The Role of Copy Number Variants in Severe Idiopathic Male Infertility

Francesco Kumara Mastrorosa

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

Biosciences Institute
Faculty of Medical Sciences
Newcastle University

August 2021

Abstract

Very few genetic variants are currently known to cause severe male infertility phenotypes in a dominant fashion. This is explained by an absence of cohorts of patient-parent trios, which impedes the study of this genetic inheritance model. It has been demonstrated that *de novo* copy number variants (CNVs) on chromosome Y cause severe male infertility, but up to now, such variants have not been identified on the autosomes. In this thesis, I used whole-exome sequencing (WES) in a large cohort of male infertility patient-parent trios to study the possible role of *de novo* and maternally inherited CNVs as causes of quantitative sperm defects. I also used the same method to identify likely pathogenic CNVs in two patients-only cohorts affected by quantitative as well as qualitative sperm defects. We identified several possibly causative *de novo* and maternally inherited CNVs, pointing us to novel candidate genes for male infertility. Moreover, my findings contributed to the identification of the first gene on chromosome X associated with a male infertility phenotype characterised by multiple morphological abnormalities of the sperm flagella. This study reveals unique insight into the genetic aetiology of severe male infertility. Also, it illustrates the value of WES as a tool for CNV detection and supports the increasing use of this genomic technique in the field of male infertility genetics.

Acknowledgements

I would like to thank my supervisors Prof Joris Veltman and Dr Mauro Santibanez-Koref for their guidance and inspiration. I thank Joris for accepting me as a PhD student when my knowledge of genomics was very basic. Thanks to you, I have learned a lot since then, and I have had many opportunities to meet other colleagues working in the field. I thank Mauro for spending a great amount of time with me during my first year as a PhD student, guiding me and advising me and for being always willing to help. You both helped me develop my critical thinking and facilitated my growth as a scientist. I hope that there will be other opportunities to work together again in the future.

I thank all the present and past members of the male infertility research group. It was a pleasure to work with all of you, and I enjoyed the time together in the office and outside. In particular, I would like to thank Dr Miguel Xavier for his day-to-day assistance during my analyses and our exchange of ideas. I believe that his support was very important to maintain my PhD work on the right path. I thank Dr Bilal Alobaidi for his contribution to the CNV validations included in this thesis. I also thank him for his kindness and for being always willing to help. I would like to thank Dr Harsh Sheth for suggesting to apply for this PhD position and his advice before starting my PhD and during my first year. I also thank the Barbour Foundation for funding this PhD.

I would like to thank all my colleagues at Newcastle University. I thank the member of my panel Dr Michael Jackson and Dr Gavin Hudson for their feedback and suggestions. I thank Dr Aneta Mikulasova for providing the CNV pipeline and technical support. I thank the Genomic Core Facility for sequencing the samples. I thank the Bioinformatic Support Unit for the data processing and, in particular, Dr Simon Cockell for answering my numerous questions on bioinformatics during my first year. I also thank Jackie Greenwood for the recruitment of the patients in Newcastle. I would like to thank Trish Dunham for helping me with office work and bureaucratic rules. I thank Dr Emily Dookun for suggesting to apply for the PhD student rep role and helping with the organisation of the PhD away day. I thank all my colleagues at Radboud University. In particular, Dr Godfried van der Heijden and Dr Manon Oud. It was a pleasure to work with you, and I thank you for showing me Nijmegen and making me feel at home. I thank Dr Christian Gilissen and Dr Alexander Hoischen for our discussions on genomics. I also thank the patient recruiting team at Radboudumc. I thank my colleagues at Monash University for our collaborations. It was nice working with you and exchange ideas during our meetings. Additionally, I would like to thank Dr Feng Zhang from Fudan University and all the colleagues that worked on the CFAP47 project. I also thank all the patients and families that participated to our studies. I thank any person not individually named here that supported me during my PhD in any way.

Lastly, I would like to express my gratitude to my girlfriend, friends and family for their unconditional love and support.

Table of contents

List of figures

List of tables

List of abbreviations

AIS	Androgen insensitivity syndrome
ART	Assisted reproductive technology
ASD	Autism spectrum disorder
AZF	Azoospermia factor
BAM	Binary alignment map
CAIS	Complete androgen insensitivity syndrome
CBAVD	Congenital bilateral absence of the vas deferens
CCDS	Consensus coding sequence
CNV	Copy number variant
CV	Chromosome View
DGV	Database of Genomic Variations
DNA	Deoxyribonucleic acid
EGA	European Genome-phenome Archive
GP	Genome Profile
GWAS	Genome-wide association study
G4BP	GATK4-based pipeline
ICSI	Intra-cytoplasmic sperm injection
ID	Intellectual disabilities
IE	Illumina-enriched
IMIGC	International Male Infertility Genomics Consortium
INDEL	Deletions and duplications
IVF	In-vitro fertilisation
MAF	Minor allele frequency
MAIS	Mild androgen insensitivity syndrome

MI	Maternally inherited
MMAF	Multiple Morphological Abnormalities of the Flagella
NAHR	Nonallelic homologous recombination
NGS	Next-generation sequencing
PAIS	Partial androgen insensitivity syndrome
PCR	Polymerase Chain Reaction
PI	Paternally inherited
q-PCR	Quantitative Polymerase Chain Reaction
SD	Standard deviation
SNV	Single nucleotide variant
TE	Twist-enriched
UTR	Untranslated region
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WHO	World Health Organization

Chapter 1. Introduction

1.1 Male Infertility

1.1.1 Definition and characteristics

Infertility is defined by the World Health Organization (WHO) as the inability of a couple to achieve pregnancy after a year or more of unprotected sexual intercourse (World Health Organization, 2019). Around 10-15% of couples in developed and developing countries suffer from this condition, and male factors are thought to account for 50% of infertility cases (Neto, Bach, Bobby Baback Najari, *et al.*, 2016; Tüttelmann, Ruckert and Röpke, 2018). The WHO established the average semen parameters based on fertile men (men whose partners had a time to pregnancy ≤ 12 months) from the general population. Normozoospemic men generally have an average semen volume of 1.5 ml; 39 million spermatozoa per ejaculate; sperm concentration of 15 million sperm/ml; a minimum of 58% of vital spermatozoa and 40% motile; at least 32% showing progressive motility and 4% of the spermatozoa morphologically normal (World Health Organization, 2010). Male infertility is divided into four aetiological categories. Quantitative defects of spermatogenesis represent the largest category, for which primary testicular failure (also known as primary hypogonadism) is the most frequent cause (Krausz, 2011; Krausz and Riera-Escamilla, 2018). The second common category is ductal obstruction or dysfunction, followed by alteration of the hypothalamic-pituitary axis (also referred to as secondary testicular failure or secondary hypogonadism) (Krausz and Riera-Escamilla, 2018). Finally, the smallest category consists of errors of the sperm productions that lead to qualitative defects of the sperm (Tournaye, Krausz and Oates, 2017; Krausz and Riera-Escamilla, 2018). Azoospermia, the condition of complete absence of sperm in the ejaculate, is the most severe form of quantitative sperm defects and is found in 10-15% of infertile men (F. Tüttelmann *et al.*, 2011). This is the most impairing form of infertility because treatments (i.e., artificial reproductive technologies) are often unfeasible due to the absence of spermatozoa both in the ejaculate and in testicular tissue. Azoospermia can be caused by the physical blockage of the male excurrent ductal system (referred to as obstructive azoospermia) or by sperm production defects (indicated as non-obstructive azoospermia). Other infertility phenotypes are associated with a reduced number of spermatozoa (oligozoospermia), reduced sperm motility (asthenozoospermia), morphological defects (teratozoospermia) or a combination of these features (Esteves *et al.*, 2018) (Figure 1.1). The last two conditions also have extremely rare subcategories, such as globozoospermia, which presents with round-headed sperm and Primary Cilia Dyskinesia, a disease caused by the impairment of different parts of the primary ciliary apparatus (Krausz, Escamilla and Chianese, 2015; Esteves *et al.*, 2018). Finally, a subtype of both is the Multiple Morphological Abnormalities of the sperm Flagella (MMAF) syndrome.

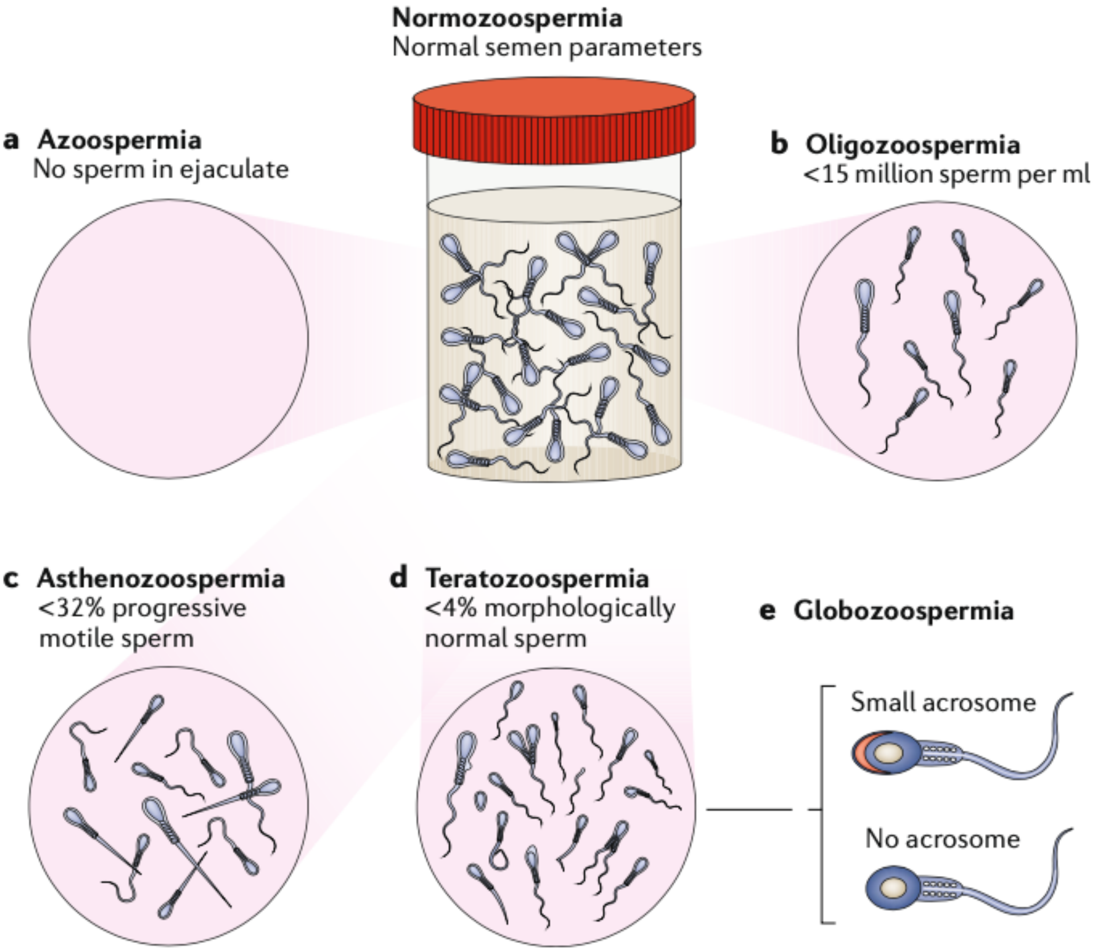


Figure 1.1. Infertility phenotypes. **a.** Azoospermia **b.** Oligozoospermia **c.** Asthenozoospermia **d.** Teratozoospermia **e.** Globozoospermia. Reprinted with permission from Springer Nature (Esteves *et al.*, 2018).

1.1.2 The biology of spermatogenesis

The heterogeneity of male infertility phenotypes is not surprising considering the complexity of spermatogenesis, the process that produces sperm from primordial germ cells. It includes the proliferation of spermatogonial stem cells, differentiation into spermatocytes as well as generation and maturation of haploid spermatids (Neto, Bach, Bobby B. Najari, *et al.*, 2016; Chen *et al.*, 2017) (Figure 1.2). A complete round of spermatogenesis requires 64 days on average and continues throughout the entire reproductive life (Misell *et al.*, 2006; Griswold, 2016). These steps are regulated by the hypothalamic-pituitary-testicular axis, which controls, through a series of concatenated events, the level of gonadotropin-releasing hormone, follicle stimulating hormone, luteinising hormone, testosterone and estradiol-17ß (Sharma and Agarwal, 2011; Chen *et al.*, 2017). The entire process of spermatogenesis is supported by the Sertoli cells. These cells contribute to testis formation under the regulation of Y chromosome-specific genes (Griswold, 1998). Also, once testicles are correctly formed, they provide structural, nutritional and paracrine support to the germ cells and (particularly during the advanced steps of spermatogenesis) to the haploid spermatids, which are metabolically limited (Mruk and Cheng, 2004). At puberty, in the basal compartment, spermatogonial stem cells start their mitotic process to renew the cell pool and generate spermatogonia that differentiate and eventually become spermatozoa (Sharma and Agarwal, 2011). The spermatogonia undergo mitosis to produce primary spermatocytes. These cells perform meiosis I to form secondary spermatocytes, which in the adluminal compartment progress into meiosis II to produce haploid spermatids. Once the latter are produced, the process called spermiogenesis leads to their maturation (Sharma and Agarwal, 2011). During this step, the spermatids acquire acrosomes and flagella. The mature spermatids then detach themselves from the Sertoli cells and migrate into the lumen of the seminiferous tubules, becoming free spermatozoa. (Sharma and Agarwal, 2011). Interestingly, spermatids originating from the same spermatogonia, until this point, remain connected by intercellular bridges that allow synchronised maturation and biochemical exchanges (Dym and Fawcett, 1971; Braun *et al.*, 1989; Sharma and Agarwal, 2011). As a result of this process, spermatozoa are morphologically mature. However, they will be able to fertilise only after the “capacitation” step, a biochemical and physiological modifications taking place in the epididymis and the female genital tract (Jin and Yang, 2017). Remarkably, Sertoli cells play a crucial role in most of these processes, and spermatogenesis would not happen without them (Sharma and Agarwal, 2011). Sertoli cells are related to a male infertility syndrome called the Sertoli-Cell-Only syndrome, characterised by the complete absence of germ cells in the testis and consequent infertility. The complexity of spermatogenesis is demonstrated by this multitude of steps, which need to occur regularly and without complications. This is further illustrated by studies using mouse genetic models that revealed more than 400 genes involved in spermatogenesis (Jamsai and O'Bryan, 2011; Krausz, Escamilla and Chianese, 2015) and by proteomics, epigenomics and genomics investigations that estimated that more than 2300 genes are associated with spermatogenesis and testis functions (Schultz, Hamra and Garbers, 2003).

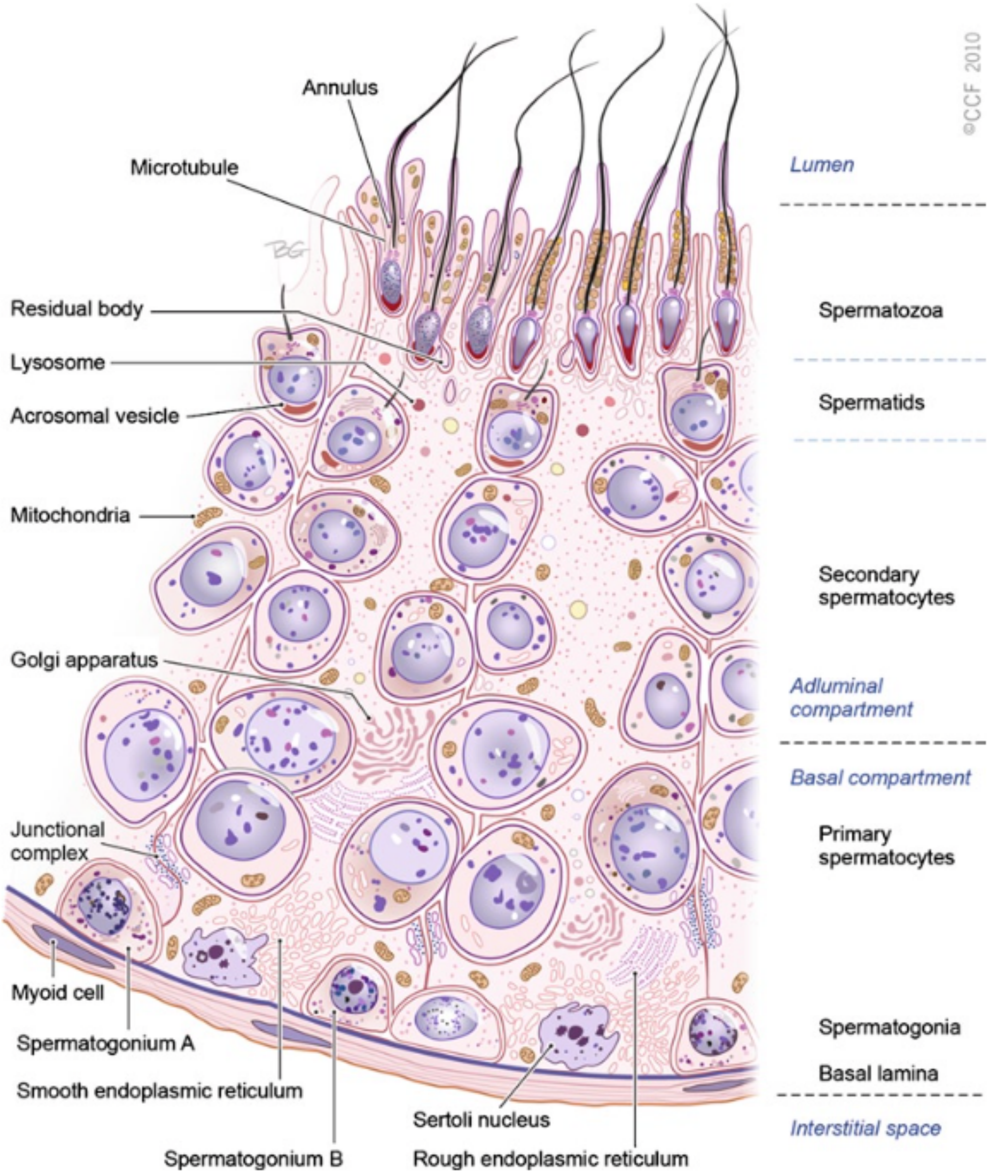


Figure 1.2. Section of the seminiferous tubules' epithelium. Every compartment is defined by the cell types present. From the basal compartment to the lumen, the spermatogonia undergo different phases to become full morphologically formed spermatozoa. Reprinted with permission from Springer Nature (Sharma and Agarwal, 2011).

1.1.3 Risk factors and non-genetic causes of male infertility

Fertility is not stable throughout a man's life and several risk factors can contribute to its decrease. Genetic and non-genetic causes can impair spermatogenesis, leading to more or less severe phenotypes. A 2016 review described the main risk factors and non-genetic causes extensively (Mahat *et al.*, 2016). The foremost risk factors are age (Pasqualotto *et al.*, 2004) and different living habits, including smoking (Nadeem, Fahim and Bugti, 2012), alcohol consumption (Emanuele and Emanuele, 1998), obesity (Kort *et al.*, 2006), exposure to harmful physical and chemical agents (Cherry *et al.*, 2001; Ten *et al.*, 2008) and long-term tiresome and resistance exercise (Tremblay, Copeland and Van Helder, 2004; Safarinejad, Azma and Kolahi, 2009). Stress, both physical and psychological, although still debated, has been suggested to be an additional risk factor (Mahat *et al.*, 2016) as well as habits that increase scrotal temperature (Mieusset *et al.*, 1987; Mieusset and B'ujan, 1994) or reactive oxygen species (Olayemi, 2010) and the consumption of certain therapeutic drugs (Mieusset and B'ujan, 1994). Risk factors aside, fertility can also be impaired by conditions such as untreated varicocele (Cozzolino and Lipshultz, 2001; Jarow, 2001), a condition characterised by the enlargement of scrotal veins; complications of mumps orchitis (Masarani, Wazait and Dinneen, 2006), a condition in which patients have pain and swelling of the testes after mumps infection; endocrinal disorders, such as hormone deficiency (Lalitha *et al.*, 2013); infections of the reproductive tract (Marconi *et al.*, 2009); ejaculatory disorders and immunological factors (Brugh and Lipshultz, 2004).

1.1.4 Treatments

Assisted reproductive technologies (ART), such as in-vitro fertilisation (IVF) and intra-cytoplasmic sperm injection (ICSI), are modern treatments for male infertility. IVF was performed successfully for the first time in 1978 (Steptoe and Edwards, 1978) and consists of oocyte fertilization through incubation with sperm in a Petri dish. ICSI is an evolution of the IVF technique and was first introduced in 1992 (Palermo *et al.*, 1992). It consists of the precise injection of a single spermatozoon into the oocyte cytoplasm with a glass micropipette. The sperm for these procedures is retrieved from the ejaculate when possible, but in severe infertility cases, epididymal sperm aspiration or testicular sperm extractions are necessary (Esteves *et al.*, 2018). In azoospermia patients, sperm extraction can be performed if residual spermatogenesis occurs in the testis (Esteves, 2015; Esteves *et al.*, 2018). Both IVF and ICSI are treatments used in more than 60 countries as reported by the International

Committee for Monitoring Assisted Reproductive Technologies (Dyer *et al.*, 2016; Adamson *et al.*, 2018), and at least 200,000 babies are conceived through ART each year worldwide (Qin *et al.*, 2015). Several questions have been raised on the possible consequences of ART on the health of the offspring, particularly for babies conceived with ICSI procedure, which bypass the natural selection of the sperm. Some published researches have reported a slightly increased risk of congenital malformation in ART-born babies (Qin *et al.*, 2015; Hoorsan *et al.*, 2017). A small increased risk for autistic disorders and intellectual disabilities (ID) has been found in ICSI-born offspring compared to IVF-born babies (Sandin *et al.*, 2013). Despite these findings, there is general uncertainty on whether ART substantially increases the risk of congenital malformations or neurodevelopmental delay as well as on the risk difference between IVF and ICSI procedures (Massaro *et al.*, 2015; Catford *et al.*, 2017; Esteves *et al.*, 2018). ICSI introduction provided a treatment option for severe male-factor infertility (Sandin *et al.*, 2013), but its influence on the reproductive health of the offspring is currently understudied. This is mainly because the first ICSI-born children are only now reaching the adult stage of their life, and not many research groups have been able to assess their health, including their reproductive fitness (Belva, Bonduelle and Tournaye, 2019). The first and so far only study investigating the semen quality of young adults conceived through ICSI was published in 2016 (Belva *et al.*, 2016). It showed decreased semen quality and quantity in a group of 54 young adults born with ICSI due to male infertility of their fathers. However, the samples size was limited, and there was a weak negative correlation between the total sperm count in the fathers and their sons (Belva *et al.*, 2016). These data highlight the need of long-term follow-up on babies born after ICSI and large studies investigating the consequence of the technique on the reproductive health of the offspring. This is a main interest of the field of male infertility genetics since IVF, and especially ICSI, could allow the transmission of pathogenic genetic variants from infertile fathers to their children. A full understanding of both the genetics of male infertility and the impact of IVF and ICSI on the reproductive health of the offspring is crucial for proper counselling and treatment.

1.2 The Genetics of Male Infertility

Genetic causes can explain male infertility cases, and the most severe phenotypes correlate with the presence of germline genetic abnormalities (Krausz and Riera-Escamilla, 2018). There are very few published studies investigating the causes of infertility on a large, unbiased selection of infertile men. With the current knowledge only 4 - 9.2% of male infertility cases can be explained by genetic causes, with 36 to 72% of the total left unexplained (Olesen *et al.*, 2017; Punab *et al.*, 2017; Tüttelmann, Ruckert and Röpke, 2018).

1.2.1 Chromosomal abnormalities

Structural and numerical chromosomal aberrations play a major role in male infertility. These variants are thought to interfere with meiosis during spermatogenesis, leading to meiotic arrest (Sun *et al.*, 2007).

Klinefelter syndrome (karyotype 47, XXY) is the most common chromosomal abnormality causing non-obstructive azoospermia (Krausz and Riera-Escamilla, 2018) and in large studies explained between 0.9 and 3.5% of the cases (Olesen *et al.*, 2017; Punab *et al.*, 2017; Tüttelmann, Ruckert and Röpke, 2018). It was first reported as a syndrome presenting with gynecomastia, aspermatogenesis and other distinctive traits by the scientist who gave his name to the condition in 1942 (Klinefelter, Reifenstein and Albright, 1942). Years later, Ferguson-Smith *et al.* and Jacobs and Strong described the 47, XXY as the karyotype typical of the syndrome (Ferguson-Smith *et al.*, 1957; Jacobs and Strong, 1959). Another chromosomal abnormality associated with male infertility is the 46, XX syndrome, also known as "de la Chapelle" syndrome from the name of the scientist who reported the first case (de la Chapelle *et al.*, 1964). Patients with this syndrome are azoospermic, and 90% of them also carry an *SRY* (sex-determining region Y) translocation on chromosome X (Capel, 1998; Zenteno-Ruiz, Kofman-Alfaro and Méndez, 2001). *SRY* is considered to be the initiator of the male developmental pathway. It leads the differentiation of the gonadal bipotential cells into testes (Zenteno-Ruiz, Kofman-Alfaro and Méndez, 2001). The remaining 10% of de la Chapelle syndrome cases are thought to be caused by either genetic variants in unknown gene/genes involved in sex determination or Y chromosome mosaicism (Zenteno-Ruiz, Kofman-Alfaro and Méndez, 2001). Other structural variants, including Robertsonian translocations and large inversions, are more frequent in severe oligozoospermia cases than normozoospermic men (Vincent *et al.*, 2002). Due to the major involvement of chromosomal abnormalities, karyotype analysis is one of the most common diagnostic tests advised for azoospermia and severe oligozoospermia cases (Jungwirth A.D.T. *et al.*, 2018).

A second diagnostic test advised for male infertility aims to detect deletions on the Y chromosome. These chromosomal aberrations are often not detectable by karyotyping due to their small size, and explain between 0.3 and 5.4% of all male infertility cases (Olesen *et al.*, 2017; Punab *et al.*, 2017; Tüttelmann, Ruckert and Röpke, 2018). Some regions on the long arm of the Y chromosome were discovered to be essential for spermatogenesis ~20 years ago, and since then, they have been called azoospermia factor (AZF) regions (Vogt *et al.*, 1996; Skaletsky *et al.*, 2003). There are three of these regions, named AZFa, AZFb and AZFc (the last two partially overlap). They contain several genes highly expressed in the testis and likely to be involved in spermatogenesis (Krausz and Riera-Escamilla, 2018). The phenotypes resulting from the deletions of these regions vary according to the region involved and range from azoospermia to oligozoospermia. Usually, the most severe condition (azoospermia with absence of sperm in the testicular tissue) is found in patients carrying large AZFa deletions, while a sub-normal quantity of sperm is found in subjects with AZFc deletions (McLachlan and O'Bryan, 2010; Krausz and Riera-Escamilla, 2018) (Figure 1.3). The AZF flanking regions contain repeated homologous sequences that increase the risk of nonallelic homologous recombination (NAHR) and consequently the risk of structural variants occurring on this chromosome (Krausz and Riera-Escamilla, 2018). Klinefelter syndrome and Y chromosome microdeletions are the two most common genetic causes of severe male infertility and together can be found in up to ~ 6% of the cases (Olesen *et al.*, 2017; Punab *et al.*, 2017). Relevant for this thesis, both these abnormalities occur mostly *de novo* in the germline of the infertile men (Krausz and Riera-Escamilla, 2018). Something that is not unexpected as both parents were fertile and therefore unlikely to carry these severe chromosome abnormalities that impact fitness.

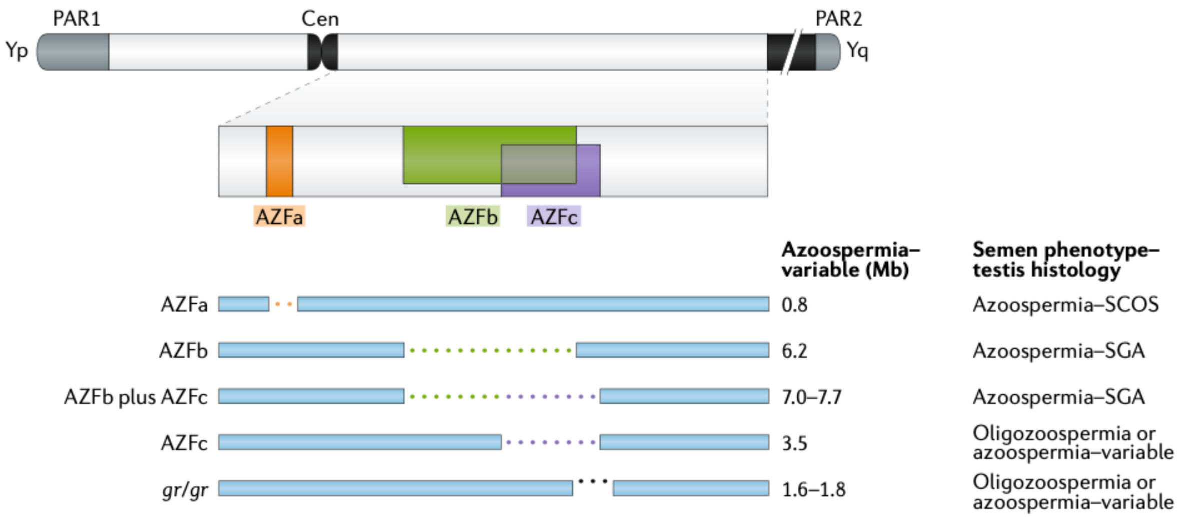

PAR1
Cen
PAR2
Yp
Yq
AZFa
AZFb
AZFc
Azoospermia–variable (Mb)
Semen phenotype–testis histology
AZFa
0.8
Azoospermia–SCOS
AZFb
6.2
Azoospermia–SGA
AZFb plus AZFc
7.0–7.7
Azoospermia–SGA
AZFc
3.5
Oligozoospermia or azoospermia–variable
gr/gr
1.6–1.8
Oligozoospermia or azoospermia–variable

Figure 1.3. Overview of the various deletions of the AZF regions on chromosome Y. The male infertility phenotype is usually more severe with AZFa deletions, while in patients with AZFc deletions often sperm is present in the ejaculate (oligozoospermia) or in the testicular tissue. CEN = centromere; PAR = pseudoautosomal regions; SCOS = Sertoli-cell-only syndrome; SGA = spermatogenic arrest. Reprinted with permission from Springer Nature (Krausz and Riera-Escamilla, 2018).

