagnosis of male infertility using spermatozoal RNA: KLHL10mutations in oligozoospermic patients impair homodimerization', *Human Molecular Genetics*, 15(23), pp. 3411–3419.

Yeo, G., Holste, D., Kreiman, G. & Burge, C.B. (2004) 'Variation in alternative splicing across human tissues', *Genome Biology*, 5(10), p. R74.

Yıldırım, Y., Ouriachi, T., Woehlbier, U., Ouahioune, W., Balkan, M., Malik, S. & Tolun, A. (2018) 'Linked homozygous BMPR1B and PDHA2 variants in a consanguineous family with complex digit malformation and male infertility', *European Journal of Human Genetics*, 26(6), pp. 876–885.

Young, J., Xu, C., Papadakis, G.E., Acierno, J.S., Maione, L., Hietamäki, J., Raivio, T. & Pitteloud, N. (2019) 'Clinical Management of Congenital Hypogonadotropic Hypogonadism', *Endocrine Reviews*, 40(2), pp. 669–710.

Yu, J., Chen, Z., Ni, Y. & Li, Z. (2012) 'CFTR mutations in men with congenital bilateral absence of the vas deferens (CBAVD): a systemic review and meta-analysis', *Human Reproduction*, 27(1), pp. 25–35.

Yu, J., Ji, G., Shi, W., Zhao, R., Shen, W., Zheng, J., Li, H. & Jiang, F. (2020) 'RBM5 Acts as Tumor Suppressor in Medulloblastoma through Regulating Wnt/β-Catenin Signaling', *European Neurology*, 83(3), pp. 242–250.

Zeng, T., Guan, Y., Li, Y., Wu, Q., Tang, X., Zeng, X., Ling, H. & Zou, J. (2021) 'The DNA replication regulator MCM6: An emerging cancer biomarker and target', *Clinica Chimica Acta*, 517pp. 92–98.

Zhang, Z., Zariwala, M.A., Mahadevan, M.M., Caballero-Campo, P., Shen, X., Escudier, E., Duriez, B., Bridoux, A.-M., Leigh, M., Gerton, G.L., Kennedy, M., Amselem, S., Knowles, M.R. & Strauss, J.F. (2007) 'A Heterozygous Mutation Disrupting the SPAG16 Gene Results in Biochemical Instability of Central Apparatus Components of the Human Sperm Axoneme1', *Biology of Reproduction*, 77(5), pp. 864–871.

Zhu, F., Liu, C., Wang, F., Yang, X., Zhang, J., Wu, H., Zhang, Zhiguo, He, X., Zhang, Zhou, Zhou, P., Wei, Z., Shang, Y., Wang, L., Zhang, R., Ouyang, Y.-C., Sun, Q.-Y., Cao, Y. & Li, W. (2018) 'Mutations in PMFBP1 Cause Acephalic Spermatozoa Syndrome', *The American Journal of Human Genetics*, 103(2), pp. 188–199.

Zirkin, B.R. & Papadopoulos, V. (2018) 'Leydig cells: formation, function, and regulation†', *Biology of Reproduction*, 99(1), pp. 101–111.