1.2.2 Dominant and recessive male infertility genes

Currently, very few genes have been associated with male infertility and even less are routinely tested in the diagnostic follow-up of male infertility cases. Cystic fibrosis transmembrane regulator (*CFTR*) gene is known to cause cystic fibrosis when both the alleles are impaired. Most (>95%) male patients with cystic fibrosis are infertile due to congenital bilateral absence of the vas deferens (CBAVD) (obstructive azoospermia), the ducts that allow the spermatozoa to pass from the testes to the urethra (Sokol, 2001; Popli and Stewart, 2007). There are also patients with *CFTR* mutations that are infertile but do not show cystic fibrosis symptoms (Sokol, 2001). In these men, the condition is thought to be caused by milder mutations and that isolated CBAVD is a milder form of cystic fibrosis, affecting only the genital tract (Patrizio *et al.*, 1993; de Souza *et al.*, 2018). A similar scenario was found investigating the contribution of the Androgen Receptor (*AR*) mutations to male infertility. Pathogenic variants of the *AR* gene cause Androgen Insensitivity Syndrome (AIS), which can have different forms, from complete (CAIS) to mild (MAIS) (O'Hara and Smith, 2017). If CAIS cases have an external female phenotype and impaired masculinization, patients with milder forms, MAIS and partial AIS (PAIS), instead tend to have an external male phenotype, advanced masculinization and testicular maturation, but they suffer from male infertility (Ferlin *et al.*, 2006; O'Hara and

Smith, 2015). An expansion of the CAG trinucleotide repeat in the first exon of the *AR* gene (8-37 repeat units in unaffected individuals), has been also associated with male infertility (Mobasseri *et al.*, 2018). However, since the results of the investigations have not always been concordant and often ethnicity-specific, this expansion is usually considered just a risk factor, especially for the Caucasian population (Mobasseri *et al.*, 2018). The *AR* and *CFTR* genes are now classified as male infertility genes (Jungwirth A.D.T. *et al.*, 2018), and testing is being introduced in the common diagnostic practice. Despite this, diagnostic procedures in male infertility genetics have remained almost unchanged since the AZF regions were discovered ~ 20 years ago, and are mainly limited to karyotyping, Y chromosome deletion detection and *CFTR* and *AR* mutation screening (Tüttelmann, Ruckert and Röpke, 2018; Oud *et al.*, 2019).

Research output in male infertility genetics has grown substantially in the last few years. A recent systematic review published at the beginning of 2019 evaluated research papers reporting candidate male infertility genes (Oud *et al.*, 2019). The authors showed that the level of research on male infertility genetics grew steadily since the 90s. Starting from 2012, next-generation sequencing (NGS) has increasingly been used in laboratories worldwide. Despite increased research and many candidate genes reported, the improvements in the field have not been substantial, as illustrated by the already discussed unchanged diagnostic practice. Oud *et al.* concluded that at the beginning of 2019, there were only 16 high-confidence autosomal recessive (*AURKC*, *CFAP43*, *CFAP44*, *CFAP69*, *CFTR*, *DNAH1*, *DPY19L2*, *FANCA*, *FANCM*, *PLCZ1*, *PMFBP1*, *SPATA16*, *SUN5*, *TEX15*, *WDR66*, *XRCC2*), 4 autosomal dominant (*DMRT1*, *HSF2*, *KLHL10*, *SYCP3*) and 4 X-linked (*ADGRG2*, *AR*, *NR0B1*, *TEX11*) genes known to cause isolated male infertility when mutated (Oud *et al.*, 2019). Most of the genes acting in a recessive manner were found in patients from consanguineous families with specific forms of male infertility, particularly in qualitative sperm defects. Variants in these genes are unlikely to explain a major fraction of the more common quantitative forms of male infertility in the outbred population. For the few autosomal dominant and X-linked genes, it is unknown whether the pathological variants were inherited or represented *de novo* mutations (Oud *et al.*, 2019).

1.2.3 Different approaches to study male infertility genetics

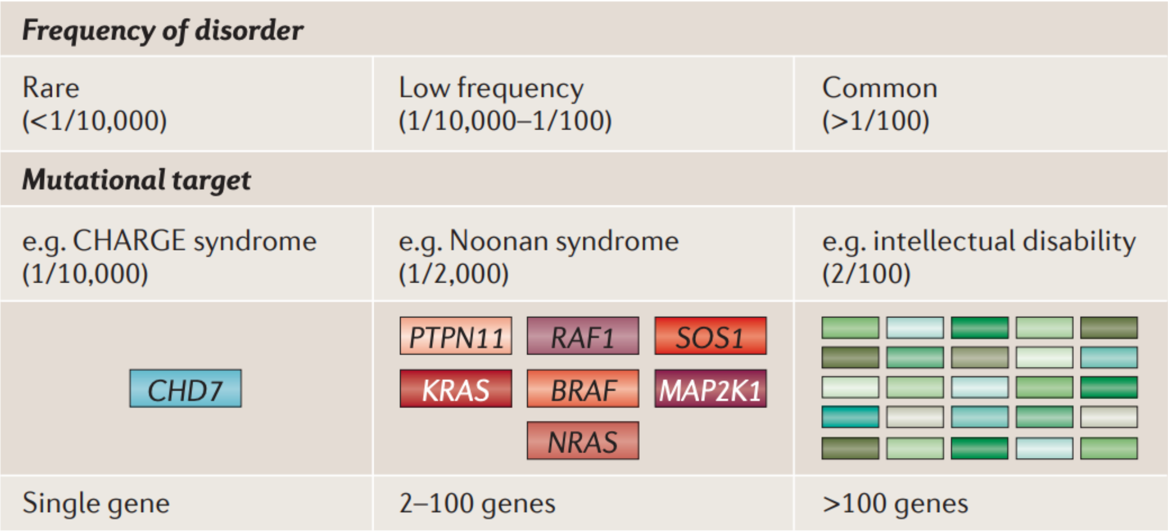
The percentage of unexplained isolated male infertility cases can reach >70% (Tüttelmann, Ruckert and Röpke, 2018). In the past, different approaches were attempted to identify novel genetic causes of male infertility.

Several Genome-Wide Association Studies (GWAS) were performed for the disease in the last ~10 years and revealed significant association with a small group of SNPs (Aston and Carrell, 2009; Aston *et al.*, 2010; Dalgaard *et al.*, 2012; Hu *et al.*, 2012; Kosova *et al.*, 2012; Zhao *et al.*, 2012). A 2014 review showed that the results from these GWAS investigations were discordant, and associated SNPs conferred only a modest risk (Aston, 2014). Thus, the researchers concluded that common SNPs do not independently contribute to a clinically significant risk of severe male infertility (Aston, 2014).

Numerous mutant mouse models with a reproductive phenotype were produced in the last 20 years. They revealed hundreds of candidate male infertility genes and helped associating these genes to specific male infertility phenotypes (Matzuk and Lamb, 2008). The resequencing of these candidate male infertility genes identified in animal models is another approach used in the field for disease gene discovery. This method led to the identification of human male infertility genes, such as *SYCP3* (Miyamoto *et al.*, 2003) and *TEX11* (Yang *et al.*, 2015; Yatsenko *et al.*, 2015; Xavier *et al.*, 2020). However, it is not an unbiased method. It depends on the knowledge of disease genes in other species, which often have a similar but not identical reproductive system compared to humans (Jamsai and O'Bryan, 2011).

Another approach involves cases-controls studies, which have been mainly used in male infertility to investigate structural variants and mutations in specific candidate genes (Tüttelmann *et al.*, 2011; Krausz *et al.*, 2012; Lopes *et al.*, 2013; Yatsenko *et al.*, 2015). This method is helpful in the discovery of new disease genes, but requires very large cohorts of patients and controls and detailed clinical and phenotypical information (Jamsai and O'Bryan, 2011; Xavier *et al.*, 2020).

As mentioned in paragraph 1.1.2, there are ~2300 genes predicted to be involved in testis function and spermatogenesis (Schultz, Hamra and Garbers, 2003). Therefore, it would require very large-sized unbiased genomic studies to obtain a complete overview of all genes involved in the aetiology of male infertility. Equally important, most research so far has been focused on recessive and X-linked forms of male infertility, with little attention to dominant causes. This is reflected in the very few autosomal dominant genes confidently linked to the disease (Oud *et al.*, 2019) and the fact that the parental origin of their deleterious mutations has not been clarified. Unless transmitted through the maternal side, dominant mutations cannot cause male infertility since the disorder would prevent these variants from being passed on from generation to generation. However, research into other disorders with a negative effect on reproductive fitness showed that *de novo* germline mutations and structural variants affecting the coding region frequently result in dominant diseases. As an example, whole-exome (WES) and whole-genome sequencing (WGS) studies demonstrated that damaging *de novo* germline mutations in the coding region of the genome explain the majority of all severe ID (IQ < 50) (Vissers, Gilissen and Veltman, 2016). An important aspect of these diseases is their mutational target, i.e., the number of genes that can cause the condition when *de novo* mutated. The size of the mutational target is positively correlated with the disease frequency in the population. The rarer is the condition, the smaller the mutational target, contrarily, relatively frequent conditions, such as ID, have high genetic heterogeneity, and isolated mutations in many genes could cause them (Veltman and Brunner, 2012) (Table 1.1).



Frequency of disorder		
Rare (<1/10,000)	Low frequency (1/10,000–1/100)	Common (>1/100)
Mutational target		
e.g. CHARGE syndrome (1/10,000)	e.g. Noonan syndrome (1/2,000)	e.g. intellectual disability (2/100)
CHD7	PTPN11 RAF1 SOS1 KRAS BRAF MAP2K1 NRAS	🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩
Single gene	2–100 genes	>100 genes

Table 1.1. The relationship between the mutational target size and the frequency of genetic disorders caused by *de novo* mutations. Disorders caused by mutations in a single gene are rare because of the low probability of a mutational event in that specific gene. In contrast, disorders caused by isolated mutations in many different genes are more common in the general population. Reprinted with permission from Springer Nature (Veltman and Brunner, 2012).

To demonstrate this concept, in 2016, there were more than 700 genes known to cause intellectual disability and related disorders when mutated (Vissers, Gilissen and Veltman, 2016). A similar scenario could be hypothesised for male infertility. It affects ~7% of all men (Krausz and Riera-Escamilla, 2018), and numerous genes are involved in spermatogenesis, making a high genetic heterogeneity very likely. A *de novo* paradigm would also explain the persistence of the disorder in the outbred population, despite its reproductive lethality. This is consistent with the observation that the two most common genetic causes of non-obstructive azoospermia, Klinefelter syndrome and Y chromosome deletions, represent *de novo* variants in affected individuals (Thomas and Hassold, 2003; Krausz and Riera-Escamilla, 2018). Nevertheless, no large-scale studies are investigating *de novo* single nucleotide variants (SNVs) and copy-number variants (CNVs) in male infertility patients.

In order to investigate possible dominant forms of male infertility, it is crucial to conduct studies involving a large number of patients and parents to identify deleterious *de novo* and maternally inherited variants (that affects male fertility exclusively). This task is substantially more difficult in male infertility, where patients have already reached the adult stage of their life and often access to parental DNA is not possible. Sometimes parents are unreachable (deceased or not willing to participate). Other times patients do not consent to participate in the study or share their fertility status with their parents since male infertility can still carry a social stigma. In fact, even in modern days, based on obsolete stereotypes, it is relatively common to attribute the failure of conceiving to the woman and/or to consider men with fertility problems as lacking masculinity and virility (Gannon, Glover and Abel, 2004; Mumtaz, Shahid and Levay, 2013; Ergin *et al.*, 2018). To overcome this problem, international collaborations are critical as they can facilitate the recruitment of large numbers of patients and parents. An example is the International Male Infertility Genomics Consortium (IMIGC) (http://www.imigc.org/), which includes, in addition to our group at Newcastle University, groups working in three different continents on the genetics of male infertility. The design of these studies should also take advantage of the modern NGS techniques that allow an unbiased investigation of the entire coding region of the genome, both for SNVs and CNVs. Discovering novel disease genes acting in a dominant manner and novel forms of transmission would improve not only diagnostic, but also drug discovery, genetic counselling and personalised medicine.

1.3 WES in Mendelian Diseases

NGS started to play a more routine role in infertility research since 2012 (Oud *et al.*, 2019). Amongst its advantages, it provides an unbiased approach and the ability to investigate multiple classes of genetic variations (chromosomal abnormalities, structural variants, SNV and small deletions and duplications, also known as INDELs) with a single test. A 2018 systematic review investigated the cost of both WES and WGS (Schwarze *et al.*, 2018). The authors reported that the cost of WES ranged between $555 (£382) and $5169 (£3592) per test and that of WGS between $1906 (£1312) and $24810 (£17243). WGS allows the investigation of non-coding regions, provides a more consistent sequencing coverage and a more reliable identification of structural variants (Lelieveld *et al.*, 2015). WES, however, is cheaper and has multiple advantages in terms of speed and resources needed for data processing, storage and interpretation (Lelieveld *et al.*, 2016). Its diagnostic yield reached 25% for Mendelian diseases (Yang *et al.*, 2013), a percentage much higher than karyotype analysis (5 to 15%) (Shevell *et al.*, 2003; Shaffer, 2005), microarray analysis (15 to 20%) (Miller *et al.*, 2010) and Sanger sequencing of single genes, which depends on prior knowledge of disease genes. It is also an affordable and effective technique for Mendelian disease gene discovery since most of these disorders are caused by variants disrupting the coding region, and to study diseases with a high genetic heterogeneity (Bamshad *et al.*, 2011; Lelieveld *et al.*, 2015). All these characteristics make WES a convenient first-tier approach to investigate the genetics of male infertility. In this thesis, WES is used to study different cohorts of patient-parent trios and single patients affected by quantitative and qualitative forms of male infertility.

1.4 CNVs and male infertility

CNVs are the focus of this thesis. In the following chapters, *de novo* and inherited CNVs are characterised in large groups of infertile men. At the same time, my colleagues investigated the role of *de novo* and inherited SNVs in the same cohorts using the same WES data. Previous studies have investigated the contribution of CNVs (outside the chromosome Y) to the origin of severe male infertility (Frank Tüttelmann *et al.*, 2011; Lopes *et al.*, 2013; Chianese *et al.*, 2014; Eggers *et al.*, 2015; Luo *et al.*, 2019). The authors looked for CNVs (mainly using microarrays) present more frequently in infertile men than controls or screened specific loci associated with male infertility in patients and unaffected individuals. These studies led to the discovery of several patient-specific variants associated with the disease. For instance, performing a large case-control study, Lopes *et al.* provided important information on the frequency of mutations in *DMRT1*, a gene implicated in the disease aetiology (Lopes *et al.*, 2013; Lima *et al.*, 2015; Oud *et al.*, 2019). For many CNVs reported in the literature, a lack of information on inheritance does not allow one to assess whether these CNVs deserve further studies. Parental information can help determine whether a CNV is inherited from a fertile father (and therefore less likely to explain male infertility in the proband), whether it is maternally inherited or *de novo* in the germline of the infertile man. To date, only one article has been published on the role of *de novo* and maternally inherited mutations in male infertility (Hodžić *et al.*, 2020). The authors used WES to detect and interpret *de novo* and inherited SNVs in a cohort of 13 men affected by idiopathic azoospermia and their parents. They identified *de novo* variants in 5 novel candidate disease genes (*SEMA5A, NEURL4, BRD2, CD1D* and *CD63*) and 3 potentially pathogenic maternally inherited mutations in genes previously associated to male infertility. Unfortunately, CNV analysis was not conducted by these authors. Thus, this thesis represents the first study on the role of *de novo* (outside the chromosome Y) and maternally inherited CNVs in severe male infertility.

1.5 CNV Detection from WES Data

In 2015, the rate of *de novo* CNV events was estimated to be 0.0154 per generation for CNVs larger than 100kb (Kloosterman *et al.*, 2015). This was, however, a rough estimate, and a recent population study (Collins et al., 2019) demonstrated that our knowledge is highly biased by the sequencing techniques and the structural variants detection methods utilised. Therefore, the actual rate of *de novo* CNVs in the general population is likely higher, especially if we consider *de novo* CNVs of all sizes. This rate might further increase in cohorts of patients if *de novo* CNVs play a major role in the disease. As an example, a 2014 study (Gilissen *et al.*, 2014) showed that ~20% of severe intellectual disability cases caused by *de novo* events were explained by a *de novo* CNV. Also, a recent investigation on large cohorts of unaffected and autism spectrum disorders (ASD) families showed a significantly higher rate of *de novo* structural variants in ASD patients (Belyeu *et al.*, 2021).

In these studies, *de novo* variant discovery was performed using WGS data, which provides greater sensitivity than WES. Nevertheless, WES is an undervalued resource for CNV analysis. Many studies, also outside the field of male infertility, limit their WES analysis to SNVs and INDELs. This is likely to be due to the bioinformatic resources and experience needed to conduct CNV analysis from this type of data and the absence of a largely shared standard methodology, contrarily to the well-established SNV detection workflows. Even in my personal experience, a standard pipeline for CNV detection was not established for the WES data generated at the Faculty of Medical Science at Newcastle University before 2018. Although not always exploited, CNV analysis from WES data is helpful to improve the diagnostic yield as well as to identify pathogenic and candidate pathogenic CNVs, as several studies have demonstrated (Epilepsy Phenome/Genome Project Epi4K Consortium, 2015; Gambin *et al.*, 2017; Pfundt *et al.*, 2017; Corbett *et al.*, 2018; Marchuk *et al.*, 2018). Microarrays were for a long time the primary genomic technology used for CNV analysis (Miller *et al.*, 2010). Compared to them (and WGS), WES covers only the 1-2% of the genome corresponding to the coding region. The discontinuous signal in WES data results is systematic biases such as inaccurate read mappability, biases introduced by the PCR assay needed for the exome library amplification (e.g., GC-content bias) and sequencing batch effects (Teo *et al.*, 2012). For this reason, CNV detection from WES data is limited to depth-of-coverage methods (Lelieveld *et al.*, 2016). This technique uses the WES data from multiple individuals to normalise the sequencing read counts within genomic windows for each sample. This normalisation step reduces the variability caused by the systematic biases. The normalised read counts for each window of a single individual are then compared to those of the other samples. Finally, CNVs are inferred correlating the increases or decreases of the normalised reads counts to gains and losses, respectively.

CNV discovery from WES data does not allow the characterisation of the CNV breakpoints within the introns, nor the detection of the position, the orientation, and the number of additional copies. The sensitivity of WES for CNVs has been compared to that of widely used medium resolution microarrays (de Ligt *et al.*, 2013; Lelieveld *et al.*, 2016). A 2013 study that compared the performance of WES and microarrays in CNV discovery found a 88% concordance (de Ligt *et al.*, 2013). In that analysis, WES failed to detect a single exon deletion, and the authors suggested that its minimum resolution is three exons. This conclusion has been supported by other studies (Fromer, Jennifer L Moran, *et al.*, 2012; Krumm *et al.*, 2012). More recent WES-based analyses identified smaller CNVs encompassing 1 or 2 exons (Gambin *et al.*, 2017; Marchuk *et al.*, 2018; Tsuchida *et al.*, 2018). This suggests that the minimum resolution is dependent on the quality of the WES data and the algorithms used for CNV detection. For this reason, researchers might rely on several CNV detection tools when analysing their WES data. This possibly increases the sensitivity for CNVs but has the downside of extending the length of the analysis due to the greater number of calls and potential false positives. With improved sequencing performances and better algorithms for CNV detection, it has been suggested that WES will eventually substitute conventional microarrays for CNV analysis (Tsuchida *et al.*, 2018). Such strategy would allow SNV and CNV detection in the coding region with a single test and eliminate the cost of a dedicated assay for CNVs, such as a microarray test, which cost ranges approximately between $200 and $500 (Aston, 2014).

1.6 Project Aims and Outline of the Chapters

From this introduction, it is hopefully clear that male infertility genetics is an understudied field in medicine, and many questions are yet to be answered. The work presented in this thesis explores the role of CNVs in male infertility, using WES data from various patient (and parents) cohorts. My goal is to improve our understanding of the genetics of male infertility.

In the next chapter, Chapter 2, I describe the materials and the methods used throughout this thesis.

In Chapter 3, I assess the suitability of our WES data for CNV analysis. Also, I investigate the consequences of performing WES using different exome enrichment kits and DNA extracted from different tissues on WES coverage and CNV detection.

In Chapter 4, I investigate the role of *de novo* CNVs in male infertility using the WES data from 183 azoospermia or severe oligozoospermia patient-parent trios. The findings presented in this chapter are part of our pre-print paper currently in the BioRxiv database (Oud *et al.*, 2021) (see Appendix A).

In Chapter 5, I analyse the WES data from the same trio cohort to explore the difference between paternal and maternal CNVs in male infertility patients and assess the possible role of maternally inherited CNVs in the aetiology of severe quantitative male infertility.

In Chapter 6, I investigate the CNV burden in another cohort of 142 patients with quantitative male infertility to identify CNVs of potential diagnostic interests or affecting the novel candidate disease genes discovered in the previous chapters.

In Chapter 7, I use the WES data from 24 patients with asthenoteratozoospermia to identify CNVs of potential clinical interest that might cause this disease. The findings in this chapter are included in a publication in the American Journal of Human Genetics (Liu *et al.*, 2021) (see Appendix A).

Finally, in Chapter 8, I discuss the results of all the projects conducted during my PhD and their implications for the field. I also discuss how male infertility diagnostics could be improved and ideas on the design of future studies.

Chapter 2: Materials and Methods

2.1 Recruitment of Patients With Azoospermia and Severe Oligozoospermia

From 2007 to October 2017, a total of 331 male infertility patients with unexplained non-obstructive azoospermia and severe oligozoospermia, with or without asthenozoospermia, were recruited at Radboudumc outpatient clinic in Nijmegen (the Netherlands). Additionally, 45 patients with similar phenotypes were recruited at the Fertility Clinic, NHS Foundation Trust in Newcastle upon Tyne (UK) between January 2018 and January 2020. The reference values and the semen nomenclature used follow the WHO guidelines (World Health Organization, 2010). A blood sample was collected from the patients during the clinical evaluation and their respective fertility centre. All patients underwent karyotyping and AZF deletions screening with negative results as inclusion criterium. We also obtained a saliva sample from the parents of 201 patients. DNA was extracted from blood and saliva samples. All the participants in the study gave written consent for the analysis of their DNA and the evaluation of their clinical data. The study protocol was approved by the respective Ethics Committees and/or Institutional Review Boards (Nijmegen: NL50495.091.14 version 4, Newcastle: REC Ref: 18/NE/0089).

2.2 Recruitment of Patients With Asthenoteratozoospermia

A total of 24 patients with asthenoteratozoospermia were recruited at Monash University in Melbourne (Australia). These patients did not show any other obvious pathological phenotype and their hormone levels, secondary sexual characteristics and development of male external genitalia were normal. The patients were negative for chromosomal abnormalities and AZF deletions. Sample collection was approved by the human ethics panels at three different sites: Monash Day Surgery (Clayton), Monash Medical Centre and Monash University, Australia. DNA samples from the 24 patients were sent to the International Centre for Life in Newcastle upon Tyne for sequencing.

2.3 WES and Exome Enrichment Kits

WES was performed by the Genomic Core Facility (Newcastle University, UK) at the International Centre for Life on an Illumina NovaSeq 6000 platform. Two exome enrichment kits were used: the Nextera DNA Exome sequencing kit (previously called TruSeq Rapid Exome) from Illumina and the Human Core Exome sequencing kit from Twist Bioscience. The Nextera kit targets 45 Mb of exonic content (covers ≥ 98% of the RefSeq, Consensus coding sequence (CCDS), and Ensembl coding content), and according to the manufacturer, it achieves an excess of 75% of on-target sequencing reads (https://support.illumina.com.cn/content/illumina-marketing/en/products/by-type/sequencing-kits/library-prep-kits/truseq-rapid-exome.html). The Twist kit targets the CCDS and captures 33 Mb. According to manufacturer, it should cover 99% of ClinVar variants (https://www.twistbioscience.com/products/ngs/fixed-panels/human-core-exome?tab=overview).

2.4 WES Data Processing and Basic QCs

The initial processing of the sequencing data was performed by the Bioinformatics Support Unit (Newcastle University, UK). The sequencing reads were aligned to the human genome assembly GCRh37.p5 (hg19) using Burrows-Wheeler Alignment (BWA-mem) version 0.7.17 (Li and Durbin, 2009). Indexing, duplicate marking, and recalibration were performed with SAMtools version 1.6 (Li *et al.*, 2009). Picard version 2.21.1 (*Picard Toolkit*, no date) and GATK version 4.1.4.1 (Van der Auwera *et al.*, 2013).

The Bioinformatics Support Unit also performed checks to confirm the relatedness of the trio members using Peddy (Pedersen and Quinlan, 2017). The package uses the genotype data of the samples to compare pairs of individuals. For each pair, the IBS2 and IBS0 metrics are calculated. The first represents the number of genomic sites in which the individuals share two alleles, while the second represents the number of sites in which no allele is shared. These parameters allow separating related and unrelated individuals. The same tool and the same genotype data were also used to confirm the sex of each individual. This analysis confirmed that the data from each sample corresponded to the correct individual and family. It also excluded labelling errors and sample swapping during DNA extraction and sequencing.

The Bioinformatic Support Unit and Dr Miguel Xavier used Picard version 2.21.1 (Picard Toolkit, no date) to determine the coverage for each sample.

The sequencing reads of the samples were examined using the corresponding binary alignment map (BAM) format file and the IGV visualisation tool (Thorvaldsdottir, Robinson and Mesirov, 2013).

The WES data from the trio cohort (reference code: EGAS00001005417) and the asthenoteratozoospermia patients (reference code: EGAS00001005018) are publicly available at the European Genome-phenome Archive (EGA) (https://ega-archive.org) both as BAM files and in FASTQ format. The fertility status of each individual is indicated.

2.5 Individual Target Coverage Comparison Between Samples Sequenced From Saliva-Derived and Blood-Derived DNA

GATK3 DiagnoseTargets version 3.8.1.0 was used to determine the coverage of each target of the Twist exome enrichment kit in 43 probands (sequenced from blood-derived DNA) and 43 fathers (sequenced from saliva-derived DNA), each with an average exome coverage > 50X (see chapter 3). The coverage of the targets labelled by the tool as "poor-quality", which indicated a high probability of mapping errors, was converted to 0. The coverage of each target in each sample was normalised dividing it by the average exome coverage of the corresponding sample. The normalised coverage of each target was compared between the groups of probands and fathers using the Wilcoxon rank sum test (two-sided). The p-values obtained for each comparison were corrected using the Bonferroni correction for multiple tests.

2.6 CNV Detection Tools for WES Data

Before performing the CNV analysis, I tested three CNV discovery bioinformatic tools and chose the one that was most suitable for my analyses. The first two, called CoNIFER and XHMM, have been used in multiple publications (Poultney *et al.*, 2013; Epilepsy Phenome/Genome Project Epi4K Consortium, 2015; Pfundt *et al.*, 2017; Tsuchida *et al.*, 2018) and evaluated in different analyses (Tan *et al.*, 2014; Yao *et al.*, 2017). The third one was an unpublished tool developed by Dr Aneta Mikulasova (Newcastle University, UK).

CoNIFER was originally published in 2012 (Krumm *et al.*, 2012). This package has the advantage of requiring a minimum of only 8 samples to perform the CNV calling (20 advised for optimal results) and provides a tutorial with a test dataset of 26 sample on its website (http://conifer.sourceforge.net/). It offers a quality check procedure to identify and exclude poorly performing samples and a visualisation tool. CoNIFER has not been updated since 2012. Due to incompatibility with the most recent versions of the libraries required to run CoNIFER, we were unable to install all its functions on our machines, and the website providing support is no longer available.

XHMM was published in 2012 (Fromer, Jennifer L. Moran, *et al.*, 2012). A detailed tutorial was later published in a separate paper (Fromer and Purcell, 2014). The package is designed to work with at least 50 samples and offers a tool to identify poor quality ones, similarly to CoNIFER. Other functions include a visualisation tool and a CNV statistical genotyping procedure to identify recurrently altered loci and batch biases. XHMM, despite being as old as CoNIFER, was more user-friendly to install and use, thanks to the detailed tutorial provided in the second paper. However, we experienced similar libraries incompatibilities, and some functions such as the visualisation and *de novo* CNV calling tools were not available to us. The support for XHMM is limited to a Google group where users can interact, and its webpage (http://atgu.mgh.harvard.edu/xhmm) is currently offline.

The third CNV tool integrated GATK4 package's read counts normalisation with custom R-based segmentation and visualisation (Mikulasova *et al.* unpublished) (www.github.com/AnetaMikulasova/CNVRobot). The GATK4-based pipeline (G4BP) was still in development when it was initially tested, but with support from Dr Mikulasova and technical assistance from Dr Miguel Xavier (Newcastle University, UK), it was easily installed and used. G4BP, contrarily to CoNIFER and XHMM, exploits a panel of controls to normalize the read count and allows to specify female and male controls to improve the CNV detection on the sex chromosomes. G4BP CNV visualisation is the most advanced of the three CNV tools, allowing detailed visualisation of the CNV profile for each chromosomal position. The graphical representations integrate SNPs data to identify and visualize genomic stretches of homozygosity and heterozygosity loss. Also, it integrates the CNV data of the Database of Genomic Variations (DGV) (MacDonald *et al.*, 2014), allowing visual assessment of the number of variants identified in the healthy population for each locus. Finally, it aggregates the data of the sequencing targets in the control samples to show frequently altered loci. G4BP visualisation was easily adaptable for both singletons and trio analyses and enabled trio-based visualisation and controls-case comparison.

A detailed benchmark and extensive comparison of the three packages were not performed as it was not the aim of this project. However, the easier integration, the trio-based options and the advanced visualisation provided by the G4BP tool made it the default package for the CNV analyses performed during my PhD.

2.7 CNV Detection With G4BP CNV Tool

The G4BP requires the BAM files of the samples to detect the CNVs. The tool used ~200 parental samples as unaffected controls for the read count normalisation. The sequencing read depth of each target of the exome enrichment kits is influenced by several factors such as the PCR amplification performed during the library preparation, the GC content of the genomic region sequenced, and the quality of the reads mapping against the reference genome. The normalisation step uses the GATK4 algorithm to remove the noise caused by these systematic biases of WES based on a model constructed from the panel of controls. Using the normalised read depth data, then a custom R-based segmentation procedure infers the copy number status of each locus. The visualisation step generates the CNV plots, described in the next paragraph, using a custom script that uses the R package KaryoploteR (Gel and Serra, 2017). Other packages required for using the pipeline are part of the Tidyverse collection (Wickham *et al.*, 2019).

To obtain similar results as the ones presented in this thesis, I recommend following the same sample exclusion criteria described in section 2.11 and using a similar number of controls (~200) for the read count normalisation.

The data processing was performed on a 24 CPU Linux-based machine with ~1.5 terabytes of memory ram. The average processing time, including CNV calling and visualisation, is estimated to be between 3 to 5 hours per trio or per patient with male and female controls. Processing time is influenced by the exome coverage of the sample under examination and consequently by the size of the corresponding BAM file.

2.8 G4BP Visualisation and Plot Description

In this paragraph, I introduce the reader to the different types of plots that the G4BP CNV tool produces, and I explain how to interpret them as well as the information that they provide.

2.8.1 Genome profile – Chromosome view – CNV plots

The G4BP pipeline produces three types of plots: Genome Profile (GP) plots (Figure 2.1), Chromosome View (CV) plots (Figure 2.2) and CNV plots (Figure 2.4). Each graph presents information for three individuals. The sample in the lower section (proband MI_01974 in Figure 2.1) is the individual under examination, while the one in the middle (Father MI_01974 in Figure 2.1) and the one on the upper section (Mother MI_01974 in Figure 2.1) can either be the parents (respectively father and mother) or male and female controls. In the GP plot, all the chromosomes are represented; in the CV plot, only a single one; and the CNV plot is a representation of a specific genomic region containing a deletion or a duplication. In every graph, for each sample, there is a log2ratio (number of observed normalised reads/number of expected normalised reads) (Log2R in the plots) profile and minor allele frequency (MAF) track. Here, MAF is defined as min(A, B)/(A+B), where A and B represent the frequencies of the two alleles.

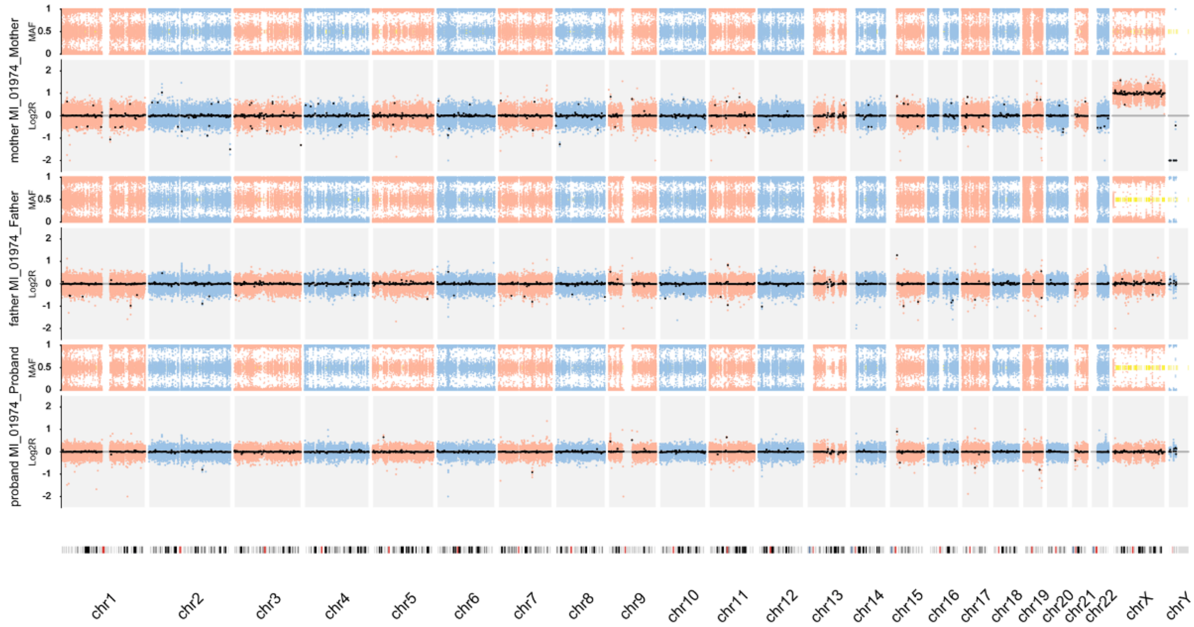


Figure 2.1. Example of a Genome Profile plot for a trio (trio 1974). For each member of the trio (from the top Mother, Father and Proband), there are a Log2ratio, representing the copy-number status of the sequenced segments on each chromosome, and a minor allele frequency plot, representing the allele frequency at each sequenced genomic position across all the chromosomes.

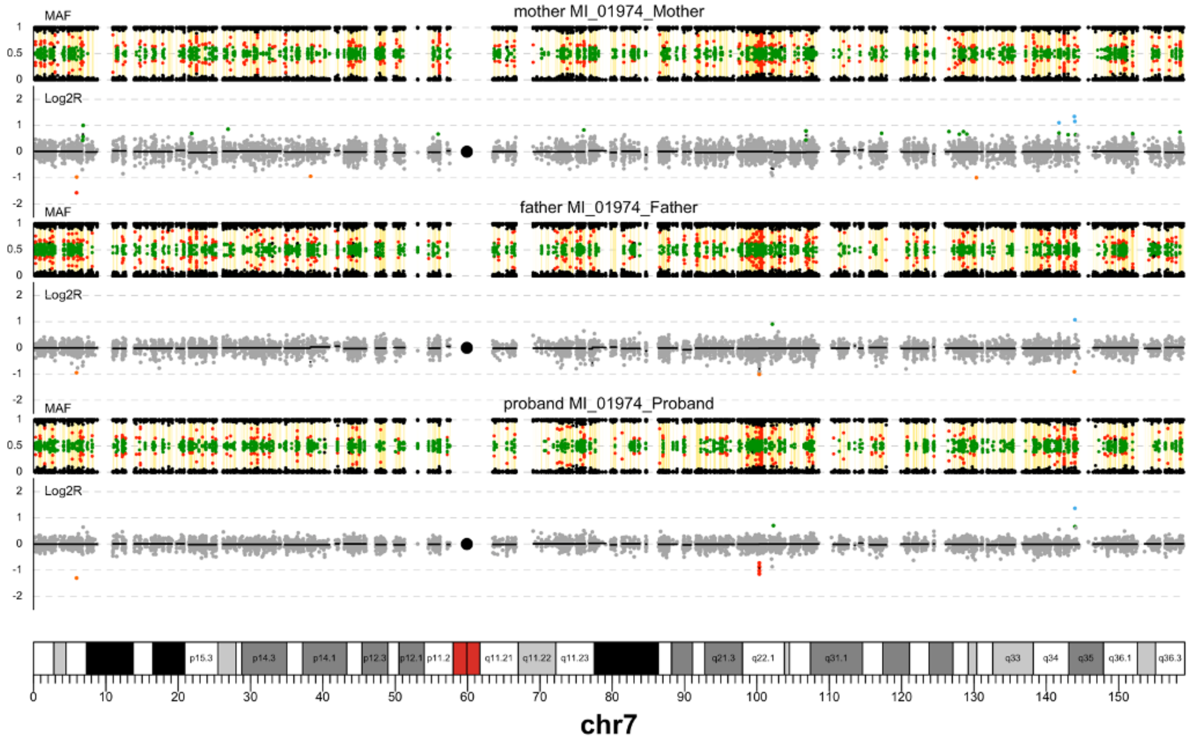


Figure 2.2. Example of a Chromosome View plot for chromosome 7 of trio 1974. For each member of the trio (from the top Mother, Father and Proband), there are a Log2ratio plot, representing the copy-number status of all the sequenced segments on the single chromosome, and a minor allele frequency plot, representing the allele frequency at each sequenced genomic position across the single chromosome.

The number of copies for each sequencing target is inferred from the number of observed normalised sequencing reads, which are compared to a reference built from the control dataset (i.e., unaffected individuals). Log2ratio = 0 indicates that a specific target in the sample has the same average coverage as the reference and consequently the same number of copies. An increase of the log2ratio represents the presence of additional copies, while a decrease indicates a loss. On the autosomal chromosomes, a log2ratio = -1 indicates the loss of one allele. The absence of the locus is instead conventionally indicated by a log2ratio of -2. The sex chromosomes are expected to have only one copy in the male reference, and the loss of that single copy will be represented by a log2ratio of -2. In females, the chromosome X will have a log2ratio = 1, due to the two copies of the chromosome, while the absent chromosome Y will have log2ratio = -2. Each log2ratio profile is paired with a MAF track positioned above. This graph indicates the SNP status in the corresponding region. Homozygous SNPs have MAF = 1 or 0, while heterozygous SNPs are indicated with MAF = 0.5. The SNP data is useful to confirm heterozygosity loss, reflected as the absence of heterozygous SNPs in a deleted region, and to identify stretches of homozygosity. These characteristics are common to all the graphs generated by the G4BP tool. The GP plot is used to identify the presence of chromosome-wide abnormalities (e.g., additional/missing chromosomes) and to confirm the sex of each individual. The CV plots show the profile of an entire chromosome and can be used to investigate the regions surrounding a CNV. These two types of graphs were also used to identify samples with a high level of noise. In noisy samples (Figure 2.3), the log2ratio signal is highly variable across the targets. Thus, the copy number status of the genomic segments is less clear than for other individuals (Figure 2.1 and 2.2). This characteristic can affect specificity and sensitivity and generate an extremely high number of CNV calls. The samples with this type of GP and CV plots and with a CNV count > 3 standard deviations (SD) above the mean (calculated from the corresponding cohort), such as Proband 1 (CNV calls = 60, average in probands = 11, SD = 5) in Figure 2.3, were excluded as done in other studies (Ruderfer *et al.*, 2016; Wang *et al.*, 2016).

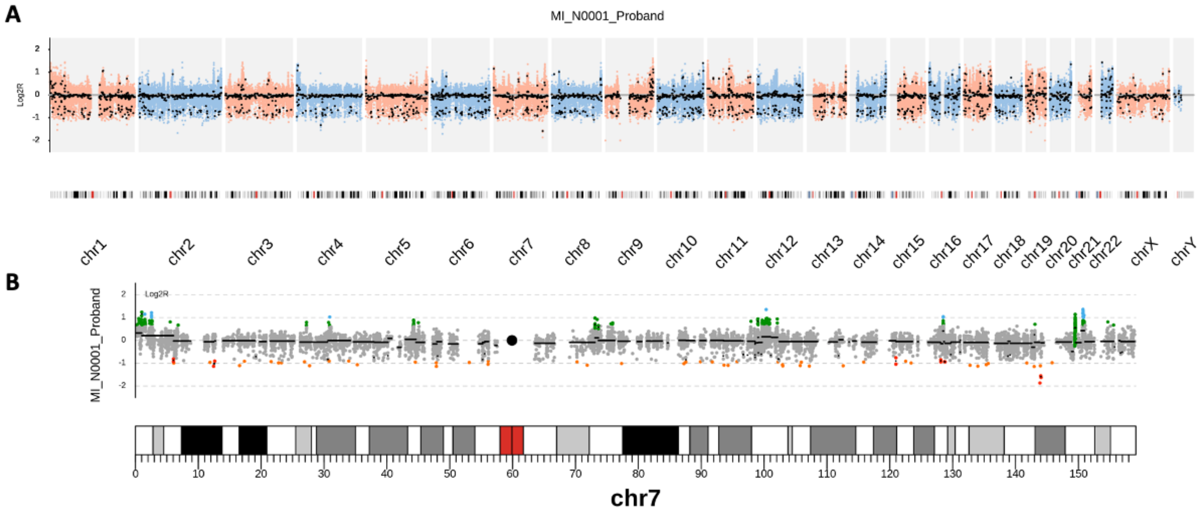


Figure 2.3. Example of a noisy sample. (**A**) GP plot and (**B**) CV plot of chromosome 7 of a noisy individual (proband 1). The high level of noise is indicated by the high Log2R variability of the genomic segments and the individual sequencing targets.

The CNV plots are used to investigate specific deletions or duplications, and they allow the visualisation of the single sequencing targets in the log2ratio profiles (Figure 2.4).

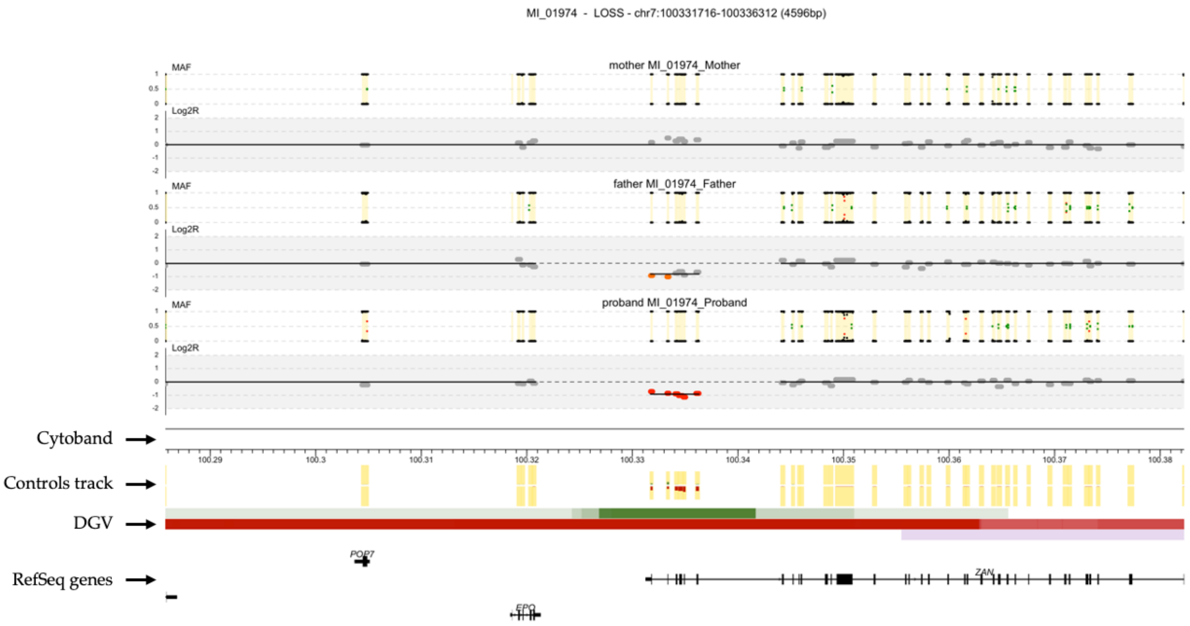


Figure 2.4. Example of a CNV plot for a paternally inherited deletion detected in proband 1974. For each member of the trio (from the top Mother, Father and Proband), there are a Log2ratio, representing the copy-number status of each sequenced segment within the genomic window, and a minor allele frequency plot, representing the allele frequency at each genomic position of the sequenced segments. In proband and father 1974 a heterozygous deletion has been detected (6 adjacent sequencing targets with a log2R = -1).

The CNV plots also display additional elements: beneath the log2ratio profiles, there are the corresponding cytoband and a control track, represented as yellow bars. These latter are one per target, and they schematically represent the frequency of loss and gain in the control samples. The height of a yellow bar indicates how many controls had high quality reads for that specific target. The more controls have the specific target indicating a loss, the more its corresponding yellow bar will be coloured in red from the middle to the bottom, while the more samples have the target indicating a gain, the more its yellow bar will be coloured in green from the middle to the top. In Figure 2.4, the targets within the deletions have shorter yellow bars compared to the surrounding targets. This indicates that those targets were uninformative (no reads mapped or low-quality mapping) in some controls, which were therefore excluded. The red colour in the bars shows that those targets showed a loss in multiple controls. This information is useful to identify frequently variated loci in the control cohort. In Figure 2.4, for example, the heterozygous deletion is likely to be a frequent polymorphism. The horizontal bars beneath the control track represents the variants from the DGV. The green bars represent the duplications, while the red represents the deletions. The purple bars indicate a complex structural variation. The more intense is the colour of the bars, the more corresponding variants were listed in the database for that location. This gives an immediate indication of the variability of the locus in a healthy population dataset. In the genomic loci represented in Figure 2.4, several large deletions and duplications have been reported in the samples of the DGV. Finally, the section on the bottom shows the RefSeq genes in the corresponding genomic region.

2.9 Bar Charts and Box Plots

In this thesis, I often use bar charts and box plots to present my data. All these plots were generated with the R package ggplot2 version 3.3.3, which is part of the Tidyverse collection (Wickham *et al.*, 2019).

2.10 Statistical Test

To test the differences between two groups, I used the Wilcoxon rank sum test.

2.11 Samples' Exclusion Criteria for the CNV Analysis

Samples for which a 30X average exome coverage was not achieved were excluded. Samples with a genome profile and chromosome view plots showing a high level of noise (example in Figure 2.3) as well as a total number of CNVs > 3 SD above the average (calculated from the corresponding cohort) were excluded as done in other studies (Ruderfer *et al.*, 2016; Wang *et al.*, 2016). Patients who naturally conceived with their partner after being enrolled were also excluded from the study. A total of 183 patient-parent trios with azoospermia or severe oligozoospermia, 142 patients with the same phenotypes for which parental samples were not available and 24 patients with asthenoteratozoospermia were retained for the CNV analysis.

2.12 CNV Annotations

All the CNVs identified were annotated using the tool AnnotSV version 2.0 (Geoffroy *et al.*, 2018) with default settings. This tool provides information for each CNV identified, including the genes it encompasses as well as the number and IDs of the CNVs listed in the DGV Gold Standard (MacDonald *et al.*, 2014) that have at least a 70% overlap. Every gene involved in a CNV was also annotated with the corresponding pLI score provided by the GnomAD database v2.1.1 (Karczewski *et al.*, 2020).

2.12.1 CNV prioritisation and classification criteria

To identify CNVs of potential clinical interest for male infertility, I first prioritised CNVs present in < 1% of the samples included in the DGV Gold Standard to select those rare in a dataset of healthy individuals. Secondly, I selected the CNVs encompassing at least 10 sequencing targets as the more exons are involved in a variant, the higher is its probability of being deleterious. To identify genes possibly affected by heterozygous dominant CNVs, I looked for those with a pLI score > 0.9. Such a score indicates a high probability of that gene to be intolerant to loss-of-function variants (Lek *et al.*, 2016).

For the trio cohort, I visually inspected the CNV plots of all the variants identified. I classified as "*de novo*" the CNVs present in the proband but absent in both his parents. I classified as "paternally inherited" the CNVs present only in the proband and his father and as "maternally inherited" those present only in the proband and his mother.

I also prioritised CNVs encompassing genes carrying a *de novo* mutation (in the probands of the trio cohort) classified as possibly causative of the disease or of unknown significance by Dr Manon Oud and Hannah Smith (see Appendix B). The classification criteria are described in Oud *et al.* 2020.

At the end of 2019, the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen) published a scoring framework to classify CNVs with an evidence-based method (Riggs *et al.*, 2020). These guidelines are advised to report constitutional CNVs maintaining consistency and transparency across clinical laboratories. Despite recognizing the importance of using these standards in a clinical setting, our studies aim to identify novel candidate genes for male infertility and not to provide a definitive genetic diagnosis to the patients. Also, the ACMG and ClinGen guidelines are often used to evaluate CNVs encompassing several megabases identified in patients with severe neurodevelopmental disorders or multiple congenital abnormalities, as mentioned in

Riggs *et al.*, 2020. We expect the candidate causative CNVs identified in patients with isolated male infertility to be smaller, since these men do not have other observable pathogenic phenotypes and a single impaired spermatogenesis gene might be sufficient to cause the disease, and to occur in regions not previously studied in the field, considering the very few genomic loci confidently associated to severe male infertility so far. For this reason, here I provide the criteria that I used to classify the CNVs identified in our studies. A heterozygous CNV was defined as likely pathogenic for male fertility if:

1) the CNV was either *de novo* or maternally inherited and it was not present in any other father of the trio cohort.

2) the CNV was present in less than 1% of the individuals of the DGV Gold Standard.

3) the CNV involved at least 10 exons.

4) the CNV affected at least 1 gene with pLI score > 0.9 that may play a role in spermatogenesis (according to published literature or gene databases), or that has been previously associated to a male infertility phenotype similar to that described in the patient under examination.

A hemizygous CNVs was classified as likely pathogenic for male fertility if it met criteria 1 and 2 and encompassed a gene that may play a role in spermatogenesis (according to published literature or gene databases), or that has been previously associated to a male infertility phenotype similar to that described in the patient under examination.

The CNVs with unknown inheritance (i.e., when parental samples were unavailable) that met the criteria 2, 3, and 4 (pLI score and size were not considered when the CNV was hemizygous) were classified as possibly pathogenic for male fertility.

2.12.2 Databases employed

As mentioned previously, I used the DGV Gold Standard to select the rare CNVs identified in our cohorts of patients. The entire DGV includes structural variants data from healthy individuals only, and to date (4th July 2021), it comprises 983,845 CNVs and 4,083 inversions from 75 studies (http://dgv.tcag.ca). The DGV Gold Standard is a subset of the DGV that includes only CNVs that were found in two separate studies and in at least two different samples. To investigate the frequency of the CNVs identified in our cohorts in other dataset, I also used the GnomAD-SV version 2.1

database (Collins *et al.*, 2020). This dataset includes structural variants data from 10,847 genomes of unrelated individuals from disease-specific and population genetic studies (4th July 2021) (https://gnomad.broadinstitute.org). Another large database that contains CNV data is the DECHIPER database (Firth et al., 2009) (https://www.deciphergenomics.org). It contains data from 39,910 patients (7th October 2021) with different pathogenic phenotypes. The database includes a collection of CNVs involved in microdeletion and microduplication syndromes associated with developmental disorders. These CNVs often encompass several megabases and genes. It is unclear whether the impairment of these numerous genes cause infertility as well as neurodevelopmental disorders. Therefore, it would be complex to use this data to prioritise candidate causative variants among the CNVs identified in our cohorts. For this reason, the DECHIPER database was not used for CNV prioritisation.

Throughout the thesis, I report the RNA and protein expression levels of specific genes in human tissues. This information was retrieved from the Protein Atlas database. This project aims to map all the human proteins in cells, tissues and organs (Uhlén *et al.*, 2015).

Also, I often cite that impairment of specific genes causes pathogenic phenotypes in mice. The mouse data were retrieved from the Mouse Genome Informatics database (http://www.informatics.jax.org). This website aggregates the data from studies on mice and provides the list of pathogenic phenotypes observed in mice when a specific gene of interest is disrupted.

To investigate whether the candidate disease genes identified had been previously associated with spermatogenesis or male infertility in published studies, I used the PubMed website (https://pubmed.ncbi.nlm.nih.gov). This database contains more than 32 million citations of biomedical literature (4th July 2021).

Protein interaction data were retrieved from the STRING database (Szklarczyk *et al.*, 2019). The database includes data from known and predicted protein-protein interactions. It covers 24,584,628 proteins from 5,090 organisms (30th July 2021).

2.13 CNV Validation

The *de novo* autosomal CNV identified in proband 953 (see chapter 4) was validated with a microarray assay. The assay was outsourced to the Northern Genetics Service at the International Centre for Life, Newcastle upon Tyne (UK). It was performed using the whole genome Illumina Infinium CytoSNP-850K v1.1 microarray platform, and the CNV analysis was conducted with the BlueFuse Multi v4.4 software.

The X-linked *de novo* CNV identified in proband 584 (see chapter 4) was validated with a q-PCR assay. The experiment was performed by Dr Bilal Alobaidi using the TaqMan Copy Number Assay designed for the *NXT2* gene (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's protocol.

The X-linked deletion identified in patient 2603 (see chapter 7) was validated with a PCR assay. Dr Bilal Alobaidi designed the PCR primers using the tool Primer3Plus (Untergasser *et al.*, 2007) and performed the experiment. In chapter 7 the position of the PCR products is described.

Chapter 3. Influences on WES coverage and CNV Detection

3.1 Introduction

NGS is progressively substituting Sanger sequencing for disease-gene discovery in male infertility research (Oud *et al.*, 2019). While Sanger sequencing is highly dependent on previous knowledge of disease genes, WES and WGS offer an unbiased approach and allow the investigation of all coding regions and of the entire genome, respectively. Despite its limitations, such as non-uniformity of coverage, limited structural variants detection ability and exclusion of the non-coding region, WES is an extremely valuable tool for studying Mendelian diseases (Bamshad *et al.*, 2011), and its cost is substantially lower than that of WGS (Schwarze *et al.*, 2018). WES has been used successfully for variant analysis in several fields, revealing pathogenic variants in known and novel genes (Pfundt *et al.*, 2017; Boraldi *et al.*, 2019; Kherraf *et al.*, 2019; Tong *et al.*, 2020). In male infertility, it mainly led to the identification of autosomal recessive candidate disease genes, of which 16 are to date considered disease genes (Oud *et al.*, 2019). However, it was never systematically used to investigate dominant disease genes and currently, only 4 dominant genes have unambiguously been linked to the disease (Oud *et al.*, 2019). The difficulties in acquiring parental samples for male infertility studies complicates the discovery of pathogenic *de novo* and maternally inherited variants, and the large mutational target expected for the disease requires large cohorts to be collected. By recruiting patient-parent trios for azoospermia and severe oligozoospermia cases, our research group hopes to address these problems and provide a first insight on the role of *de novo* and inherited dominant variants in a large group of families. In addition, we collected DNA from patients for whom parents were unavailable for both quantitative and qualitative male infertility.

3.1.1 CNV detection from WES data

One of the main topics of this thesis is CNV detection from WES data. This type of analysis is often neglected, and most researchers use WES data exclusively for SNV identification. This is likely due to the bioinformatic resources and knowledge needed to perform such analyses and the lack of widely accepted standards for CNV discovery workflows. Due to the limitations of CNV detection from WES data, explained in paragraph 5 of Chapter 1, the CNV detection is restricted to depth-of-coverage methods. Current literature demonstrates that, despite the limitations, it is possible to perform a robust CNV analysis from WES data (Epilepsy Phenome/Genome Project Epi4K Consortium, 2015; Pfundt *et al.*, 2017; Corbett *et al.*, 2018). Moreover, recent studies suggest that good quality WES data and algorithms even allow the detection of single or two-exons CNVs (Marchuk *et al.*, 2018; Tsuchida *et al.*, 2018).

Since coverage is a crucial factor in CNV detection from WES data, in this chapter, I explore the differences in coverage between WES data generated from DNA extracted from the two sources of material that were available to us (i.e., blood for patients and saliva for parental samples), as well as the impact of using two different exome enrichment kits for the sequencing. In the literature, different studies evaluated the consequences of using saliva-derived DNA instead of blood-derived DNA for WES (Kidd *et al.*, 2014; Zhu *et al.*, 2015; Ibrahim *et al.*, 2020). These analyses suggested that the DNA from both the tissues can generate high-quality data. However, none of these studies investigated the impact on CNV detection specifically. Here, I inspect the WES data of our samples to identify putative loci where coverage is systematically different in these two tissues. If detected, such differences could affect the sensitivity and specificity of CNV detection, particularly for *de novo* variants, where the same locus must be analysed in patient and parents.

3.2 Aims

In this chapter, I investigate the impact of using DNA extracted from different tissues and the impact of using different exome enrichment kits on WES coverage and CNV detection. This study aims to:

- Compare the WES coverage between samples sequenced with two different exome enrichment kits, evaluate their suitability for variant detection and assess the influence of the two kits on CNV calling.
- Investigate the consequences of using blood and saliva as starting material on WES coverage and evaluate the impact on CNV detection.

3.3 Results

From 2018 to 2020, our research group (specifically Dr Bilal Alobaidi and Lois Batty) in collaboration with the Genomic Core Facility at the International Centre for Life (Newcastle University), sequenced 764 DNA samples from different projects using WES. These samples were collected from men affected by severe isolated forms of male infertility. For 201 patients, parental samples were also sequenced. The Bioinformatic Support Unit of Newcastle University and Dr Miguel Xavier, a senior post-doc of our group, used the Picard package version 2.21.1 to generate the WES coverage data included in the next paragraph.

3.3.1 Influence of the exome enrichment kit choice on coverage and CNV detection

Initially, we wanted to evaluate the difference in coverage for WES data generated with different exome enrichment kits. The first 329 samples were processed using the Nextera DNA Exome enrichment kit from Illumina, while the subsequent 435 samples with the Human Core Exome enrichment kit from Twist Bioscience (see Chapter 2 for design and other characteristics). These samples were sequenced in 10 sequencing runs on the Illumina Novaseq 6000 platform. In each run, 96 samples were sequenced. These included samples from our laboratory and samples from other research groups when we could not provide all 96 samples. The two exome enrichment kits used have different designs. The Illumina kit targets 45 Mb, while the Twist kit 33 Mb. Their designs are both based on the consensus coding sequence (CCDS) (Pruitt *et al.*, 2009); however, the Illumina kit's design is also based on the RefSeq genes and the Ensembl coding content, according to the manufacturer. The mean target coverage (in this chapter called "average exome coverage") was used to compare the two groups. This metric represents the average sequencing coverage of the targeted region sequenced in the experiment (i.e., all the exons targeted by the exome enrichment kit). Figure 3.1 shows the coverage distribution of the samples of the two groups.

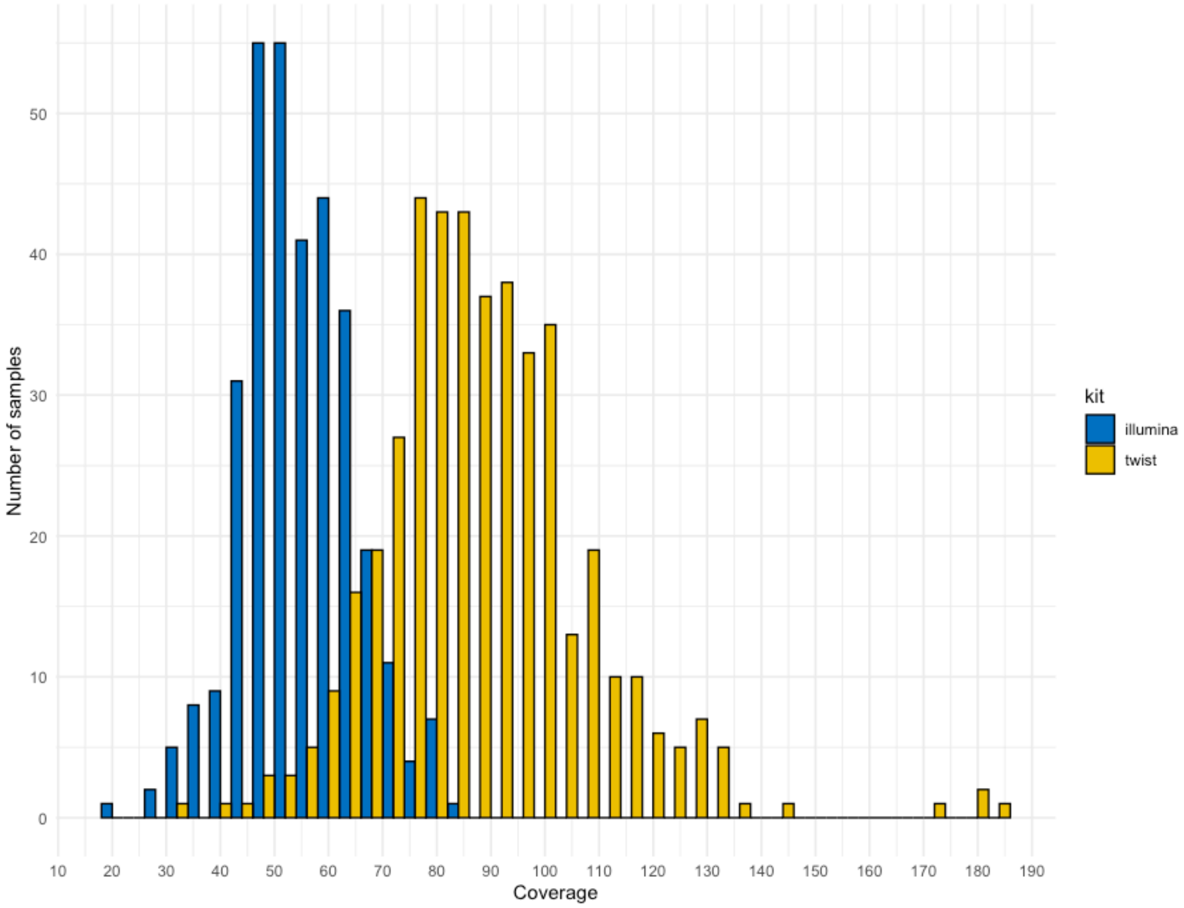


Figure 3.1. Distribution of the Nextera (blue) (329) and Twist (yellow) (435) samples according to their average exome coverage. The Twist samples have a generally higher average exome coverage compared to the Nextera samples.

The data showed that the average exome coverage is significantly lower when using the Illumina enrichment (average 55-fold) kit instead of the Twist enrichment kit (average 87-fold) (p-value < 2.2e-16). Three Illumina-Enriched (IE) samples (proband 927, father 546 and father 873) had average exome coverage < 30X and were excluded from further analyses.

We also wanted to investigate whether the different enrichment kits influenced the number of CNVs identified. For this test, we compared the CNVs detected from the WES data of 183 probands that were part of the patient-parent trio cohort. This cohort included 92 samples enriched with the Illumina kit and 91 enriched with the Twist kit. For the first group, a total of 1312 CNVs were called, while 742 CNVs for the second (Figure 3.2A), with an average of 14 and 8 CNVs per sample, respectively. In the IE samples, more CNVs were detected and a slightly higher percentage of small CNVs (i.e., CNV involving 3 or 4 targets) (Figure 3.2B). However, IE samples had a general higher noise level in the CNV data compared to Twist-Enriched (TE) samples. To better explain this difference, I used as an example proband 1724. The exome of proband 1724 was enriched and sequenced initially with the Illumina kit as part of the trio analysis and, in a separate experiment, with the Twist kit. The genome profile and chromosome view plots produced by the CNV pipeline (see Chapter 2 for plot description) for the IE sample show a higher Log2R variability amongst the targets compared to the TE sample's plots (Figure 3.3A-3.3B). Figure 3.3B shows how the more uniform Log2R of the targets in the TE sample allowed the identification of a duplication not segmented in the IE sample. For proband 1724, 5 CNVs were identified using the TE sample data, and 16 using the IE sample data (Table 3.1). 7 CNVs identified in the IE sample encompassed 3 or 4 targets, in contrast to only 1 CNV detected in the TE sample. Importantly, 3 CNVs in total were identified in both samples, indicating that they are likely real and not a detection artefact.

Overall, the CNV detection was possible for samples enriched with both exome enrichment kits. The higher amount of noise in the IE samples data suggests a lower specificity compared to TE ones. However, IE samples produced a much higher number of CNV calls. These differences must be considered when comparing CNV data from IE and TE samples in future analyses.

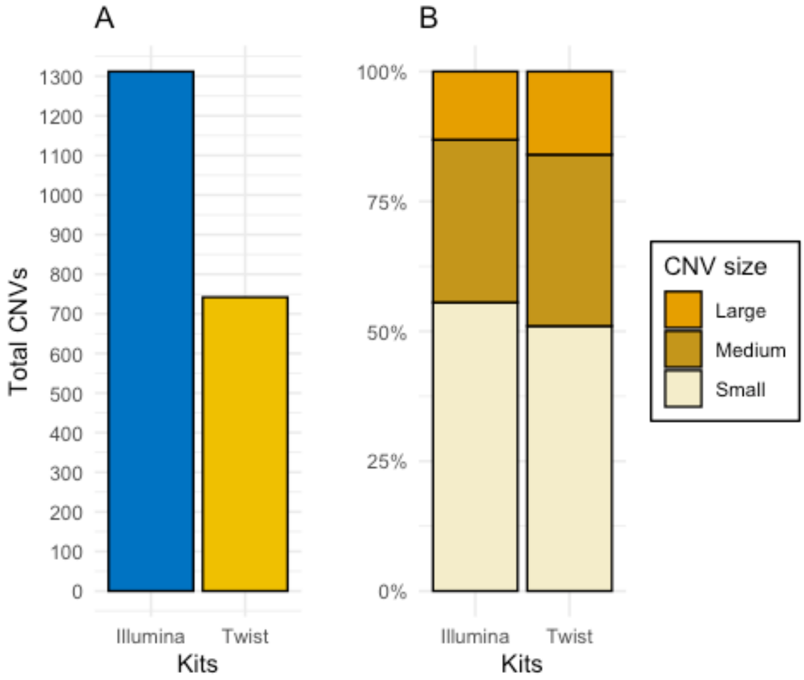


Figure 3.2. **A.** Number of CNVs identified in the IE and TE samples. The number of CNVs identified in the IE samples is substantially higher than the number of CNVs detected in the TE samples. **B.** Percentage of Small (3-4 targets), Medium (> 4 targets and < 10 targets), Large (>= 10 targets) CNVs identified in the IE and TE samples. The IE samples have a slightly higher percentage of Small CNVs.

	Proband 1724	
	Illumina enriched sample	Twist enriched sample
Average exome coverage	61X	91X
CNV calls	16	5
Small CNVs (3-4 probes)	7	1

Table 3.1. Average exome coverage and CNV data for proband 1724, enriched and sequenced with both Illumina and Twist exome enrichment kits. The Twist enriched sample have a lower number of total CNV calls and Small CNVs, despite having a higher average exome coverage than the Illumina enriched sample.

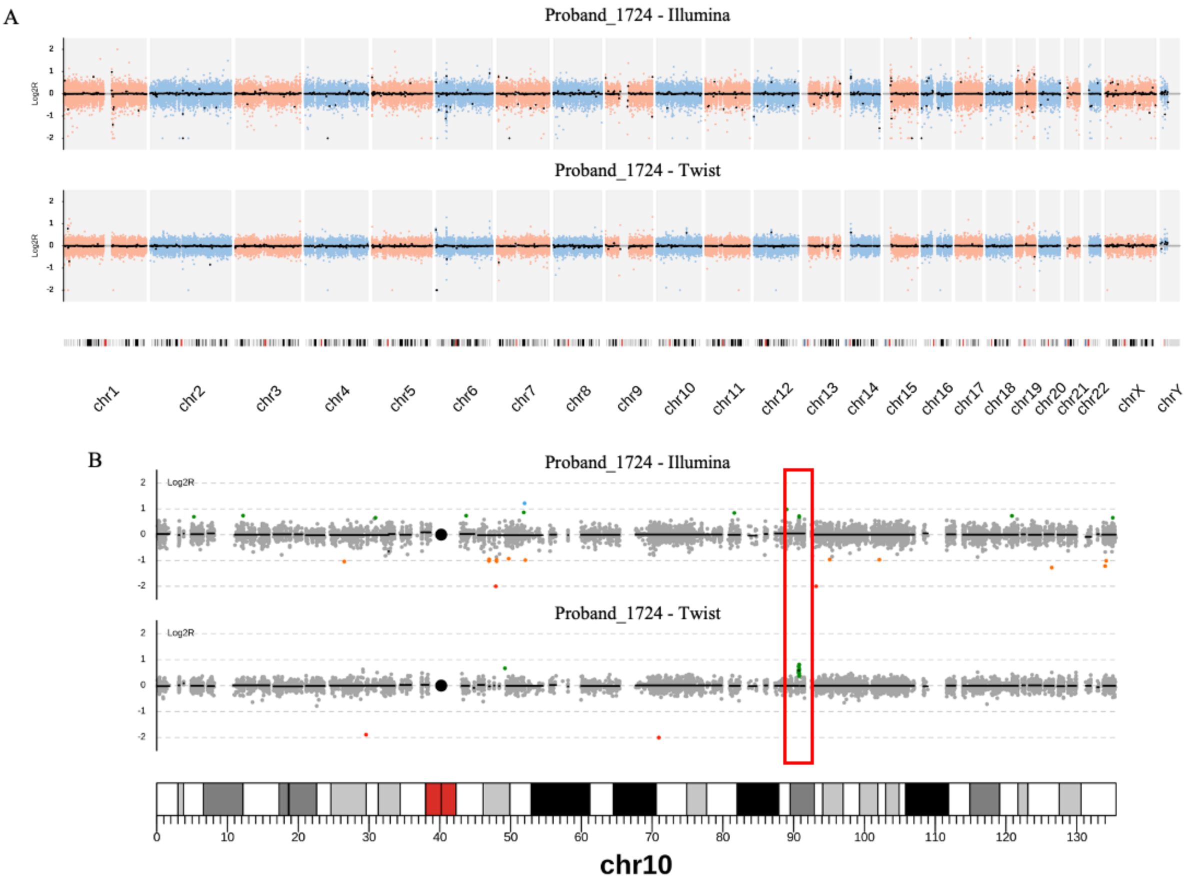


Figure 3.3. **A.** Genome profile of proband 1724 from the IE and TE samples. **B.** Chromosome 10 profile of proband 1724 for the IE and TE samples. In the red box, a duplication that was detected in the TE sample, but failed the detection in the IE sample due to the noise.

3.3.2 Influence of the DNA origin on WES coverage

The infertile patients in our studies provided a blood sample in the clinics where they were referred. These samples were used to extract DNA and perform WES. The parents that agreed to participate in the study received a saliva collection kit at home. Then they returned it to us for DNA extraction and WES. We compared the average exome coverage distribution generated from blood-derived DNA to that generated from saliva-derived DNA (Figure 3.4). For this comparison, we used the WES data of all the collected trios regardless of their final inclusion in the variant analysis (Table 3.2). The samples excluded in the previous paragraph were not included in this evaluation.

	Illumina	Twist
Probands (blood)	107	95
Fathers (saliva)	109	92
Mothers (saliva)	110	91

Table 3.2. The number of samples sequenced from blood- and saliva-derived DNA for each exome enrichment kit used for comparison.

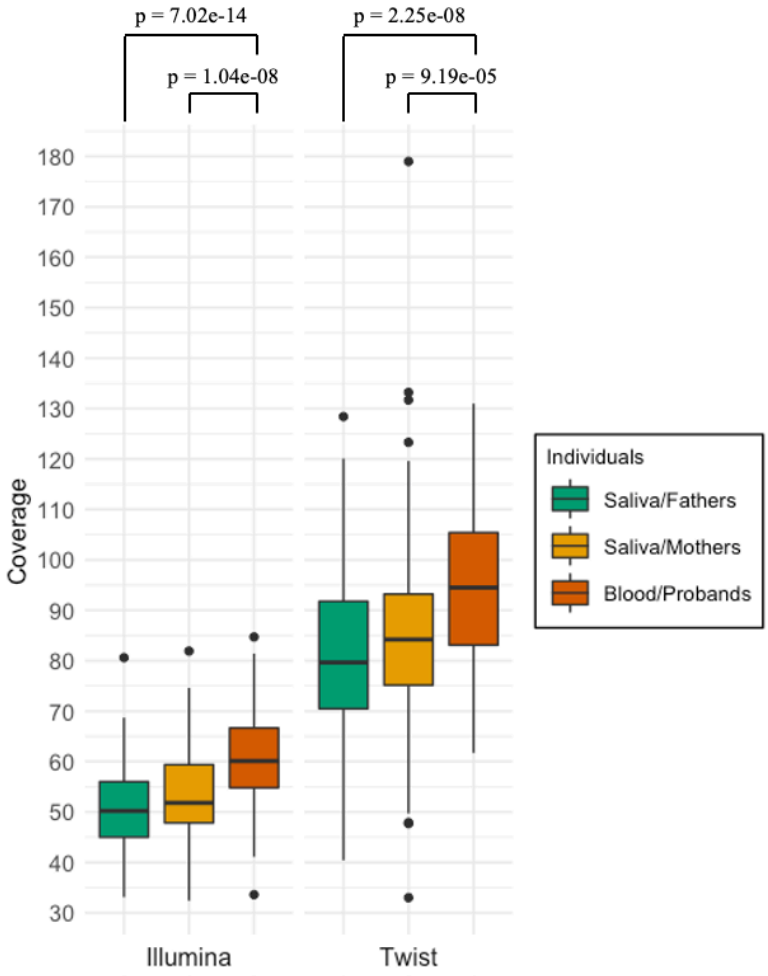


Figure 3.4. Average exome coverage distribution of the trio members separated by exome enrichment kit. WES was performed with saliva-derived DNA for mothers and fathers. Proband samples were sequenced exclusively from blood-derived DNA. The samples sequenced from blood-derived DNA showed a significantly higher average exome coverage than the samples sequenced from saliva-derived DNA, independently from the enrichment kit used.

The results showed a significantly higher average exome coverage for samples sequenced from blood-derived DNA (probands) compared to those sequenced from saliva-derived DNA (mothers and fathers), independently from the enrichment kit used (Figure 3.4). Average exome coverage > 60X for all the samples was achieved only for blood-derived DNA samples enriched by the Twist enrichment kit. For saliva-derived DNA, the average exome coverage obtained using the Twist kit was higher than that obtained using the Illumina kit. The percentage of samples with average exome coverage > 60X increased from 18 to 91% when using the Twist kit instead of the Illumina kit. Importantly, this data shows that, despite the differences, using both DNA sources and both the exome enrichment kits, we were able to generate WES data with an average exome coverage > 30X for the large majority of the samples.

3.3.3 Influence of the DNA origin on individual gene coverage

While it was clear that the saliva-derived DNA generated WES data with a lower average exome coverage than blood-derived DNA, we wondered whether there was also a systematic variation in coverage for individual genes. To test this hypothesis, the sequencing coverage of each target of the Twist enrichment kit was compared between groups of samples sequenced with DNA extracted from both tissues. WES data from 43 fathers (saliva-derived DNA) and probands (blood-derived DNA) with average exome coverage > 50X was selected for the comparison. The GATK3 DiagnoseTargets tool was used to retrieve the single sequencing targets coverage information (see Chapter 2). For each target, the average coverage in the blood and saliva group was calculated and used for evaluation (Table 3.3).

	Blood-derived DNA samples (43 probands)	Saliva-derived DNA samples (43 fathers)
Average exome coverage	94X	82X
Total number of targets in the exome enrichment kit (AUT - X - Y)	194283 - 6819 - 548	
Targets with no informative reads in both groups (AUT - X - Y)	2096 - 426 - 337	
Targets with >=10X coverage but <10X in the other group	106	6
Number of genes encopassed by targets with >=10X coverage but <10X in the other group	93	6

Table 3.3. Comparison of two groups of samples sequenced from blood- and saliva-derived DNA. The samples of both groups were sequenced using the Twist exome enrichment kit. The average exome coverage was higher for the blood-derived DNA samples as well as the number of targets with >=10X coverage but less than 10X in the other group. AUT = autosomal chromosomes, X = chromosome X, Y = chromosome Y.

The number of targets with no informative reads (no mapped reads or low-quality mapping) in both groups accounted for 1% on the autosomes, 6% for the chromosome X and 61% for the Y chromosome. The chromosome Y had no informative reads for more than half of its targets, probably due to limitations of short-read mapping. To test the presence of genes better covered in one group compared to the other, I looked for sequencing targets with >= 10X coverage in one group and < 10X in the other. This threshold is commonly used in the SNV analysis to exclude loci that might not provide sufficient coverage for calling heterozygous variants (Rehm *et al.*, 2013; Kong *et al.*, 2018). If a gene was sufficiently sequenced in one group (e.g., probands) but not in the other (e.g., parents), the accuracy for calling *de novo* mutations would likely decrease. In total, 106 targets were sequenced at a minimum of 10-fold coverage in the blood group but at less than 10-fold in the saliva one. This represents 0.079% of the total targets present in the exome enrichment kit. Most of these targets were distributed in different genes, with 90 out of 106 affecting each a single gene. 8 genes were encompassed by two of these targets and only one gene (*NOMO1*, 31 exons in total) by 3 targets. Vice versa, only 6 targets were sequenced at a minimum of 10-fold coverage in saliva-derived DNA samples and less in blood-derived DNA samples. They affected 6 different genes. Overall, very few targets were sufficiently covered in one group but not in the other. The spread of these targets amongst several genes suggests that this difference will have a limited impact on the variant analyses of the coding region. The different number of targets identified in the two groups might be the consequence of the different average exome coverage.

CNV detection from WES data might be influenced by systematic biases of the sequencing (Hehir-Kwa, Pfundt and Veltman, 2015). Some sequencing targets may be systematically deeper sequenced in blood-derived DNA than in saliva-derived DNA or vice versa due to the tissue of origin. Such differences might influence the results of the CNV workflow, which is based on the depth of coverage and uses parental samples as unaffected controls (see Chapter 2 for detail). To test this hypothesis, I looked for sequencing targets that showed a higher coverage in one group compared to the other using the data of the 43 patient-father pairs. I first normalised each individual target coverage of each sample with the respective sample's average exome coverage to account for the variability within the groups. Then I compared the normalised target coverage of the 201,650 targets between the blood and saliva groups. I performed the Wilcoxon rank sum test (two-sided) for unpaired groups and subsequently corrected the p-values obtained with Bonferroni correction for multiple tests. The targets that showed a significant bias (Bonferroni corrected p-values < 0.05) toward one group were 4 (0.002% of the total). Specifically, 4 targets covering 4 out of 5 exons of *PRSS1*, a gene that encodes a trypsinogen (Koshikawa *et al.*, 1994), were sequenced at higher coverage in the saliva group (Figure 3.5). Only a minor fraction (4 out of 201,650) of the targets presented a sequencing bias in one group. Thus, the tissue difference between the starting material used for WES is unlikely to influence the CNV analysis.

Target	1	2	3	4	5
Chromosome	chr7				
Start	142457336	142458406	142459625	142460282	142460719
End	142457375	142458565	142459878	142460418	142460871
P-value Bonferroni corrected	**1.38E-03**	1	**1.09E-06**	**2.39E-02**	**8.89E-05**

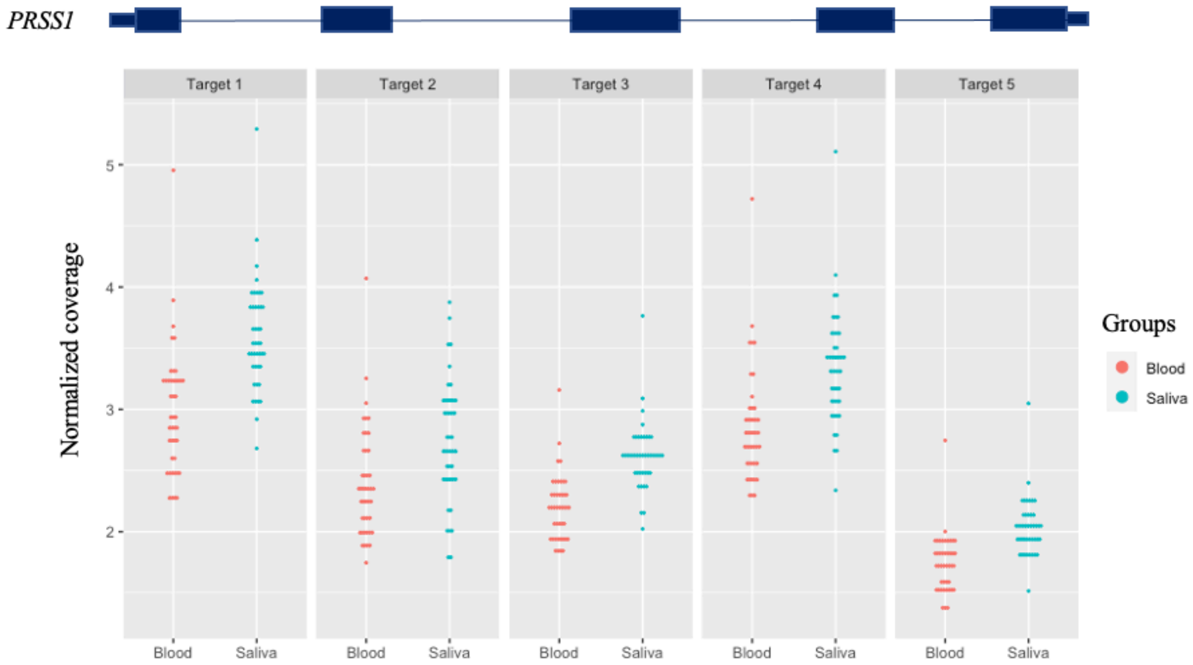
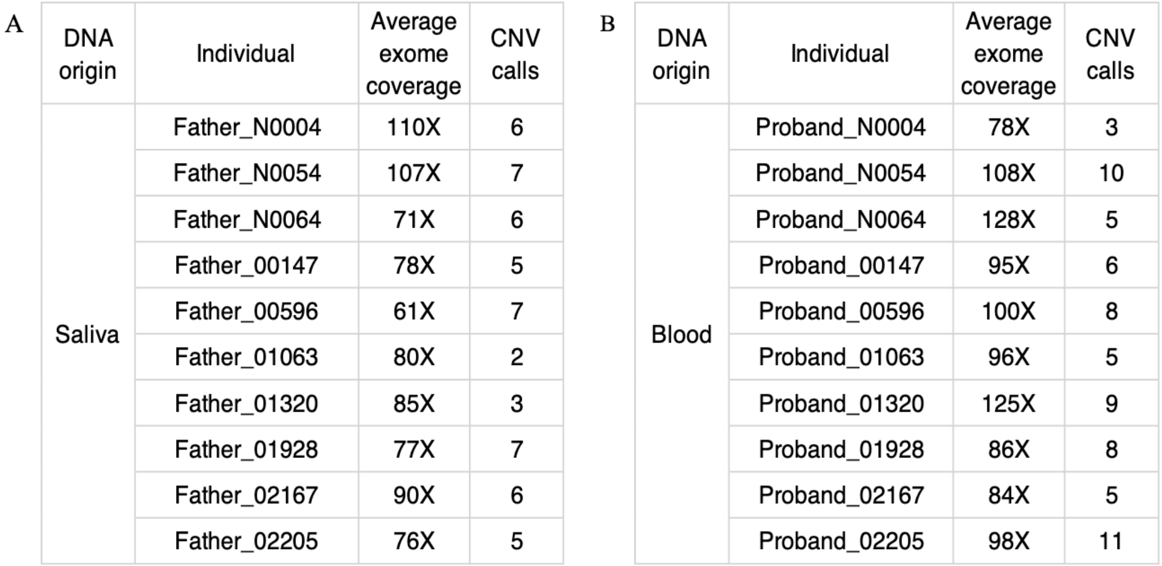


Figure 3.5. Location of the targets on the *PRSS1* gene with the respective Bonferroni corrected p-values (significant values in bold). Below, the normalised target coverage in blood- and saliva-derived DNA samples for each *PRSS1* target. For each target, the wider the segments, the more samples (out of the 43 of the group) had the corresponding normalised coverage.

3.3.4 Influence of the DNA origin on CNV detection, a practical comparison

Based on the results presented in the previous paragraph, we did not expect the DNA origin to influence the CNV calling. To test this hypothesis, I performed the CNV calling with the WES data of 10 samples sequenced from saliva-derived DNA, all using the same WES enrichment kit (Twist kit). These samples consisted of 10 fathers from our trio cohort that were not included as unaffected controls in the CNV detection pipeline. I compared the number of CNVs detected in these fathers with that of the corresponding probands. Table 3.4 shows the number of CNV calls in each sample and the respective average exome coverage.



A

DNA origin	Individual	Average exome coverage	CNV calls
Saliva	Father_N0004	110X	6
	Father_N0054	107X	7
	Father_N0064	71X	6
	Father_00147	78X	5
	Father_00596	61X	7
	Father_01063	80X	2
	Father_01320	85X	3
	Father_01928	77X	7
	Father_02167	90X	6
	Father_02205	76X	5

B

DNA origin	Individual	Average exome coverage	CNV calls
Blood	Proband_N0004	78X	3
	Proband_N0054	108X	10
	Proband_N0064	128X	5
	Proband_00147	95X	6
	Proband_00596	100X	8
	Proband_01063	96X	5
	Proband_01320	125X	9
	Proband_01928	86X	8
	Proband_02167	84X	5
	Proband_02205	98X	11

Table 3.4. **A.** CNV calls and average exome coverage of 10 fathers (saliva-derived DNA samples) that were not used as unaffected controls in the CNV workflow. **B.** CNV calls and average exome coverage of 10 probands (blood-derived DNA samples) from the same trios.

On average, the number of calls per sample in the saliva group was 5, while in the blood group was 7. Respectively, they had an average exome coverage of 84X and 100X. Interestingly, the only two probands with a lower average exome coverage than their respective fathers (Proband_N0004 and Proband_02167) had a lower number of CNV calls as well. Despite the small sample size, the data is consistent with the expectation that the DNA origin does not directly influence the CNV calling from WES data. Instead, as expected, the average exome coverage is one of the main parameters that influences the CNV detection.

3.4 Discussion

In this chapter, I explored the influences of two factors on WES coverage and consequently on CNV detection: the choice of the exome enrichment kit and the tissue used to extract the DNA.

To investigate the consequences of using different exome enrichment kits, I compared the average exome coverage of samples enriched with the Illumina Nextera DNA kit to that of samples enriched with Twist Human Core exome kit. TE samples showed a substantially higher average exome coverage than IE ones, despite being sequenced on the same platform and with the same number of samples per sequencing run. This difference is likely due to the dissimilar designs of the two enrichment kits. Illumina kit targets 12 Mb more than its Twist equivalent, according to the manufacturers. Chilamakuri *et al.* assessed the performance of the Illumina kit and demonstrated that it covers several UTRs (Chilamakuri *et al.*, 2014). It also covers more coding bases since its design is based not only on CCDS regions, as it is for Twist kit, but also on the RefSeq and Ensembl coding content data. However, a published direct comparison of the two kits is not available in the literature. The wider targeted region of the Illumina kit might explain the lower average exome coverage of IE samples as well as the greater number of CNVs identified. On the other hand, Twist kit produced high average exome coverage, and its advertised uniform coverage of the target region is reflected in a lower noise level in the CNV data (Figure 3.3A and 3.3B). Illumina kit might be a suitable choice for exploratory studies that want to investigate the coding region as well as the UTRs and detect a greater number of CNVs, despite a possibly lower specificity due to the high level of noise in the CNV data. The Twist kit appears instead better suited to sequence a smaller fraction of the genome at high coverage and to perform a more reliable CNV analysis. It might be preferentially used in routine clinical research where the interest is focused on variants in known genes since the manufacturer states that 99% of ClinVar variants are covered with this kit. Notwithstanding these differences, both the enrichment kits were able to produce an average exome coverage much higher than 30X, which is conventionally considered the minimum acceptable coverage, and were both suitable for CNV analysis

The second factor influencing the average exome coverage was the tissue of origin of the DNA used for WES. We used the WES data from our trio cohort to explore this aspect. The WES data of the probands, sequenced from blood-derived DNA, showed a significantly higher average exome coverage compared to the WES data of the parents, sequenced from saliva-derived DNA. A significant difference was found for both the exome enrichment kits. One hypothesis is that the lower average exome coverage of samples sequenced from saliva-derived DNA was caused by a different quality of the DNA extracted from this tissue (Kidd *et al.*, 2014). Residual substances in saliva, such as food or tobacco, might have disturbed the DNA extraction. More importantly, the microbial DNA possibly present in saliva (Quinque *et al.*, 2006; Lim *et al.*, 2017) might have been extracted and sequenced in a small percentage with the human DNA, even though it should not be enriched using the exome enrichment kits. The sequencing reads derived from bacterial DNA would remain unmapped against the human reference and therefore excluded, reducing the overall average exome coverage of the samples. Finally, the more laborious procedure for saliva samples collection that required the individuals participating in the study to ship their saliva samples to our laboratory might have increased the DNA degradation in these samples. Nevertheless, the samples sequenced from saliva DNA were able to reach an average exome coverage above 30X, and the vast majority (91%) of those enriched with the Twist enrichment kit had an average exome coverage above 60X, which is commonly considered an optimal coverage. These results confirmed what reported by different studies that showed an acceptable quality of the WES data generated from saliva-derived DNA (Kidd *et al.*, 2014; Zhu *et al.*, 2015).

These studies, however, did not explore what consequences the use of saliva-derived DNA for WES might have on CNV detection. To investigate this aspect, we compared the coverage of each sequencing target of the Twist exome enrichment kit between 43 samples sequenced from saliva-derived DNA and 43 samples sequenced from blood-derived DNA and looked for systematic variation in coverage for individual genes.

First, we established that a very small percentage of targets (<1%) had an average coverage >=10X in one group but <10X in the other. The number of these targets was different in the blood (106 targets) and saliva (6 targets) groups. Almost all of them were distributed across different genes and did not affect large contiguous coding regions. This suggests that the difference between the groups is not systematic for specific genes and would have a minimal impact on the variant analysis. The different number of targets identified in the two groups might be the consequence of their different average exome coverage.

Secondly, we identified 4 targets (< 0.01% of the total) that were systematically sequenced at higher coverage in saliva-derived DNA. These probes targeted 4 exons of the *PRSS1* gene (5 exons in total). This gene encodes a trypsinogen produced by the pancreas that has an active role in digestion (Koshikawa *et al.*, 1994; Teich *et al.*, 2006). Its higher coverage in saliva-derived DNA might be due to a somatic difference in the DNA. For instance, this gene might be more active in saliva than blood, and therefore its locus might have been more accessible and easier to sequence in saliva-derived DNA. Overall, the very small fraction of targets showing a systematic difference in coverage in samples sequenced from blood- and saliva-derived DNA indicates that using both these groups of samples for CNV studies is unlikely to bias the results.

Lastly, to test this expectation, I performed CNV calling in 10 samples sequenced from saliva-derived DNA, corresponding to 10 fathers from the trio cohort that were not included as unaffected controls in the CNV workflow. I compared the number of CNVs called in these samples to that detected in the respective probands, which were sequenced from blood-derived DNA. The lower average number of calls per sample identified in the saliva group is likely the consequence of the lower average exome coverage, rather than a systematic difference caused by the DNA origin. Two probands had lower average exome coverage than their respective fathers. These probands were the only ones with fewer CNVs than their parents. The cohort size used for this test was limited, but the data suggests that the average exome coverage is the only direct influence on CNV calling. Since saliva-derived DNA produces WES data with generally lower coverage than blood-derived DNA, samples extracted from this tissue could be sequenced in sequencing runs with a lower number of total samples. This might allow reaching a higher average exome coverage with a slight increase in cost per sample. Moreover, for every candidate *de novo* CNV identified in the probands, it is necessary to independently verify the absence of the variant in the parents with a separate test, as their lower average exome coverage might prevent its detection from the WES data. Nevertheless, this study demonstrates that saliva-derived DNA generated WES data suitable for CNV analysis and suggests that samples with the same average exome coverage are comparable independently of the tissue used for DNA extraction.

3.5 Conclusions

In this chapter, I instigated how using different exome enrichment kits and DNA extracted from different tissues influences WES's performance. Specifically, I explored the consequences on WES coverage, the metric on which CNV detection from WES data is based. I described how the different designs of two exome enrichment kits from Illumina and Twist Bioscience companies influence the WES coverage and their advantages and drawbacks on CNV calling. I also confirmed that performing WES with saliva-derived DNA is a valid method, as previously reported in the literature. Finally, I demonstrated that using saliva-derived DNA for WES does not produce systematic coverage bias in multiple genes. However, it produces WES data with lower average exome coverage than using blood-derived DNA in the same sequencing conditions.

Chapter 4. *De Novo* CNVs in Idiopathic Quantitative Forms of Male Infertility

4.1 Introduction

Male infertility is a frequent condition that affects ~7% of men (Krausz and Riera-Escamilla, 2018). Genetic factors can cause the most severe forms of isolated male infertility; however, a large percentage of the cases often remains unexplained (Tüttelmann, Ruckert and Röpke, 2018). Genetic mutations causing male infertility undergo negative selection, but at the same time, male infertility is still present at high frequency in the outbred human population (Krausz and Riera-Escamilla, 2018). In this population, we do not expect all idiopathic cases to be recessively inherited. Concurrently, the most severe forms of infertility cannot be passed to the offspring from the father in a dominant fashion. In other fitness-impairing diseases, the role of *de novo* mutations is prominent (Acuna-Hidalgo, Veltman and Hoischen, 2016), and we hypothesise a similar scenario for male infertility. Evidence supporting this hypothesis comes from the two most frequent genetic causes of severe male infertility: chromosome Y deletions and an additional X chromosome causing Klinefelter syndrome, which both occur *de novo*. In large cohorts of patients, these genetic variants explain up to 5.4% and 3.5% of the cases, respectively (Olesen *et al.*, 2017; Punab *et al.*, 2017; Tüttelmann, Ruckert and Röpke, 2018). Besides these two causes, the contribution of *de novo* variants to the aetiology of the disease remains unknown to date. Only one recent pilot study investigated *de novo* point mutations in 13 azoospermia patient-parent trios (Hodžić *et al.*, 2020) and identified 5 candidate disease genes acting in a dominant manner. However, the contribution of *de novo* CNVs has not been explored. In this chapter, I present the results of sequencing the exome of 183 azoospermia and severe oligozoospermia patient-parent trios and using the G4BP CNV detection tool to identify *de novo* CNVs.

4.1.1 De novo CNVs in the general population and Mendelian diseases

De novo CNVs are rare in the general population (Acuna-Hidalgo, Veltman and Hoischen, 2016). However, a recent population study revealed that the ability to detect this type of variant is often influenced by the method used and the sample size (Collins *et al.*, 2020). In 2007, a microarray-based study estimated the rate of *de novo* CNVs to be 0.01 per generation in the general population and up to 0.1 for a cohort of patients with autism (Sebat *et al.*, 2007). Later in 2015, a WGS-based study proposed a rate of 0.0154 events per generation for CNV larger than 100kb (Kloosterman *et al.*, 2015). This study was the only one that specified a *de novo* CNV rate in relation to the variants size. Recently, the rate of *de novo* structural variations, including deletion, duplications, insertions, inversions, translocations and complex variants, has been estimated by investigating a large dataset of WGS data (Belyeu *et al.*, 2021). The authors of this paper studied 33 large multigeneration families as well as a large ASD cohort, including 2000+ autism patient-parent trios and quartets. They identified *de novo* structural variations rates of 0.160 and 0.206, respectively, for unaffected and autism families. This suggests that the rate of *de novo* CNVs might be higher than previously anticipated and confirmed that cohorts of patients can be enriched for such variants. Thus, we might expect a comparable situation in our male infertility trio cohort if *de novo* CNVs occurring outside the chromosome Y are an actual cause of the disease. However, since WGS has much higher sensitivity than WES, we expect to detect fewer *de novo* CNVs in our data compared to the study by Belyeu *et al.*

4.1.2 Parent of origin of de novo variants

De novo point mutations are transmitted preferentially through the paternal side (approximately 80%) (Goldmann *et al.*, 2016; Jónsson *et al.*, 2017). This is likely due to the larger number of rounds of DNA replication and consequent errors occurring during male gametogenesis. Similarly, studies on *de novo* CNVs revealed an analogous paternal bias (Hehir-Kwa *et al.*, 2011; Ma *et al.*, 2017). In male infertility, pathogenic *de novo* variants affecting spermatogenesis genes may undergo negative selection in the male germline. Therefore, in male infertility patients, we might find more *de novo* variants transmitted from the maternal side than generally expected (~ 20%). To test this hypothesis, in this chapter, I determine (where possible) the parent of origin of the *de novo* CNVs identified in our cohort.

4.2 Aims

To investigate the role of *de novo* CNVs in male infertility, I:

- Identified *de novo* CNVs from WES data of 183 patient-parent trios.
- Evaluated the pathogenicity of the *de novo* CNVs detected.
- Determined their parent of origin.
- Identified novel candidate dominant and X-linked male infertility genes.

4.3 Results

WES data from 183 patients affected by azoospermia or severe oligozoospermia and their parents were analysed with the G4BP CNV tool to identify deletions and duplications. A total of 2054 CNVs was detected in the probands, with an average of 11 CNVs identified per sample. Common variants were excluded, and only rare CNVs (present in <1% of the samples of the DGV Gold Standard) were retained for further analysis. The duplications largely exceeded the number of deletions for both the total and rare CNVs (Figure 4.1). The plots of rare CNVs generated by the G4BP tool were visually inspected, and deletions and duplications present in a proband but absent in the respective parents were selected. Two *de novo* deletions were identified in two different probands (Table 4.1).

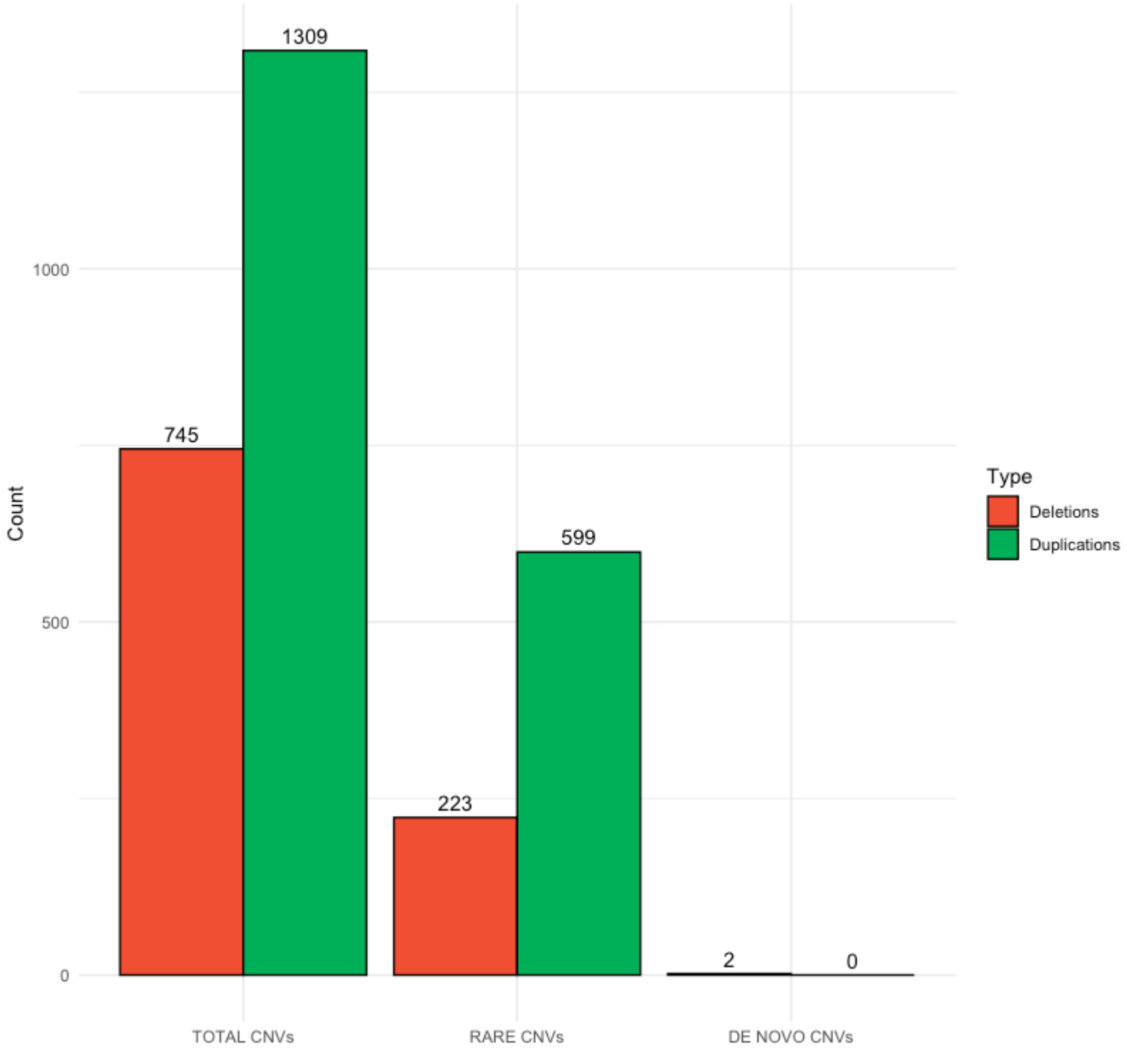


Figure 4.1. Number of CNVs identified in 183 probands with azoospermia or severe oligozoospermia. The number of duplications exceeded the number of deletions for both the total and rare CNVs. The *de novo* CNVs were deletions exclusively.

Proband	Genomic location (GRCh37)	Size	Probes	Genes	Genes with a pLI score > 0.9
953	chr11:32975325-33631588	656 kb	75	*CSTF3 - CSTF3-AS1 - DEPDC7 - HIPK3 - KIAA1549L - LINC00294 - QSER1 - TCP11L1*	*CSTF3 - QSER1*
584	chrX:108779111-108785995	7 kb	4	*NXT2*	-

Table 4.1. Proband, genomic coordinates, size, and genes involved in the two *de novo* deletions identified.

4.3.1 De novo deletions

The first *de novo* deletion was identified on chromosome 11 of proband 953. The chromosome view plot (Figure 4.2) and the CNV plot (Figure 4.3) clearly show the absence of this event in both the mother and the father of the same trio. The CNV deleted a 656 kb region and removed one copy of 8 genes (Table 4.1 – Figure 4.3). Figure 4.3 represents the location of the CNV in detail, and the MAF track shows the absence of heterozygous SNPs (MAF = 0.5) within the deleted region. This is consistent with the absence of one allele. This region has not been reported as deleted in the DGV Gold Standard nor in the GnomAD-SV database. Two genes encompassed by the deletion have a pLI score > 0.9 and are therefore likely to show haploinsufficiency: *CSTF3* (pLI = 0.98) and *QSER1* (pLI = 1). Deletions of the coding regions of these two genes have not been reported in the samples of the two population databases mentioned. *QSER1* has medium expression in the testis, according to the Protein Atlas database, and was recently recognised as a DNA methylation regulator (Dixon *et al.*, 2021). *CSTF3* is highly expressed in the testis, both at the RNA and protein level, and it is involved in the pre-mRNA processing (Grozdanov *et al.*, 2018). The literature did not reveal any link between the function of the other genes in the deleted region and the phenotype of the patient.

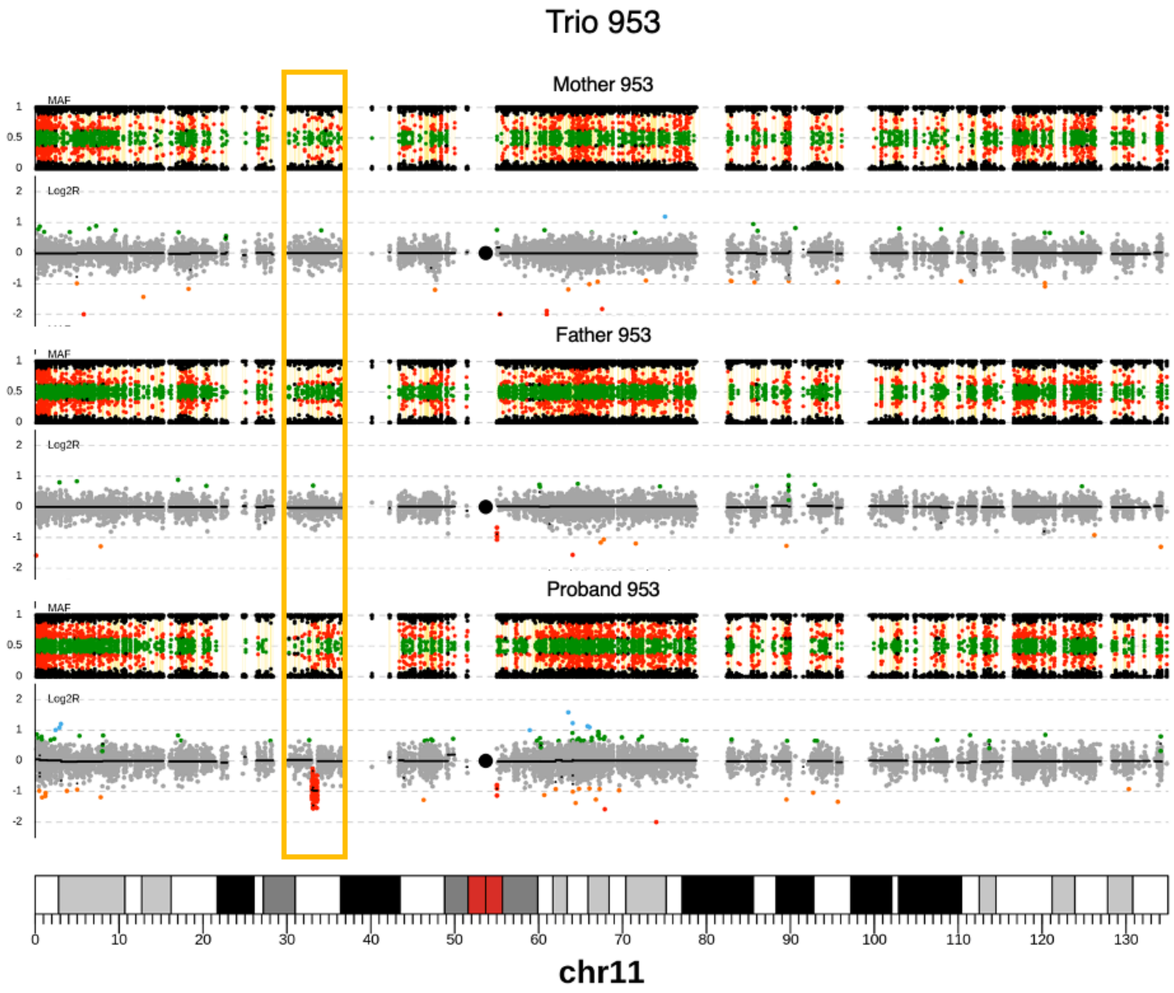


Figure 4.2. Chromosome-view plot of chromosome 11 in trio 953. In the orange box, the *de novo* deletion is highlighted as well as the correspondent parental genomic regions.

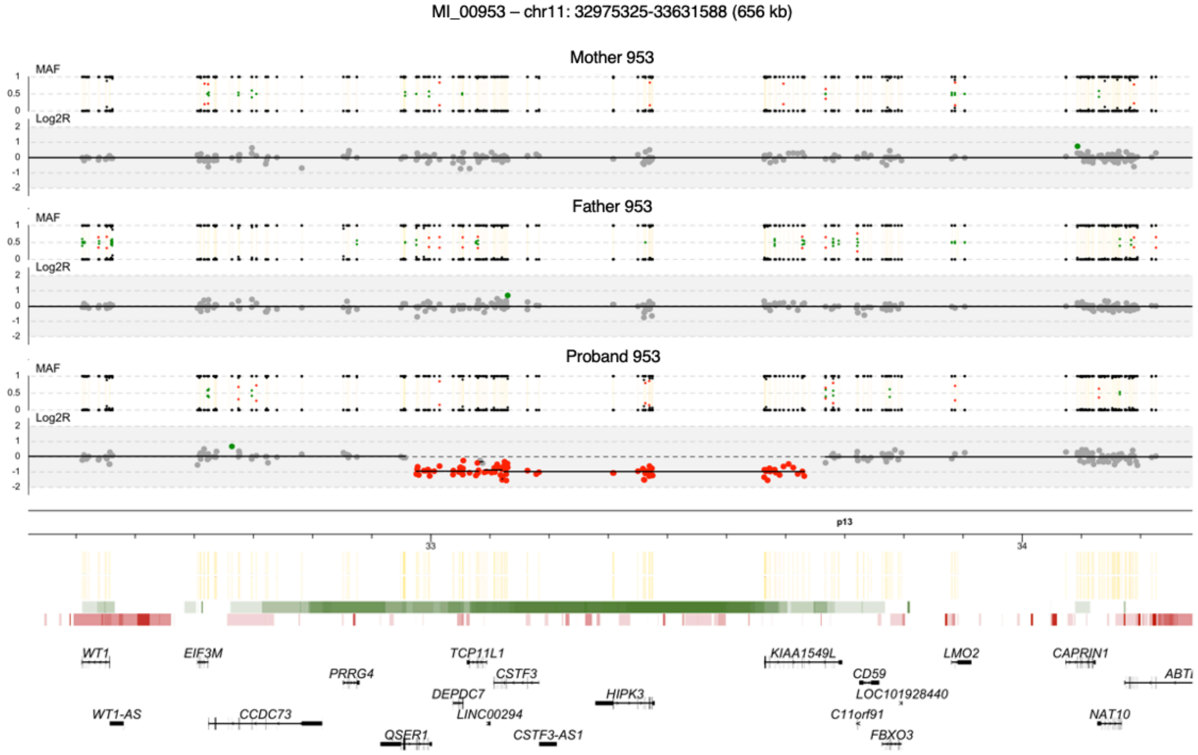


Figure 4.3. CNV plot of the *de novo* deletion identified in proband 953. The deletion removed a single copy of 8 genes. The MAF track shows absence of heterozygous SNPs (MAF = 0.5) within the deleted region in the proband, which suggests loss of heterozygosity.

The SNP data from proband 953 was investigated to determine the parent of origin of the *de novo* CNV. 9 exonic informative SNPs in the hemizygous region, sequenced at >10-fold coverage, revealed that the retained allele was maternally inherited. In contrast, the corresponding paternal SNPs were absent (Table 4.2). This data indicated that the *de novo* deletion arose on the paternal allele.

The *de novo* deletion was validated with a microarray assay carried out at Northern Genetics Service (see methods in Chapter 2). The results confirmed the presence of the deletion in proband 953 and its absence in the parents. It also confirmed that the deletion had a paternal origin and estimated a slightly larger size of 688kb. The genes involved were the same 8 identified from the WES data.

Genomic location (GRCh37)	Proband SNP	Mother SNP	Father SNP	Gene	Region	Number of sequencing reads in proband	Paternal/Maternal SNP in proband
chr11:32704946	A/C	A/C	C	*CCDC73*	Intron 7	22	Both
chr11:33053107	T	T	C	*DEPDC7*	Exon 5	13	Maternal
chr11:33054794	G	G	T	*DEPDC7*	Intron 8	18	Maternal
chr11:33065394	T	T	C	*TCP11L1*	Exon 2	36	Maternal
chr11:33078643	C	C	T	*TCP11L1*	Intron 3	17	Maternal
chr11:33080496	C	C	T	*TCP11L1*	Intron 5	12	Maternal
chr11:33097876	T	T/C	C	*LINC00294*	Exon 1	16	Maternal
chr11:33106616	T	T	C	*CSTF3*	Exon 21	31	Maternal
chr11:33129997	T	T	C	*CSTF3*	Intro 3	13	Maternal
chr11:33211922	G	G	A	*CSTF3-AS1*	Exon 4	7	Maternal
chr11:33564123	T	T	C	*KIAA1549L*	Exon 2	32	Maternal
chr11:33667333	A/G	A/G	G	*KIAA1549L*	Exon 17	40	Both

Table 4.2. 11 informative SNPs at different genomic coordinates within (in orange) and outside (no colour) the hemizygous region on chromosome 11 of proband 953. The SNPs in the hemizygous region are maternally inherited exclusively, while the paternal ones are absent.

The second *de novo* deletion was identified on chromosome X of proband 584. The chromosome view (Figure 4.4) and the CNV plots (Figure 4.5) show the hemizygous deletion in the proband and one and two alleles within the same genomic region, respectively, in the father and mother of the same trio. The CNV was 7 kb large and removed the unique copy of the gene *NXT2*. No CNVs in this region are listed in the DGV Gold Standard or the GnomAD-SV database. According to the Protein Atlas database, *NXT2* has high RNA expression levels in the testis, and it is highly conserved amongst eutherians (Khan *et al.*, 2018). It is involved in the mRNA nuclear export (Herold *et al.*, 2000).

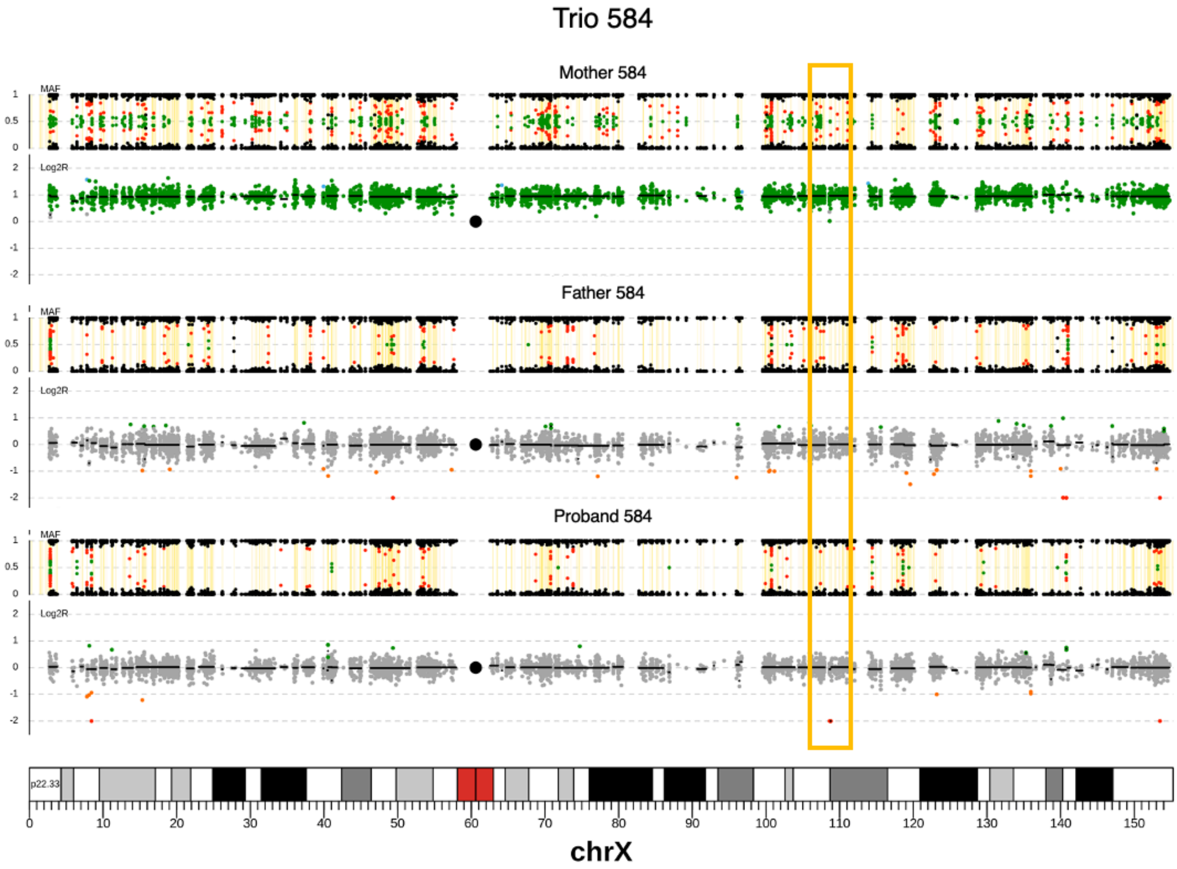


Figure 4.4. Chromosome-view plot of chromosomes X in trio 584. In the orange box, the *de novo* deletion is highlighted as well as the correspondent parental genomic regions.

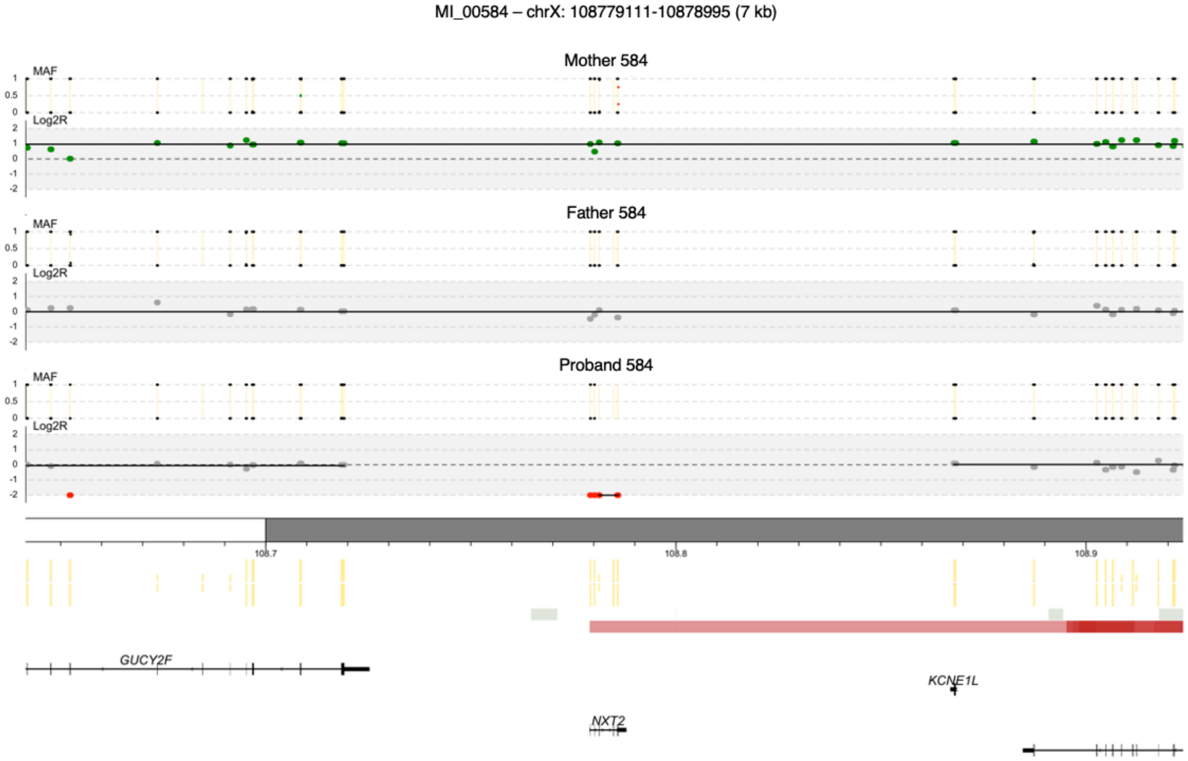


Figure 4.5. CNV plot of the *de novo* deletion identified in proband 584. The only copy of the *NXT2* gene was deleted and no other genes were involved in the CNV.

The MAF track in Figure 4.5 shows the presence of SNPs within the deleted region, a scenario incompatible with the hemizygous deletion detected. Examination of the sequencing reads from the binary alignment map (BAM) file of the sample, performed with the IGV visualisation tool, showed that only reads with low-quality mapping scores were mapped against that region. The imperfect segmentation of the first two probes of the deletion might have caused the retention of those reads and the erroneous representation of the SNPs in the plot.

The *de novo* deletion in proband 584 must have originated on the maternally inherited X chromosome since fathers do not transmit the X chromosome to male offspring. A qPCR assay, performed by Dr Alobaidi, confirmed the complete deletion of *NXT2* in the proband and the expected number of copies in his parents. (i.e., 2 in the mother and 1 in the father).

4.3.2 A mosaic deletion in a father

An interesting discovery was made in the WES data of father 1385. The same region on chromosome 11 *de novo* deleted in proband 953 showed an ambiguous segmentation of a large segment with Log2R between 0 and 1 (Figure 4.6). The segmentation did not represent a deletion (Log2R = -1), but it did not represent the signal expected for two alleles either (Log2R = 0). We commissioned a microarray assay for this sample at the Northern Genetics Service to investigate this region further. The results revealed a 3.1 Mb large mosaic deletion. This deletion was flanked by smaller mosaic deletions respectively 205 kb and 192 kb large (Table 4.3 – Figure 4.5). The largest mosaic deletion was estimated to be present in 50/60% of the cells, while the two smaller ones only in 10/30%. The CNV event completely overlapped with the *de novo* deletion in proband 953 and encompasses all the deleted genes (Figure 4.6). The segmentation of the same region in proband 1385 represented two alleles, and CNVs at that locus were not identified (Figure 4.6). Thus, the mosaic deletions were not transmitted to the proband and cannot be the cause of his infertility.

Individual	Genomic location (GRCh37)	Size	Genes	Percentage of cells estimated to carry the deletion
Father 1385	chr11:32938165-33143126	205 kb	*CSTF3 - DEPDC7 - LINC00294 - QSER1 - TCP11L1*	10-30%
	chr11:33156954-36272348	3.1 Mb	*ABTB2 - APIP - C11orf91 - CAPRIN1 - CAT - CD44 - CD59 - CSTF3 - CSTF3-AS1 - EHF - ELF5 - FBXO3 - FBXO3-AS1 - FJX1 - HIPK3 - KIAA1549L - LDLRAD3 - LMO2 - LOC100507144 - LOC101928510 - MIR1343 - MIR3973 - NAT10 - PAMR1 - PDHX - SLC1A2 - SNORD164 - TRIM44*	50-60%
	chr11:36273754-36466141	192 kb	*COMMD9 - PRR5L*	10-30%

Table 4.3. Coordinates, genes involved, and percentage of cells estimated to carry the mosaic deletions identified in father 1385.

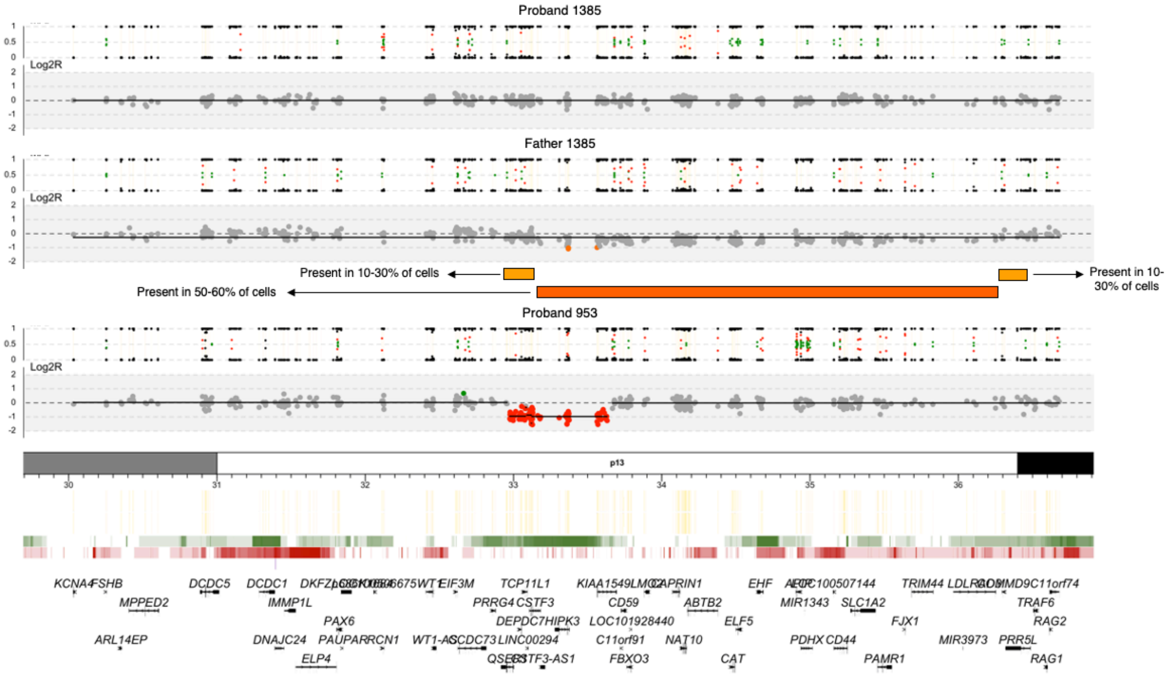


Figure 4.6. Mosaic deletions detected in father 1385 compared to the *de novo* deletion identified in proband 953. The region involved in the mosaic deletions completely overlap with the region *de novo* deleted. The CNV plot of proband 1385 shows the presence of 2 alleles in the same region, suggesting that the mosaic deletions in father 1385 were not transmitted to the proband.

4.4 Discussion

The study presented in this chapter is the first to investigate the role of *de novo* CNVs outside the AZF regions. The analysis of WES data of 183 azoospermia and severe oligozoospermia patients revealed 2054 CNVs in total. Duplications were almost twice the number of deletions for both the total and rare CNVs. Algorithms for CNV detection from WES data usually identify more duplications than deletions (Guo *et al.*, 2013). This bias could indicate the presence of false positives, which are often called in regions rich in segmental duplications and highly polymorphic loci (Rajagopalan *et al.*, 2020).

The first *de novo* deletion removed one copy of 8 genes located on chromosome 11 of proband 953. The deletion has not been reported in the databases examined. *QSER1* and *CSTF3* were the only constrained genes (pLI > 0.9) affected, and likely pathogenic SNVs were not identified on the remaining allele. *CSTF3* has high expression in the testis and plays a role in pre-mRNA processing (Grozdanov *et al.*, 2018). My colleagues Dr Manon Oud and Hannah Smith, who performed the *de novo* SNV analysis on these samples, found a similar function for several candidate disease genes that carried a *de novo* likely pathogenic SNV, such as *U2AF2*, *CDC5L* and *RBM5* (Oud *et al.*, 2021). *RBM5* is also an essential regulator of pre-mRNA splicing in mice germ cells (O'Bryan *et al.*, 2013). Alternative splicing in the testis is complex and occur at a lower level only in the brain (Song *et al.*, 2020). This process is likely one of the critical regulator mechanisms of testis development and spermatogenesis in mammals (Song *et al.*, 2020). *CSTF3* is, therefore, a compelling novel candidate male infertility gene. However, we cannot exclude that the loss of a copy of the other genes involved in the deletion might also contribute to the origin of the disease. In 2014, a partially overlapping deletion was reported in a patient exhibiting cryptorchidism and azoospermia by Seabra *et al.* (Seabra *et al.*, 2014). This 1 Mb large heterozygous deletion removed 7 genes encompassed by the *de novo* CNV identified in our study (Figure 4.7). In that case, though, parental samples were not available, and the inheritance remained unknown. The authors of the article suggested that the patient's phenotype was caused by the haploinsufficiency of *WT1*, a gene associated with male infertility (Seabra *et al.*, 2015; Xu *et al.*, 2017). *WT1* is not deleted in our patient, but genes such as *QSER1* and *CSTF3* are in common, and it is possible that their haploinsufficiency also contributed to the phenotype. Given the presence of two Alu Y repeats with high homology in the genomic region where the CNV was identified, Seabra *et al.* proposed a repeat-mediated nonallelic homologous recombination as the likely mechanism that generated the deletion. The same mechanism might have caused the *de novo* deletion found in proband 953 and the mosaic events in father 1385. However, a better characterisation of the genomic region aimed to identify repeated regions and inversions should be performed to verify this hypothesis.

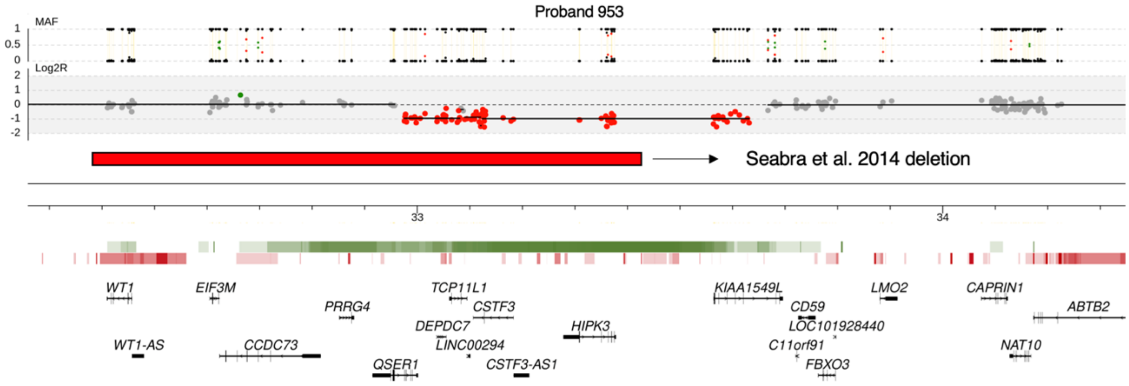


Figure 4.7. Comparison between the *de novo* deletion identified in proband 953 and the deletion reported in Seabra *et al.* 2014. 7 genes are in common between the two CNVs.

The 3 mosaic deletions identified in father 1385 affecting the same region on chromosome 11 were estimated to affect a different percentage of cells, with the largest region deleted in 50-60% of the cells. These mosaic deletions might underlie a possibly more complex rearrangement in father 1385 that cannot be fully characterised with WES or microarrays. The deletions were not transmitted to the proband and therefore cannot be the cause of his infertility. Interestingly, father 1385's partner has a history of two miscarriages. It is pure speculation, but the transmission of these large deletions to the offspring, in which they would become germline events, might have been the underlying cause of these unsuccessful pregnancies.

The second *de novo* deletion was reported on chromosome X of proband 584. It removed the unique copy of the *NXT2* gene. Deletion of this gene has not been reported in either the DGV Gold Standard or the GnomAD-SV database. *NXT2* has a medium level of protein expression and a high RNA expression in the testis, according to the Protein Atlas database. It participates in mRNA nuclear export (Herold *et al.*, 2000) and might be part of the complex alternative splicing mechanism present in the testis that was mentioned above. *NXT2* is evolutionarily conserved in eutherians but has been reported as dispensable for fertility in mice (Khan *et al.*, 2018). In that species, however, this gene's expression is not testis-enriched (Khan *et al.*, 2018), contrarily to what has been found in humans. We propose both the two *de novo* deletions identified as likely pathogenic and possibly causative of the male infertility phenotype in the respective patients. Further functional studies and screening larger cohorts of patients will clarify the frequency of these variants and the specific pathways in which the deleted genes participate.

In the cohort of 183 patient-parent trios, only two *de novo* CNVs were identified, resulting in a *de novo* CNV frequency of 0.01. This frequency is similar to one previously observed in the general population (i.e., 0.01) and lower than what found in patients with ID or autism (i.e., 0.1) (Sebat *et al.*, 2007; Kloosterman *et al.*, 2015; Vissers, Gilissen and Veltman, 2016). The large difference in the *de novo* CNV frequency between male infertility and intellectual disability patients may be partially explained by the use of WES for CNV detection, which might not have detected all the *de novo* CNVs in our cohort. Compared to ID, where large CNVs affecting multiple genes cause the disease, it is possible that, in isolated severe male infertility, the deletion of (part of) a single gene is sufficient to cause infertility without resulting in additional phenotypes. Thus, this cohort might be enriched with small *de novo* variants with a size below the minimum resolution of CNV detection from WES data (usually estimated as at least 3 exons) (Lelieveld *et al.*, 2016). It is also important to remember that patients with *de novo* CNVs on the Y chromosome were excluded from our study; therefore, our *de novo* CNV rate for male infertility patients is likely underestimated. Lastly, it is possible, of course, that *de novo* CNVs do not contribute to the origin of male infertility and ID equally. Our research group is now performing WGS for the entire trio cohort. This data is likely to reveal previously undetected *de novo* CNVs and provide a more accurate estimate of the *de novo* CNV rate in these patients.

Finally, we identified the parent of origin for both the *de novo* CNVs identified. The deletion on chromosome 11 arose on the paternal allele, while the one on chromosome X must be transmitted from the mother, who is the only parent that passes an X chromosome to male offspring. My colleague Dr Giles Holt worked on phasing the *de novo* SNVs identified in this cohort of trios by our colleagues (Oud *et al.*, 2021). He was able to phase 29% of the *de novo* point mutations found. The results revealed that 72% of the total and 75% of those classified as likely pathogenic occurred on the paternal allele. The cause of this paternal bias is often attributed to the higher rounds of replication occurring in the male germline, and similar percentages (~80%) were previously reported in the literature (Goldmann *et al.*, 2016; Jónsson *et al.*, 2017). If these likely pathogenic variants impair reproductive fitness, we wondered how they could escape negative selection in the paternal germline. We propose three possible scenarios to explain this phenomenon: (1) the *de novo* variant occurs following the temporal window in which the fertility gene is active; (2) the *de novo* variant impairs a gene that does not actively participate in spermatogenesis, but it is essential for supporting the future germline in the offspring; (3) it has been observed that the cells originating from the same spermatogonial cell are able to form cysts and share mRNA and proteins (Dym and Fawcett, 1971; Braun *et al.*, 1989; Sharma and Agarwal, 2011). This could mask the deleterious effect of a *de novo* variant that occurred in one of these cells. These 3 mechanisms might explain why these likely pathogenic SNVs and CNVs also occurred on the paternal allele and were passed to the next generation via fully functioning sperm, resulting in infertile offspring. An interesting mechanism that increases the number of pathogenic *de novo* variants that are passed to the next generation from the paternal side was identified in 2003. Goriely *et al.* described pathogenic mutations in the *FGFR2* gene that conferred a selective advantage to the spermatogonial cells where they arose but were deleterious for the embryonic development (Goriely et al., 2003). Although this mechanism allows pathogenic *de novo* variants of paternal origin to be passed to the offspring, we consider this unlikely to happen for *de novo* variants that cause male infertility, since these variants would confer a selective advantage to spermatogonial cells and, at the same time, impair spermatogenesis in the male offspring. Further investigations are required to reveal new insight on natural selection and timing of *de novo* variants affecting male fertility.

4.5 Conclusions

In this chapter, I investigated the role of *de novo* CNVs in severe idiopathic male infertility, using the WES data of 183 patient-parent trios. Two likely pathogenic *de novo* CNVs were identified in two patients, one on the paternal allele and one on the maternally inherited chromosome X. These variants highlighted 3 novel male infertility candidate genes and suggest for the first time that *de novo* CNVs outside chromosome Y can contribute to the aetiology of the disease.

Chapter 5. Maternally Inherited CNVs in Idiopathic Quantitative Forms of Male Infertility

5.1 Introduction

In the previous chapter, I investigated the role of *de novo* CNVs in 183 patient-parent trios affected by azoospermia and severe oligozoospermia. The same cohort is further analysed in this chapter to explore the possible contribution of maternally inherited CNVs to the aetiology of the disease. Inherited CNVs, transmitted through the maternal lineage, might affect spermatogenesis genes and consequently impair the reproductive fitness of the patients without affecting female fertility.

In the previous century, several papers described maternally inherited translocations (Chandley *et al.*, 1972; CHANDLEY *et al.*, 1975; Debiec-Rychter *et al.*, 1992) and other chromosomal abnormalities (Smith *et al.*, 1965) associated with impaired fertility in males but not in females. More recently, Sazci *et al.* reported a case of a translocation in a male with primary infertility that was transmitted from his fertile mother (Sazci *et al.*, 2005). These studies did not explore the contribution of single genes because of the limited resolution of the genomic techniques available at the time. Nevertheless, they highlighted the need for performing karyotyping in infertile couples as it was thought that these large rearrangements would not allow regular meiosis during spermatogenesis. A thesis that was subsequently supported by other investigations (Sun *et al.*, 2007).

In the last 10 years, multiple studies aimed to identify disease causing variants on the X chromosome (Krausz *et al.*, 2012; Chianese *et al.*, 2014; Lo Giacco *et al.*, 2014; Yatsenko *et al.*, 2015). Chromosome X is of particular interest since it is enriched in genes expressed during spermatogenesis (Vockel *et al.*, 2019). Also, this chromosome has been singled out largely because it is maternally inherited in men. An additional copy of chromosome X in males causes Klinefelter's syndrome (XXY), a main *de novo* cause of azoospermia (Krausz and Riera-Escamilla, 2018). In addition, 4 X-linked genes (*ADGRG2*, *AR*, *NR0B1*, *TEX11*) are considered disease-causing when mutated in males (Oud *et al.*, 2019). In 2015, Yatsenko *et al.* identified a pathogenic hemizygous deletion involving three exons of *TEX11* in 2 patients with non-obstructive azoospermia. Subsequent screening of the gene in a separate cohort of patients with the same phenotype revealed additional disrupting point mutations that were absent in normozoospermic controls (Yatsenko *et al.*, 2015). Disruptive *ADGRG2* mutations lead to obstructive azoospermia since the gene is essential for the normal formation of the vas deferens (Patat *et al.*, 2016). Pathogenic mutations in the *AR* and *NR0B1* genes cause androgen insensitivity syndrome (Ferlin *et al.*, 2006) and early-onset adrenal hypoplasia congenital with hypogonadotropic hypogonadism at puberty (Suntharalingham *et al.*, 2015), respectively. Both these conditions lead to endocrine disorders with severe male infertility as a consequence. Remarkably, amongst these 4 X-linked disease genes, *TEX11* is the only gene known to cause isolated non-obstructive azoospermia and the first identified carrying a pathogenic CNV (Yatsenko *et al.*, 2015).

On the other hand, a few studies have compared the CNV burden in infertile patients and controls to reveal patient-specific variants on the autosomes (Frank Tüttelmann *et al.*, 2011; Stouffs *et al.*, 2012). However, the unknown inheritance of these variants makes the prioritisation of likely pathogenic CNVs challenging. Now that we have the availability of parental samples, we can explore the role of maternally inherited CNVs on the autosomes for the first time in a systematic manner and test whether X-linked likely pathogenic variants occur *de novo* or are maternally inherited. To date, there are no pathogenic maternally inherited CNVs associated with isolated severe male infertility. In this chapter, I present an exploratory study investigating maternally inherited CNVs in 183 male infertility trios. Contrarily to *de novo* CNVs, which are extremely rare, we expect that approximately half of the CNVs identified in each proband will be maternally inherited. Therefore, prioritising those that are likely pathogenic is a challenging task. This study explores the differences between maternally and paternally inherited CNVs and identifies rare maternally inherited CNVs that might play a role in male infertility.

5.2 Aims

To investigate the role of maternally inherited CNVs in male infertility, I aimed to:

- Assess the differences in number and size between autosomal maternally and paternally inherited CNVs identified in a cohort of 183 patient-parent trios.
- Evaluate the likely pathogenicity of rare maternally inherited CNVs.
- Identify novel candidate dominant and X-linked male infertility genes.

5.3 Results

In this chapter, I used the cohort of 183 patient-parent trios affected by azoospermia and severe oligozoospermia to investigate the role of maternally inherited CNVs in severe isolated male infertility. The G4BP CNV detection tool identified a total of 2054 CNVs in the probands, with an average of 11 CNVs per sample. First, to compare maternally inherited (MI) and paternally inherited (PI) CNVs, I excluded the *de novo* CNVs reported in the previous chapter. Then, visual inspection of all the CNV plots allowed the exclusion of the CNVs with unclear inheritance and possible technical artefacts. This group included: CNVs poorly segmented in the probands and therefore possible false positives; homozygous CNVs; and variants for which inheritance could have been either maternal or paternal. I also excluded the variants identified on the sex chromosomes. The excluded CNVs constituted 53% of the total (i.e., 1099 CNVs). Most (62%) of these CNVs comprised only 3-4 sequencing targets, and 89% encompassed < 10 targets. The remaining 931 CNVs constituted a high confidence dataset of inherited autosomal CNVs. Of these, 439 were PI, and 492 were MI.

5.3.1 Maternally inherited CNVs versus paternally inherited CNVs

Initially, I wanted to test whether there was a different number of rare (present in <1% of the samples of the DGV Gold Standard) large (encompassing at least 10 sequencing targets) MI and PI CNVs in severe male infertility patients. To select large CNVs, the number of targets encompassed was preferred to the CNV size in kilobases. Since one sequencing target corresponds to a single exon, this parameter better indicates the amount of coding region involved in the CNV event. CNVs that are rare in the general population and encompass a large portion of the coding region are often associated with disease phenotypes (Li *et al.*, 2020). I compared the number of MI and PI CNVs for different CNV categories: the number of total deletions and duplications, the number of rare ones and the number of those rare and large (Figure 5.1). The total number of MI CNVs was 492, 53 more than PI CNVs. Also, the number of MI duplications (271) was higher than PI ones (225) as well as the number of rare duplications (i.e., 127 MI and 93 PI CNVs). The number of rare large deletions (16 MI and 18 PI) and duplications (30 MI and 28 PI) was similar between the two groups.

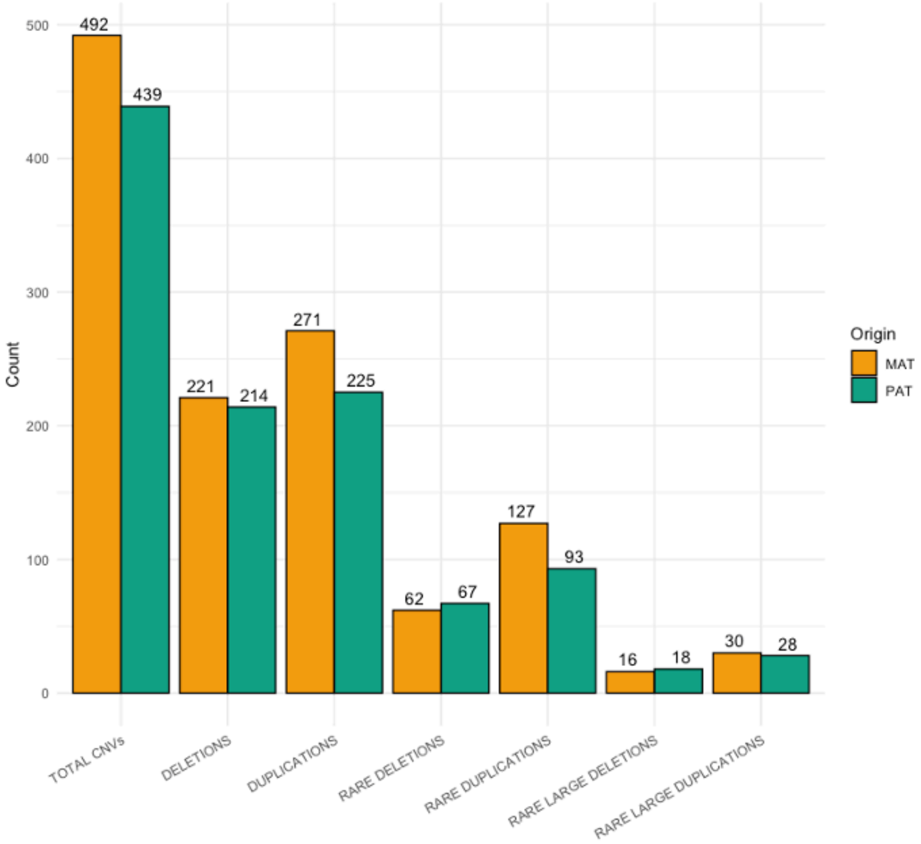


Figure 5.1. Comparison between the number of maternally (MAT) and paternally (PAT) inherited CNVs. The number of CNVs in the two groups were similar for most categories, except for the total number of CNVs and the number of total and rare duplications, where the maternally inherited ones were slightly higher. Rare = present in <1% of the samples of the DGV Gold Standard - Large = encompassing at least 10 sequencing targets.

Secondly, I tested whether PI and MI CNVs differed in their genomic size. Thus, I compared the size distribution of all deletions and duplications and the size distributions of the rare CNV events of the two groups (Figure 5.2A and 5.2B). There was not a significant difference between deletions and duplications sizes for both the compared categories.

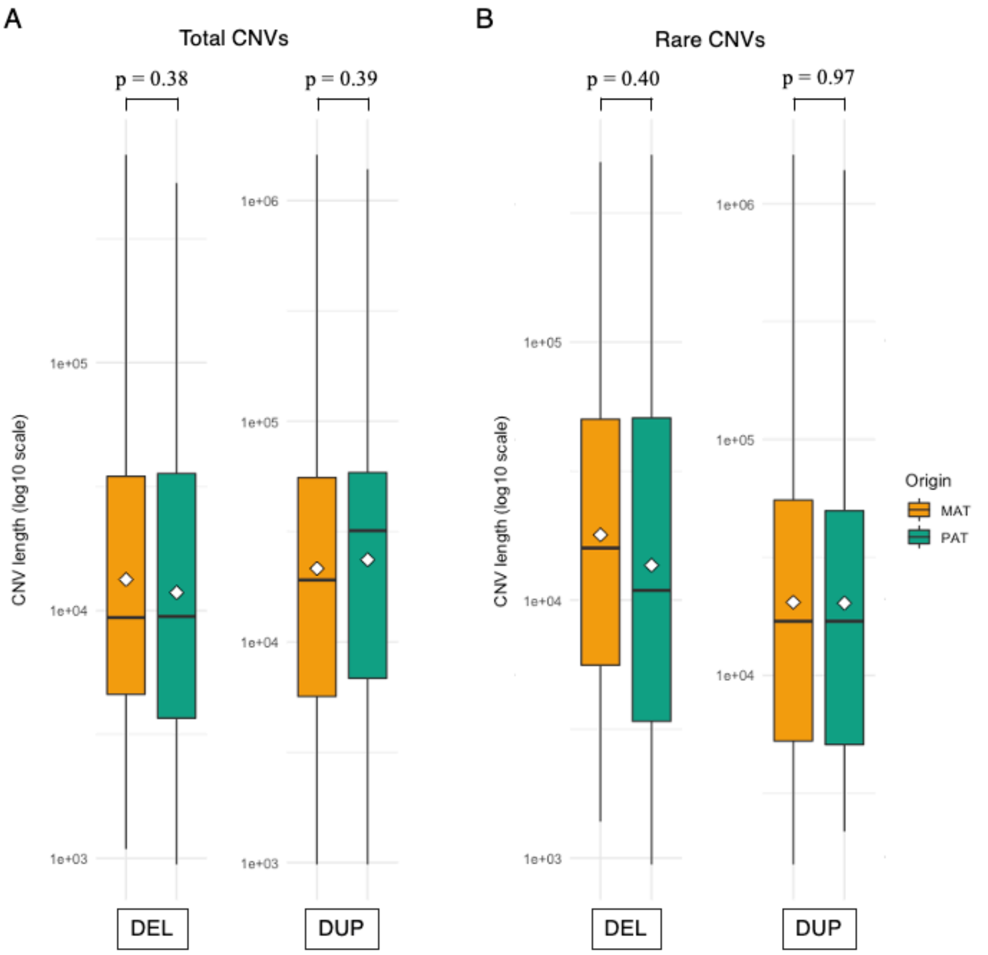


Figure 5.2. **A.** Comparison between the size distributions of all MI and PI duplications (DUP) and deletions (DEL). **B.** Comparison between the size distributions of rare MI and PI duplications and deletions. The white rhombuses within the boxes indicate the average size. There was not a significant difference in size for MI and PI deletions and duplications, both when considering all the CNVs and the rare ones.

5.3.2 Autosomal inherited deletions

Inherited CNVs are much more numerous than *de novo* CNVs, and prioritising likely pathogenic ones is challenging. We were interested in MI CNVs that might have a mild or no effect on female fertility but a substantial effect on male fertility. These CNVs could be transmitted to the offspring and would not be subjected to negative selection in the mothers. To identify the inherited CNVs that are most likely to be pathogenic, I selected the rare large deletions that encompassed at least one gene with a pLI score > 0.9. Only 3 deletions had these characteristics: one was PI, and two were MI (Table 5.1). Interestingly, these were the only deletions out of a total of 435 that affected a gene with a pLI score > 0.9, regardless of size and rarity.

Proband	Genomic location (GRCh37)	Inheritance	Size	Probes	Genes	Genes with a pLI score > 0.9
352	chr14:104978908-105478344	Maternal	499 kb	96	*ADSSL1 - AHNAK2 - AKT1 - C14orf180 - CDCA4 - CEP170B - CLBA1 - INF2 - LINC00638 - LINC02280 - MIR4710 - PLD4 - SIVA1 - TMEM179 - ZBTB42*	*AKT1 - CEP170B - INF2*
1387	chr2:54080195-54278263	Maternal	198 kb	46	*GPR75 - GPR75-ASB3 - PSME4*	*PSME4*
2133	chr22:45075604-45608300	Paternal	533 kb	46	*ARHGAP8 - KIAA0930 - LOC101927551 - LOC105373064 - MIR1249 - NUP50 - NUP50-DT - PHF21B - PRR5 - PRR5-ARHGAP8*	*PHF21B*

Table 5.1. Proband, genomic coordinates, inheritance, size, and genes involved in the rare large inherited deletions encompassing at least one gene with a pLI score > 0.9 identified in the trio cohort.

All these 3 deletions were rare, and the affected regions have not been implicated in any deletion listed in the DGV Gold Standard and GnomAD-SV database. The deletion identified on chromosome 22 of proband 2133 was PI (Figure 5.3). It was 533 kb large and encompassed 6 protein-coding genes. Amongst these, *PHF21B* has a pLI score = 0.99. Only intronic deletions have been reported in *PHF21B* in databases. *PHF21B* is highly expressed at the protein level in the brain but not in the testis, according to the Protein Atlas database. Its function is unknown; however, it might be related to stress response (Wong *et al.*, 2017). This deletion does not seem to be associated with the patient's phenotype.

One of the two MI deletions prioritised was found on chromosome 14 of proband 352. This CNV was 499 kb large and encompassed 12 protein-coding genes, of which 3 were likely intolerant to loss-of-function variants: *AKT1* (pLI = 0.98), *CEP170B* (pLI = 1) and *INF2* (pLI = 0.97) (Figure 5.4). Only intronic deletions have been reported for *AKT1* and *INF2* genes in population databases, while *CEP170B* is encompassed by deletions involving introns and exons in 40 samples of the DGV Gold Standard. Of these 3 genes, only *AKT1* shows high expression at the protein level in the testis, according to the Protein Atlas database. The function of *INF2* and *CEP170B* has not been yet clarified in humans. Instead, *AKT1* is involved in spermatogenesis and in the discussion, I present the literature information available on its function.

The other MI deletion was identified on chromosome 2 of proband 1387. This CNV was 198 kb large and encompassed 4 protein-coding genes (Figure 5.5). Only one of these genes, *PSME4*, had a high pLI score (1). The coding region of this gene is not affected by deletions in population databases. *PSME4* is highly expressed at the protein level in the human testis (and in several other tissues), according to the

Protein Atlas database, and plays an active role in spermatogenesis in mammals (Qian *et al.*, 2013). In the discussion, I review the literature papers describing its function.

Both the MI deletions presented in this paragraph, contrarily to the PI one, affected two likely haploinsufficient genes involved in testis functions and spermatogenesis.

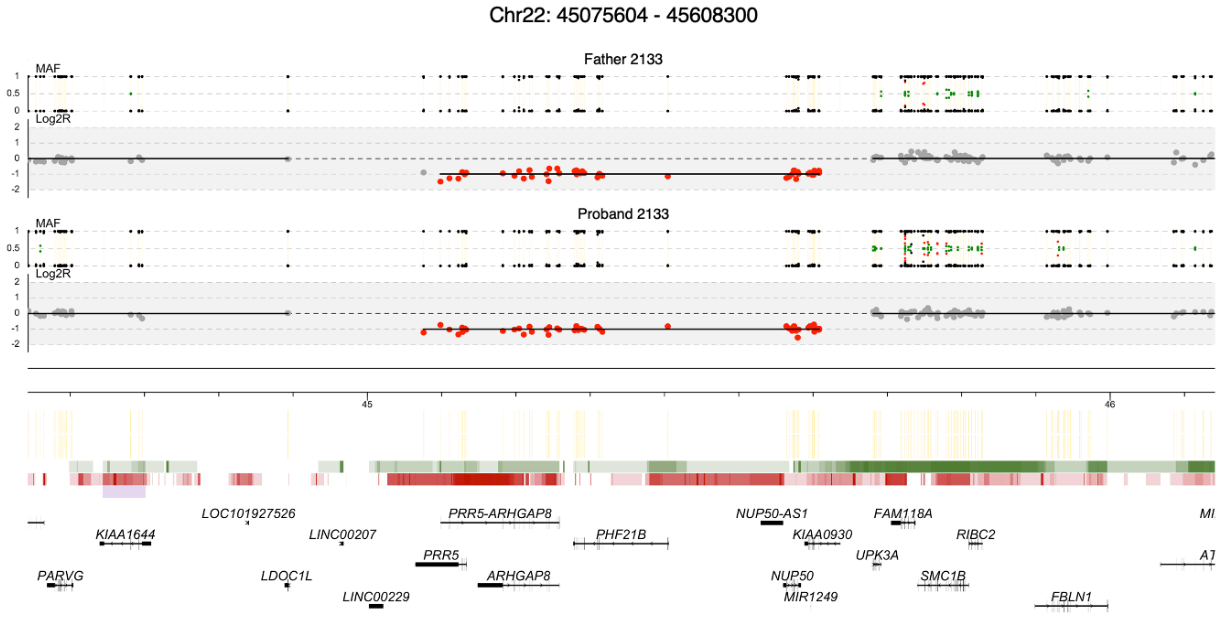


Figure 5.3. CNV plot of a rare large PI deletion identified in proband 2133. The deletion removed a single copy of 6 protein coding genes. The deletion is present in both father and proband. In both individuals, the MAF track shows absence of heterozygous SNPs (MAF = 0.5) within the deleted region, which suggests loss of heterozygosity.

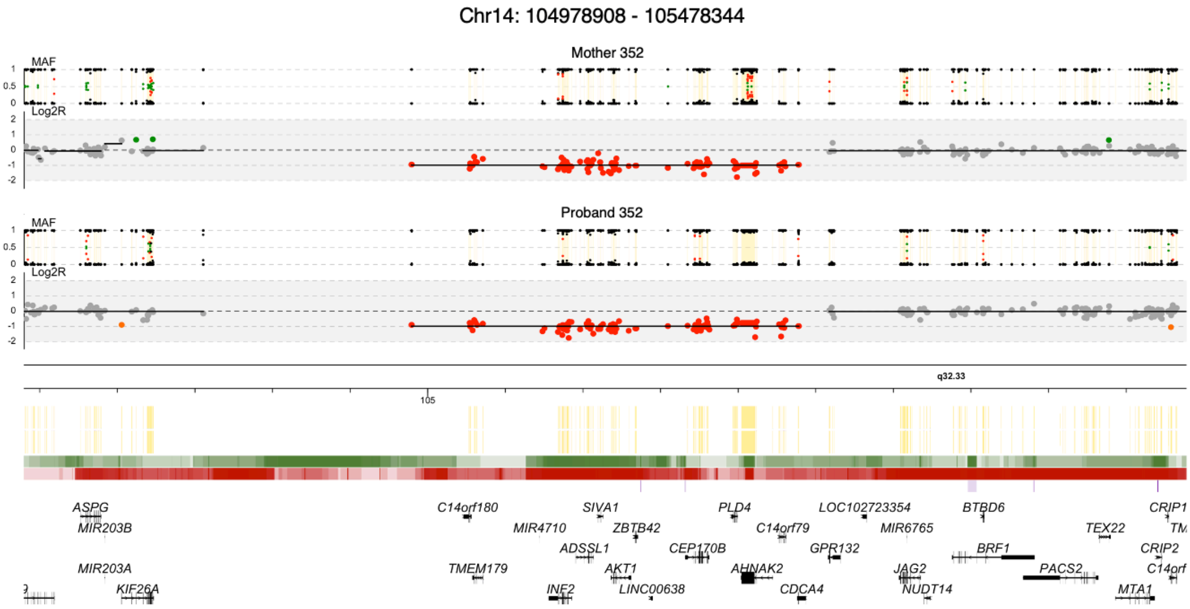


Figure 5.4. CNV plot of a rare large MI deletion identified in proband 352. The deletion removed a single copy of 11 protein coding genes. The deletion is present in both mother and proband.

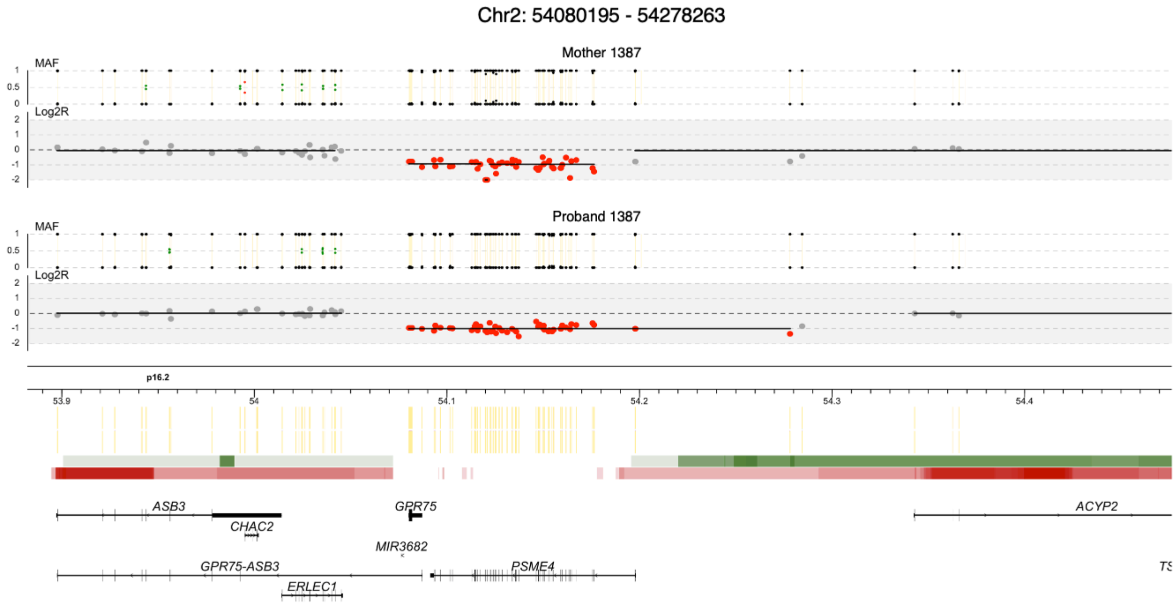


Figure 5.5. CNV plot of a rare large MI deletion identified in proband 1387. The deletion removed a single copy of 3 protein coding genes. The deletion is present in both mother and proband. In both individuals, the MAF track shows absence of heterozygous SNPs (MAF = 0.5) within the deleted region, which suggests loss of heterozygosity.

5.3.3 Autosomal inherited duplications

The clinical interpretation of duplications identified from WES data is even more complex than for deletions. For these variants, the WES data does not provide information on the number of additional copies nor on their position (tandem or interspersed) and orientation. Recently, it has been suggested that loss-of-function intolerant genes might also be sensitive to dosage increase (Collins *et al.*, 2020). For this reason, to identify autosomal inherited likely pathogenic duplications, I selected rare large gains encompassing genes with a pLI score > 0.9. High pLI score genes might be intolerant to dosage increase when entirely duplicated, or their genomic sequence might be altered when the breakpoints of a duplication are located in their coding region. Both these scenarios could impair the normal function of these genes and play a role in the origin of the disease in the patients.

Proband	Genomic location (GRCh37)	Inheritance	Size	Probes	Genes	Genes with a pLI score > 0.9	Genes with a pLI score > 0.9 disrupted by the CNV breakpoints
1697	chr10:12595148-12870962	Maternal	276 kb	10	*CAMK1D - MIR4480 - MIR4481 - MIR548Q*	*CAMK1D*	-
857	chr1:185097662-185130141	Maternal	32 kb	13	*SWT1 - TRMT1L*	*SWT1*	*SWT1*
1834	chr1:185097662-185130141	Maternal	32 kb	13	*SWT1 - TRMT1L*	*SWT1*	*SWT1*
466	chr2:29117489-29169721	Maternal	52 kb	18	*SNORD53 - SNORD53B - SNORD92 - WDR43*	*WDR43*	-
151	chr3:10327428 - 10491227	Maternal	164 kb	40	*ATP2B2 - GHRL - GHRLOS - LINC00852 - MIR378B - SEC13*	*ATP2B2*	-
2120	chr2:32631491-33246349	Maternal	615 kb	89	*BIRC6 - BIRC6-AS2 - LINC00486 - LTBP1 - MIR4765 - MIR558 - TTC27*	*BIRC6*	*BIRC6*
1487	chr12:2558065-3600955	Maternal	1 Mb	128	*CACNA1C - CACNA1C-AS1 - CACNA1C-AS2 - FKBP4 - FOXM1 - ITFG2 - ITFG2-AS1 - LINC02417 - LOC100128253 - NRIP2 - PRMT8 - RHNO1 - TEAD4 - TEX52 - THCAT155 - TSPAN9 - TULP3*	*CACNA1C - PRMT8*	*CACNA1C - PRMT8*
2133	chr12:19472879-19522797	Paternal	50 kb	14	*PLEKHA5*	*PLEKHA5*	*PLEKHA5*
1042	chr17:4007851-4077427	Paternal	70 kb	18	*ANKFY1 - CYB5D2 - ZZEF1*	*ANKFY1*	*ANKFY1*
625	chr14:35522406-35786609	Paternal	264 kb	34	*FAM177A1 - KIAA0391 - LOC101927178 - PPP2R3C - PSMA6*	*PSMA6*	-
2007	chr5:413383-619308	Paternal	206 kb	37	*AHRR - CEP72 - EXOC3 - EXOC3-AS1 - LOC100996325 - MIR4456 - PP7080 - SLC9A3 - SLC9A3-AS1*	*EXOC3 - SLC9A3*	*AHRR - CEP72*
2159	chr1:153701035-153800899	Paternal	100 kb	50	*GATAD2B - INTS3 - LOC343052 - SLC27A3*	*GATAD2B - INTS3*	-
2042	chr22:23634652-25024201	Paternal	1.1 Mb	188	*ADORA2A - ADORA2A-AS1 - C22orf15 - CABIN1 - CHCHD10 - DDT - DDTL - DERL3 - DRICH1 - GGT1 - GGT5 - GSTT1 - GSTT1-AS1 - GSTT2 - GSTT2B - GSTT4 - GSTTP2 - GUCD1 - GUSBP11 - IGLL1 - LOC391322 - LRRC75B - MIF - MIF-AS1 - MMP11 - POM121L9P - RGL4 - SLC2A11 - SMARCB1 - SNRPD3 - SPECC1L - SPECC1L-ADORA2A - SUSD2 - UBP1 - VPREB3 - ZNF70*	*SMARCB1*	-

Table 5.2. Proband, genomic coordinates, inheritance, size, and genes involved in the rare large inherited duplications encompassing at least one gene with a pLI score > 0.9 identified in the trio cohort. 6 duplications have at least one breakpoint within the coding region of a high pLI score gene.

Compared to the number of deletions involving genes with a pLI score > 0.9 (i.e., 3), the number of rare large duplications comprising the same class of genes was higher, with a total of 13 CNVs (Table 5.2). These were similar in number for MI and PI CNVs, with 7 and 6, respectively. 4 MI duplications disrupted the coding sequence of a high pLI score genes with the position of their breakpoints, while only 2 were PI. Contrarily to what I found for the MI deletions presented in the previous paragraph, we could not find, amongst these duplications, any that affected a high pLI score gene with a possible function in spermatogenesis.

For instance, both proband 857 and 1834 had the same MI duplication with a breakpoint within the *SWT1* gene (pLI score = 1). 35 partial duplications involving only the first exon or the first two exons of this gene have been listed in the DGV Gold Standard and GnomAD-SV database, while only one deletion affecting the second and third exon has been reported. According to the Protein Atlas database, this gene is highly expressed at the RNA level in the testis. However, its function is unknown. The MI duplications of both probands involve the first exon only (Figure 5.6), and they might be tolerated as indicated by the numerous gains comprising the first exons of the gene reported in population databases.

Another example is the 164 kb large MI duplication that entirely duplicated the gene *ATP2B2* (pLI score = 1) in proband 151 (Figure 5.7). The coding region of this gene has never been reported as duplicated in the population databases, and only one single-exon deletion has been described. Its function is related to ATP processing and calcium homeostasis (Fernandes *et al.*, 2007), but it has very low expression in the human testis, according to the Protein Atlas database.

Amongst the PI duplications, the breakpoint of the 50 kb rare large gain found in proband 2133 disrupted the gene *PLEKHA5* (pLI score = 1) (Figure 5.8). This gene has been reported as partially duplicated 103 times in the population databases, but no deletions affecting its coding regions have been listed. Disruption of the *PLEKHA5* leads to abnormal germ cell apoptosis and Sertoli cell morphology in mice (Xiao *et al.*, 2012), but in humans, the gene does not have high expression in the testis as reported in the Protein Atlas database. Also, *PLEKHA5* is altered in the fertile father 2133, who passed the duplication to the proband. Therefore, this gain is unlikely to be the origin of the patient's disease, assuming 100% penetrance.

These examples show that although some of the genes involved in these rare large gains might be associated with testis functions in other species or have high expression in the human testis, we cannot predict the consequences of their duplications. Some genes have unknown functions, such as *SWT1*. Others might not be involved in spermatogenesis, such as *ATP2B2*. Some might be dispensable for normal spermatogenesis since altered in fertile fathers, such as *PLEKHA5*.

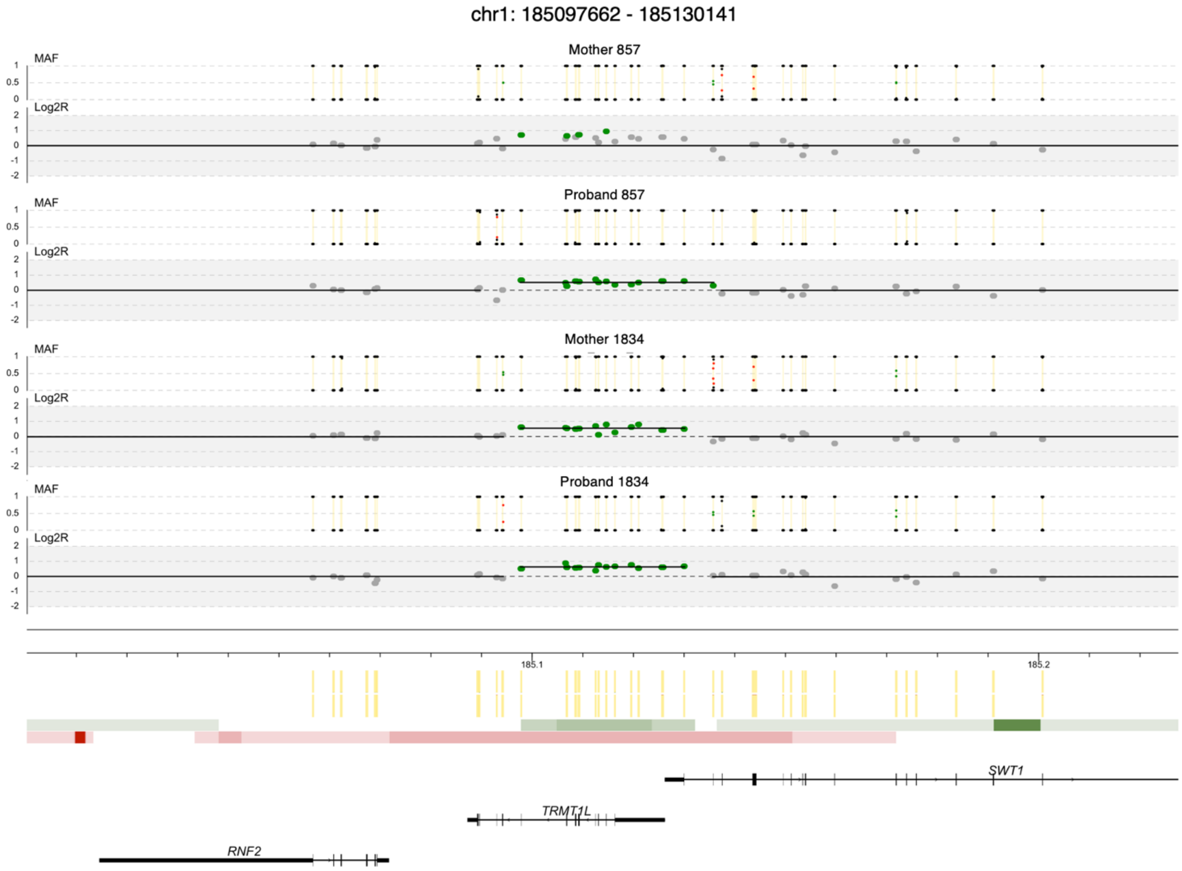


Figure 5.6. CNV plots of rare large MI duplications identified in probands 857 and 1834. The breakpoints of these CNVs affect the coding region of the gene *SWT1* (pLI score = 1), specifically, they encompass the first exon only. The duplications are shown in the respective mothers as well. In mother 857, the duplication is not segmented but the targets within the duplicated region have a log2R similar to those in the proband (> 0).

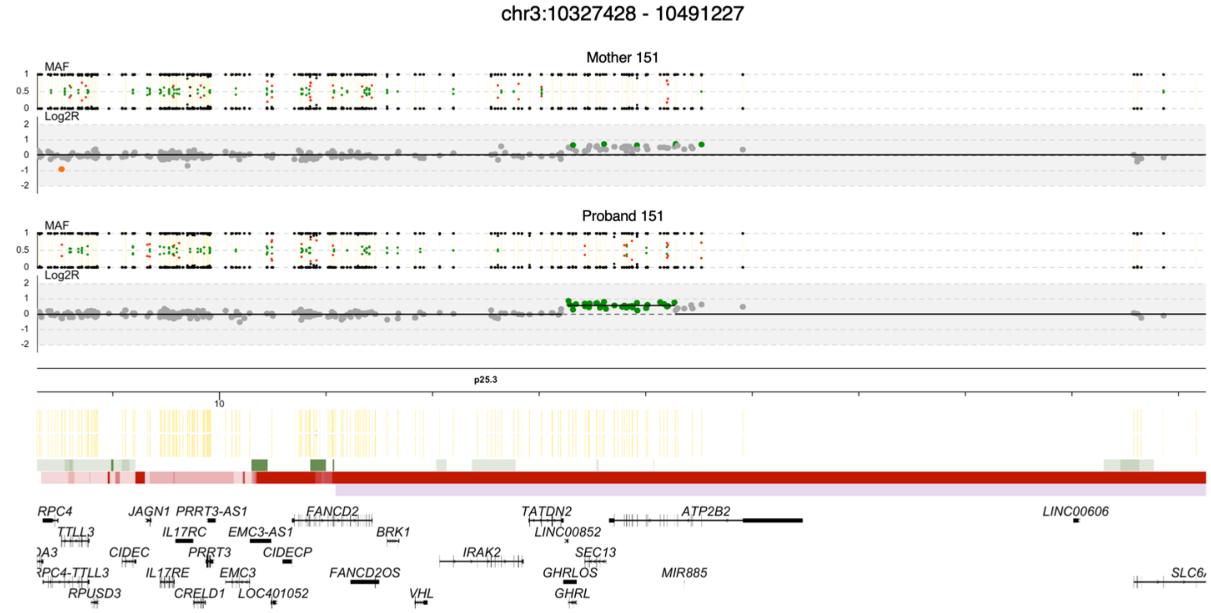


Figure 5.7. CNV plot of a rare large MI duplication identified in proband 151. The gene *ATP2B2* (pLI score = 1) is entirely duplicated in mother and proband. Some targets in the proband are not segmented in the duplication, but their log2R (> 0) is similar to that of the targets within the inferred gain. In mother 151, the duplication is not segmented but the targets within the duplicated region have a log2R similar to those in the proband.

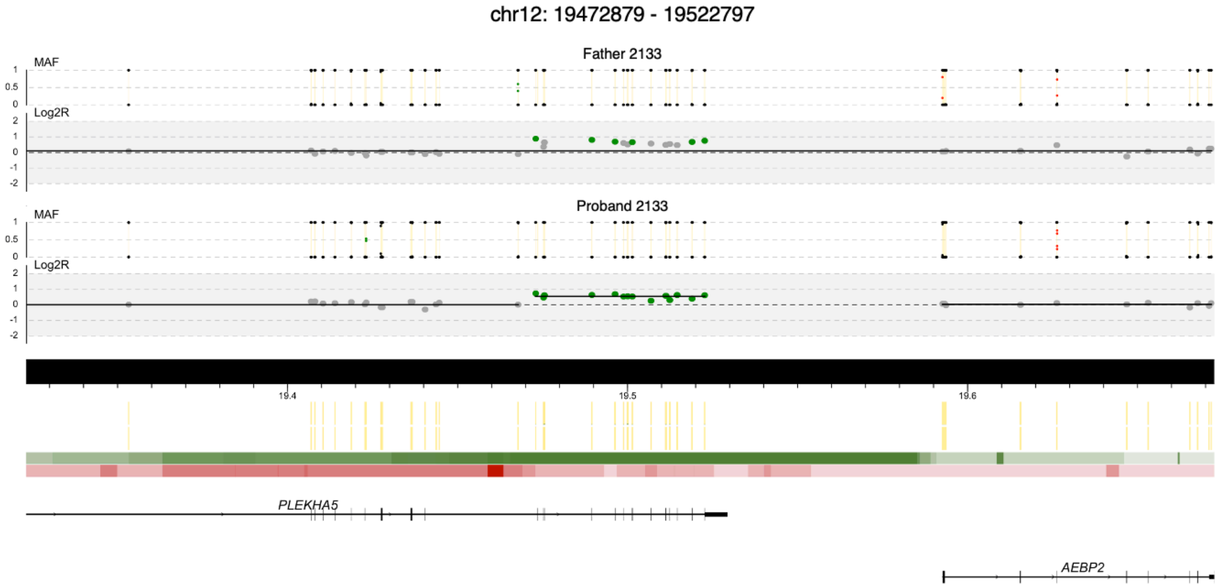


Figure 5.8. CNV plot of a rare large PI duplication identified in proband 2133. One of the breakpoints of this CNV affects the coding region of the gene *PLEKHA5* (pLI score = 1) in both proband and father. In father 2133, the duplication is not segmented but the targets within the duplicated region have a log2R similar to those in the proband (> 0).

5.3.4 Inherited CNVs on the sex chromosomes

The CNVs identified on the sex chromosomes of the probands were initially excluded from the comparison of MI and PI CNVs. In total, 25 inherited CNVs were identified on these chromosomes. The only CNVs identified on chromosome Y were 3 small (encompassing 3 targets) PI duplications. The remaining 22 CNVs were detected on the X chromosome, and only one was a deletion. Assuming that male infertility cannot be inherited from fertile fathers, the small PI duplications on the Y chromosomes were excluded from further analysis. We examined instead the CNVs found on the X chromosome of probands and their mothers. To identify the most likely pathogenic MI CNVs, I selected the only deletion detected and all rare large duplications (2) (Table 5.3).

Proband	Genomic location (GRCh37)	Type	Size	Probes	Genes	Genes disrupted by the CNV breakpoints
24	chrX:134946157-134995055	Deletion	49 kb	8	*LOC102723631 - SAGE1*	-
368	chrX:6451786-8138461	Duplication	1.7 Mb	26	*MIR4767 - MIR651 - PNPLA4 - PUDP - STS - VCX - VCX2 - VCX3A*	-
992	chrX:105139371-105451592	Duplication	312 kb	25	*NRK - PWWP3B - SERPINA7*	*NRK*

Table 5.3. Proband, genomic coordinates, CNV type, size, and genes involved in the X-linked MI deletion and the rare large X-linked MI duplications identified in the trios.

The only MI deletion identified on chromosome X was found in proband 24, and it was 49 kb in size (Figure 5.9). It removed the unique copies of two genes: *LOC102723631* and *SAGE1*. *SAGE1* has high expression exclusively in the testis and its protein has been found in spermatogonia specifically, according to the Protein Atlas database. Its coding region has been reported as deleted only in one female individual of the GnomAD-SV database. However, the function of *SAGE1* is unknown.

The first rare large MI duplication identified was found in proband 368. The CNV involved a 1.7 Mb region (Figure 5.10) that has not been reported as entirely duplicated in the samples of the DGV Gold Standard and GnomAD-SV database.

Three of the genes involved (*VCX*, *VCX2* and *VCX3A*) are part of the *VCX* genes family. *VCX* is a multi-copy gene exclusively expressed in male germ cells, specifically during spermatogenesis (Van Esch *et al.*, 2005). Duplications of this gene have been associated with male infertility and apoptosis in a 2016 study (Ji *et al.*, 2016). Other reports of variants found in this region are reviewed in detail in the discussion.

The second MI rare large duplication on chromosome X was identified in proband 992 (Figure 5.11). The CNV was 312 kb large and encompassed 3 genes. One of the predicted breakpoints of this duplication is located in the coding region of the gene *NRK* and disrupt its sequence. Only intronic deletions have been reported for this gene in the DGV Gold Standard and GnomAD-SV database. Its function is unclear, but it has been suggested to be related to late embryogenesis (Nakano *et al.*, 2000). Its protein is expressed exclusively in the ovary and at high level in the placenta, according to the Protein Atlas database.

Overall, we identified a MI deletion (proband 24) and a MI duplication (proband 368) on chromosome X involving two genes (*SAGE1* and *VCX*) that are associated with male fertility, however, their specific role in spermatogenesis and testis function is unclear.

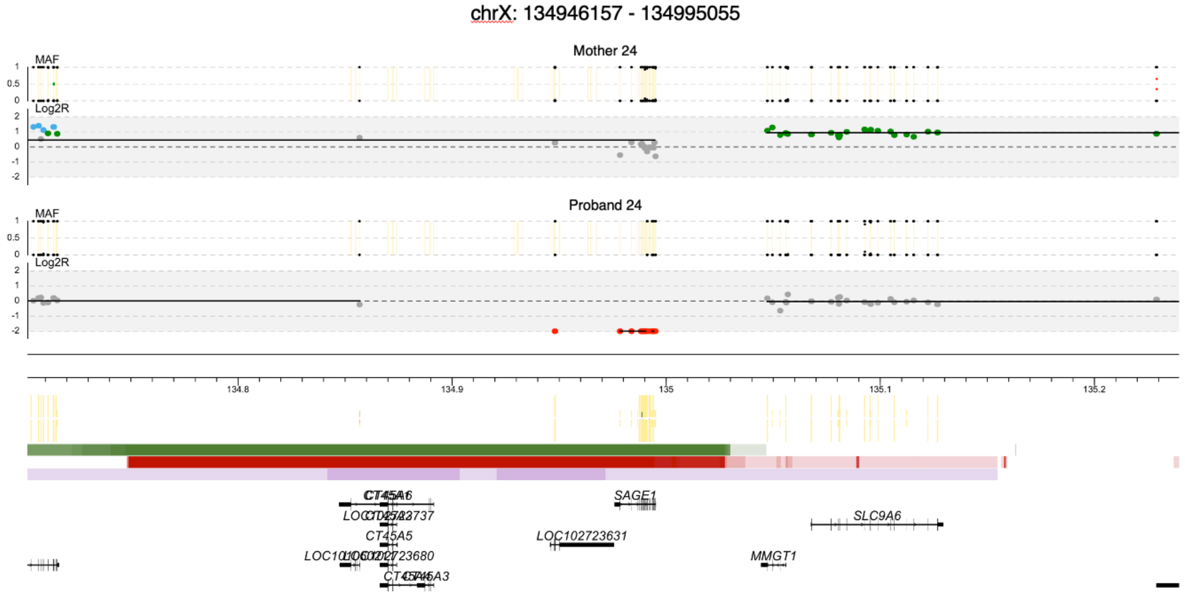


Figure 5.9. CNV plot of a rare MI X-linked deletion identified in proband 24. The CNV affects 2 protein coding genes. The Log2R tracks show a single copy of the deleted region (Log2R = 0) in the mother and 0 copies in the proband (Log2R = -2). The MAF track shows the absence of SNPs for only a part of the deletion. Examination of the read alignment file showed that these SNPs were mostly supported by 1 or 2 reads only or by low-quality mapping reads and that the majority of *SAGE1* exons were not covered by any read. Therefore, the SNPs represented are likely due to imperfect segmentation during the CNV workflow and erroneous reads mapping.

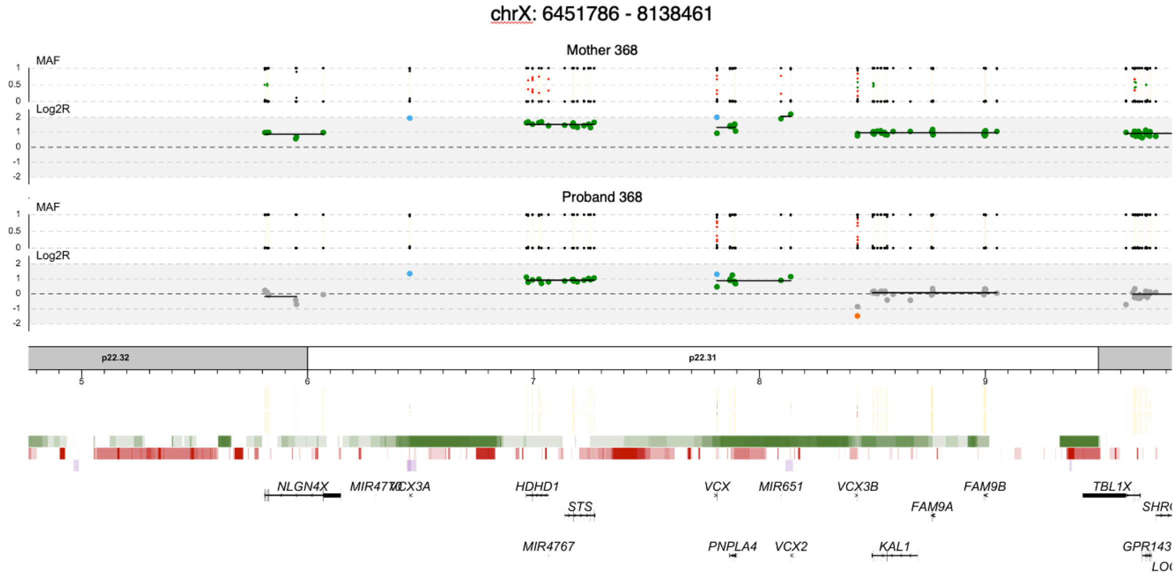


Figure 5.10. CNV plot of a rare large MI X-linked duplication identified in proband 368. The CNV encompassed 7 protein coding genes. 3 genes were part of the *VCX* family (*VCX*, *VCX2* and *VCX3A*). The Log2R tracks show additional copies in both mother and proband. In the proband, the Log2R for the duplicated region is equal to 1 because there is only one copy of the X chromosome in males. In the mother, the log2R > 1 indicates more than 2 copies.

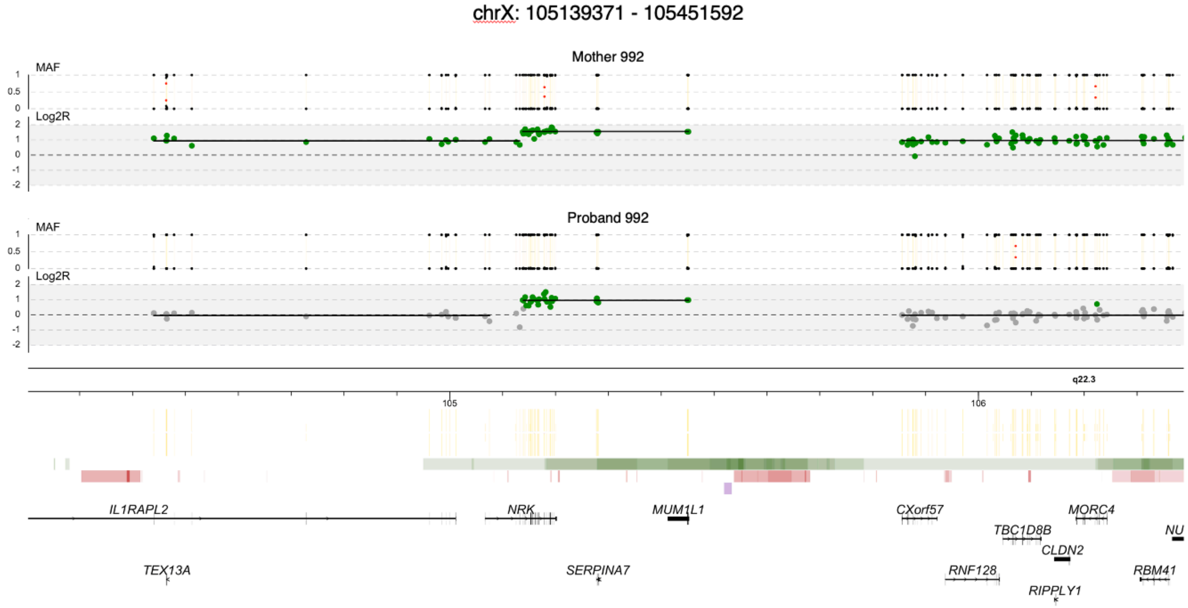


Figure 5.11. CNV plot of a rare large MI X-linked duplication identified in proband 992. The CNV encompassed 3 protein coding genes. One of the predicted breakpoints disrupts the sequence of *NRK* gene. The Log2R tracks show additional copies in both mother and proband. In the proband, the Log2R for the duplicated region is equal to 1 because there is only one copy of the X chromosome in males. In the mother, the log2R > 1 indicates more than 2 copies.

5.3.5 Inherited CNVs overlapping de novo mutated genes

In addition to prioritising inherited CNVs according to size and rarity in population databases, I looked for inherited CNVs that involved genes carrying a *de novo* mutation classified as possibly causative of the disease or of unknown significance (98 *de novo* mutations in 96 different genes) (see Appendix B). These mutations were identified and classified during the *de novo* SNV analysis conducted by my colleagues Dr Manon Oud and Hannah Smith. Also, I looked for inherited CNVs overlapping with the genes affected by the 2 *de novo* CNVs reported in chapter 4 (*QSER1, DEPDC7, TCP11L1, CSTF3, CSTF3-AS1, HIPK3, KIAA1549L, NXT2*). Three CNVs were prioritised in total. Two have already been described in this chapter: (1) the MI deletion in proband 1387 (Table 5.1 – Figure 5.5) involved the gene *GPR75-ASB3*, *de novo* mutated in proband 2010; (2) the PI duplication in proband 2042 encompassed *SPCC1L* (Table 5.2), *de novo* mutated in proband 2080. In addition, a 13kb PI duplication found in proband 94 encompassed *SMC2*, which carried a *de novo* mutation in proband 1296 (Figure 5.12).

GPR75-ASB3 gene has an unknown function, but it might be involved in protein ubiquitination (UniProt - A0A6D2WFD3). This gene is deleted in proband 1387 together with 3 other genes, of which *PSME4* was deemed the most likely candidate disease gene. (See paragraph 5.3.2).

SPECC1L gene may play a role in stabilising microtubules and cytoskeleton organisation (Saadi *et al.*, 2011). In proband 2042, the gene is entirely duplicated, and the CNV breakpoints do not disrupt its sequence. Considering that the duplication is PI, we know that the duplication of *SPECC1L* did not lead to infertility in father 2042, and it is unlikely to cause the disease in proband 2042, assuming 100% penetrance.

SMC2 is part of a condensin complex that regulated chromatin status (Kimura, Cuvier and Hirano, 2001). In proband 94, the gene's coding sequence is possibly disrupted by a breakpoint of a PI duplication (Figure 5.12). However, the same duplication in father 94 again suggests that the gain is not causative of infertility, assuming 100% penetrance.

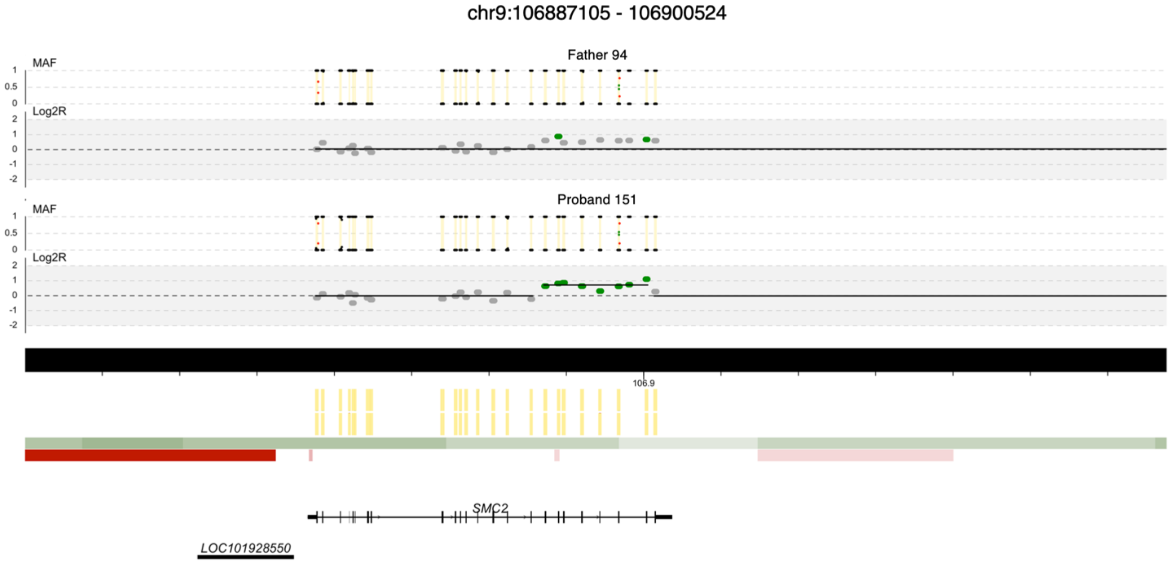


Figure 5.12. CNV plot of a PI duplication identified in proband 94. One breakpoint of the gain disrupts the coding sequence of the gene *SMC2*, which is also *de novo* mutated in proband 1296.

5.4 Discussion

In this chapter, I investigated the inherited CNVs identified in 183 patient-parent trios affected by azoospermia and severe oligozoospermia. First, I tested whether the number of autosomal rare large MI CNVs detected in the cohort exceeded the equivalent inherited from the fathers. This required creating a high-confidence dataset of autosomal inherited CNVs that excluded possible technical artefacts and CNVs for which inheritance was unclear. This method led to the exclusion of 53% of the total variants detected (2054). Excluding such a large percentage of CNVs is not ideal and likely excluded some true positive CNVs. However, the selection of a dataset of high-confidence inherited CNVs was necessary to perform an objective comparison between MI and PI CNVs. Since 89% of the excluded CNVs encompassed less than 10 sequencing targets, the majority of the CNV excluded are not likely to represent large genomic events, which are more likely to be pathogenic (Li *et al.*, 2020). The results of the comparison suggested that rare (present in <1% of the samples of the DGV Gold Standard) and large (encompassing at least 10 sequencing targets) CNVs were not disproportionally inherited from the mothers (46 MI – 46 PI). A slightly higher number of total and rare MI duplications was identified. This difference could have been caused by chance and/or it could reflect the manual curation of the inherited CNVs, as the number of variants in these two categories was compared regardless of the CNV size. The size distribution of all the deletions and duplications and the size distribution of rare ones were not significantly different between MI and PI variants. We did not expect an increase in the size or number of MI CNVs, since generally, autosomal CNVs are half MI and half PI. Although this comparison seems to support this expectation also in male infertility patients, the results would require confirmation using a more accurate method for CNV detection, such as WGS.

To identify likely pathogenic MI CNVs and novel dominant candidate male infertility genes, I selected, amongst the rare large CNVs, those that affected genes with a pLI score > 0.9. First, I investigated the 3 inherited deletions that fell into this category. Remarkably, amongst all the deletions detected, only these 3 affected a gene with a high pLI score. This suggests that this type of variants might be deleterious, and therefore, only a few of them are transmitted to the offspring. The only PI deletion in this category could not be associated with the patient's phenotype, as expected. In contrast, we propose the two MI deletions as likely pathogenic for the reproductive fitness of the respective patients.

One MI deletion (in proband 352) encompassed *AKT1* (pLI score = 0.98). Aitken and Koppers, in 2011, reviewed previous studies on DNA damage in spermatozoa and suggested that apoptosis is the default pathway for spermatozoa. In fact, there seems not to be endogenous chemical triggers for this process, and only prosurvival factors prevent these cells from following this route (Aitken and Koppers, 2011). In the same year, Koppers *et al.* described *AKT1* as a key factor for spermatozoa survival in humans and suggested that its phosphorylated status prevents spermatozoa from following the default apoptotic pathway (Koppers *et al.*, 2011; Aitken, 2018). It has also been demonstrated that the *AKT1* gene is essential for normal spermatogenesis in mice (Chen *et al.*, 2001; Kim, Omurtag and Moley, 2012). This deletion also removed one copy of the high pLI score genes *INF2* (pLI score = 0.97) and *CEP170B* (pLI score = 1). The function of these genes is unknown in humans. Therefore, we could not assess the consequences of their deletion.

The other MI deletion (in proband 1387) removed one copy of *PSME4* (pLI score = 1). *PSME4* is a component of the spermatoproteasome that participates in the histone exchange during spermatogenesis in mammals (Qian *et al.*, 2013). It is also required for normal male fertility in mice, but it is dispensable for female reproductive fitness (Khor *et al.*, 2006; Qian *et al.*, 2013; Huang *et al.*, 2016).

Deletions of the coding regions of *PSME4* and *AKT1* have not been reported population databases. The disruption of these genes might be the cause of severe isolated male infertility in patients 352 and 1387, and we classify these two MI deletions as possibly pathogenic. These CNVs could impair specifically male fertility genes without affecting female fertility, as suggested by the experiment on KO mice for *PSME4* (Khor *et al.*, 2006). This could be the reason why such CNVs were not purified by negative selection in the mothers of probands 352 and 1387. These variants are the first identified MI likely pathogenic CNVs in isolated severe male infertility cases and could represent a novel cause of the disease.

It is worth mentioning that proband 352 carries another candidate pathogenic variant identified by the SNV analysis conducted by my colleagues: a likely pathogenic *de novo* mutation in *ABLIM1*, a gene supporting the function of the Sertoli cells in the testis (Hu *et al.*, 2014). In proband 1387, instead, the MI deletion is the only candidate causative variant identified so far.

As well as studying likely pathogenic deletions, I investigated the rare large inherited duplications that encompassed genes with a pLI score > 0.9. These genes could be dosage-sensitive as well as loss-of-function intolerant (Collins et al., 2019).

In general, the analysis of duplications identified from WES data is more difficult than the analysis of the deletions since the data do not provide information on either the number of additional copies or their position and orientation. Also, it is difficult to establish the consequences of additional gene copies exclusively from publicly available gene information. Amongst the 13 duplications comprised in this category (7 MI and 6 PI), 7 gains increased the number of copies of high pLI score genes, while 6 affected their exon sequence as inferred from the position of the breakpoints. Although we cannot exclude that MI duplications involving dosage-sensitive genes might cause the disease in some patients, we could not identify duplications that impaired a high pLI score gene associated with testis function and spermatogenesis in humans. The incomplete duplication data that CNV detection from WES data provides does not allow us to prioritise these gains and assess their potential pathogenicity effectively.

I also assessed the potential pathogenicity of CNVs on the sex chromosomes of the probands. The 25 inherited CNVs identified on these chromosomes represented 1.2% of all the CNVs detected in this study. This is logical considering the higher number of autosomal chromosomes compared to the unique X and Y chromosomes in males. In addition, the X and Y chromosomes share large regions of homology, such as the PAR1 and PAR2 loci at the telomeres (Helena Mangs and Morris, 2007), and short-read sequencing produces reads that cannot be mapped accurately to these regions, precluding their analysis with this technology. Also, the probands of the cohort underwent AZF deletion screening, hence, patients with large losses on the long arm of the Y chromosomes were excluded from the study.

The 25 inherited CNVs were excluded from the previous MI and PI CNV comparisons since X and Y chromosomes are different in size and gene content. Amongst them, I investigated variants on the unique copy of the X chromosome, which is transmitted from the mother in males. The inheritance of these CNVs was confirmed by the visual inspection of the CNV plots. Only one deletion was identified on chromosome X, as well as two rare large duplications. The deletion was found in patient 24 and removed the unique copy of *SAGE1*, a gene with high protein expression specifically in spermatogonia, according to the Protein Atlas database. This gene might be important for fertility in males; however, its function is unknown, and we could not assess whether its loss could cause the disease in patient 24. Identifying a single deletion on chromosome X suggests that this chromosome is particularly intolerant to loss-of-function variants, as any gene involved in a loss would not have a second copy in males.

Of the two gains prioritised, only one might be involved in testis functions: the duplication found in patient 368 that encompassed 3 genes from the *VCX* family. *VCX* gene is expressed in male germ cells during spermatogenesis (Lahn and Page, 2000; Zou *et al.*, 2003; Van Esch *et al.*, 2005). *VCX* duplications were found to be more frequent in infertile men than controls by Ji *et al.*, who also performed *in vitro* experiments that demonstrated that upregulation of *VCX* could lead to apoptosis (Ji *et al.*, 2016). Duplications in Xp22.31, the region in which *VCX* is located, have been reported as benign (Shaw-Smith *et al.*, 2004; Zhuang *et al.*, 2019) but also associated with intellectual disability (Esplin *et al.*, 2014; Pavone *et al.*, 2019). A recent study investigating this region reported a duplication involving *VCX* in a fertile father, but the exact number of gene copies was not established (Zhuang *et al.*, 2019). The consequences of variants in this region have been debated for a long time (Li *et al.*, 2010), but since the role of additional copies of *VCX* in severe male infertility patients and *in vitro* was assessed by Ji *et al.*, no follow-up studies have been published on the topic. Further studies allowing more accurate characterisation of *VCX* duplications and copy number assessment in fertile and infertile men are essential.

Lastly, I prioritised the CNVs that involved genes that carried *de novo* mutations classified by my colleagues as possibly causative of the disease or of unknown significance in the trios. The gene list and classification were created by my colleagues who worked on the analysis of the *de novo* SNVs found in the trio cohort. At the same time, I searched for CNVs encompassing the genes involved in *de novo* deletions in the trios. This analysis revealed three CNV overlapping with such genes. The only MI CNVs that comprised one of such genes (i.e., *GPR75-ASB3*) was the MI deletion found in proband 1387 and discussed above. The loss was already considered likely pathogenic because it impaired *PSME4*, a compelling novel candidate dominant male infertility gene, but there are no data supporting the involvement of *GPR75-ASB3* in spermatogenesis. The other two CNVs encompassing *de novo* mutated genes were PI duplications. Hence, they might have a different consequence on the gene than a *de novo* mutation, and they are unlikely to be causative considering their presence in a fertile father. This analysis did not reveal additional candidate disease genes affected by both a *de novo* mutation and an inherited CNVs. Given the high genetic heterogeneity of severe male infertility, such outcome is not unexpected and finding candidate disease genes recurrently mutated in only 183 probands was unlikely.

In summary, this exploratory study revealed novel information on the role of MI CNVs in severe male infertility. Our results showed for the first time that there is no difference in the number of MI and PI rare large CNVs, but the genes encompassed might differ. In the cohort of 183 patient-parent trios, the only two MI deletions involving a gene with a pLI score > 0.9 both affected a gene involved in spermatogenesis and possibly important for normal fertility in males. The same situation was not found for the unique equivalent PI deletion. These variants could arise as *de novo* in the mothers and have mild or no deleterious effects on female fertility. Therefore, they would not be purified by negative selection and could be transmitted to the offspring. This study highlighted two novel candidate dominant male infertility genes: *AKT1* and *PSME4*. The MI autosomal duplications seem to not play a prominent role in the disease. At the same time, the analysis of MI X-linked CNVs highlighted two genes of potential interest for further studies: *SAGE1* and *VCX*. Further studies should be designed to investigate CNVs in other trio cohorts to test the inheritance of potentially pathogenic variants since our results suggest that maternally inherited CNVs might explain idiopathic cases of severe male infertility.

5.5 Conclusions

In this study, I used the cohort of 183 patient-parent trios affected by severe idiopathic male infertility, presented in the previous chapter, to systematically explore for the first time the role of MI CNVs in the disease aetiology. No substantial difference was identified between the number of rare large MI and PI CNVs, nor a difference in size between the total and the rare CNVs of the two groups. The only two MI deletions encompassing a high pLI score gene were found to be likely pathogenic for male fertility as they involved two different novel male fertility candidate genes (*PSME4* and *AKT1*). These losses are the first identified MI CNVs that could cause a severe male infertility phenotype. Moreover, the analysis of the MI X-linked CNVs revealed two loci of potential interest for the field that should be further studied. This study demonstrates the potential of studying genetic variations in cohorts of male infertility patients for which parental samples are available, and we hope that this study, as well as the investigation of *de novo* variants presented in the previous chapter, will push the field of male infertility genetics to pursue this approach.

Chapter 6. CNVs in a Cohort of Patients With Idiopathic Quantitative Forms of Male Infertility

6.1 Introduction

In addition to the 183 patient-parent trios described in the earlier chapters, we recruited 142 patients affected by azoospermia or severe oligozoospermia for which parental samples were unavailable. DNA from parents is often unavailable for several reasons (see paragraph 2.3 of Chapter 1) and, for this reason, male infertility research, so far, focused on patient-only cohorts.

The analysis of singleton cohorts does not provide information on the inheritance of the genetic variants identified; thus, *de novo* or maternally inherited CNVs cannot be prioritised. Nevertheless, in the literature, several articles investigated the CNVs in cohorts of male infertility patients and contributed significantly to the current knowledge on male infertility genes. For instance, two SNP-array based studies in 2011 revealed that homozygous deletions in *DPY19L2* contribute to the aetiology of globozoospermia, a male infertility condition in which the spermatozoa present round heads and no acrosome (Harbuz *et al.*, 2011; Koscinski *et al.*, 2011). In the following years, Lopes *et al.* and Lima *et al.* reported different deletions in non-obstructive azoospermia patients affecting the *DMRT1* gene (Lopes *et al.*, 2013; Lima *et al.*, 2015), one of the four autosomal dominant genes (*DMRT1*, *HSF2*, *KLHL10*, SYCP3) that, to date, have been reliably associated with isolated male infertility (Oud *et al.*, 2019). As discussed in the previous chapter, a few studies have focused on chromosome X (Krausz *et al.*, 2012; Chianese *et al.*, 2014; Lo Giacco *et al.*, 2014; Yatsenko *et al.*, 2015) since this chromosome is maternally transmitted in males. A few others have tried to compare the CNV burden on the autosomes between cases and controls (Frank Tüttelmann *et al.*, 2011; Stouffs *et al.*, 2012). For example, in 2011, Tüttelmann *et al.* performed the first array-CGH based study in male infertility patients and reported CNVs specific to infertile men with severe oligozoospermia and Sertoli-cell-only syndrome. This study provided the first indication that heterozygous variants might contribute to the aetiology of the disease and pointed to new genomic loci and novel candidate genes for further investigations. All these studies demonstrate that CNV analysis in patient-only cohorts may still provide useful findings.

The CNV studies mentioned above were primarily microarray-based. In contrast, we use WES data, which, as discussed before, allows robust CNV and SNV detection with the same genomic test. Variant interpretation without parental samples remains challenging. However, we can use bioinformatic tools to identify variants of clinical interest. This strategy was used in the previous chapter, where the pLI score, which indicates the probability of a gene being intolerant to loss-of-function variants (Lek *et al.*, 2016), was used to select likely pathogenic maternally inherited CNVs. In this chapter, I investigate the CNV burden in 142 male infertility patients and prioritise rare and large CNVs that affect likely haploinsufficient genes with a possible function in spermatogenesis. Moreover, I use the findings presented in the previous two chapters to look for recurrently mutated genes.

6.2 Aims

In this chapter, I aimed to:

- Identify and assess the potential pathogenicity of rare and large CNVs affecting likely haploinsufficient genes identified in the cohort of 142 patients.
- Identify CNVs involving genes affected by *de novo* variants or likely pathogenic maternally inherited CNVs in the trios.
- Identify novel candidate dominant, X- or Y-linked male infertility genes.

6.3 Results

A total of 1388 CNVs were identified in 142 patients affected by azoospermia or severe oligozoospermia, with an average of 10 CNVs per patient. The total number of CNVs identified in the autosomes was 1379. Of these, 677 were deletions and 702 duplications (Figure 6.1). The number of CNVs drastically decreased when we selected rare (present in < 1% of the samples of the DGV Gold Standard) and large (involving >= 10 targets) CNVs. Only 29 deletions and 41 duplications. 5 deletions and 12 duplications of this group encompassed a gene with pLI score > 0.9. On the sex chromosomes, no deletions were detected. 5 of the 9 duplications identified were rare and large.

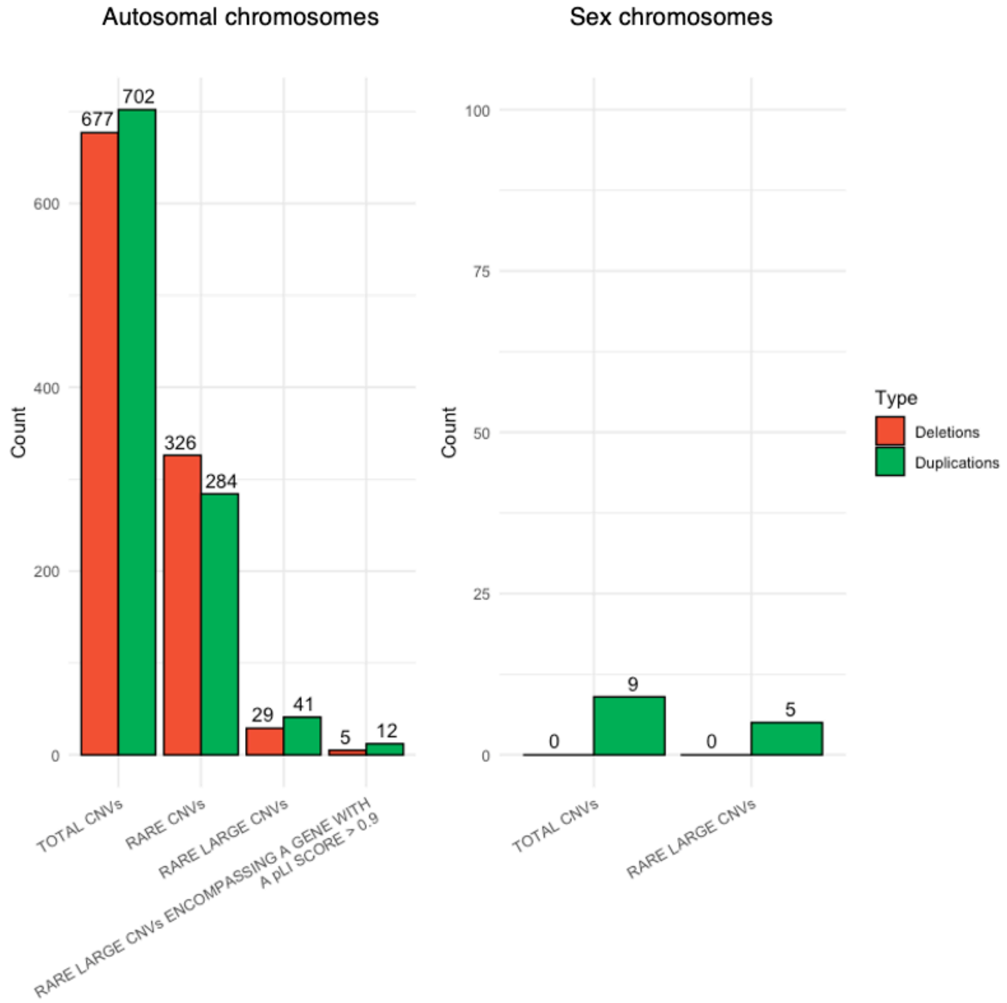


Figure 6.1. Number of CNVs identified 142 patients with azoospermia or severe oligozoospermia. The CNVs were divided into different categories according to size, rarity and the pLI score of the genes encompassed. On the autosomes, there were slightly more duplications than deletions in total as well as for the rare large CNVs and for those encompassing a gene with a pLI score > 0.9. On the sex chromosomes, only duplications were detected. Rare = present in < 1% of the samples of the DGV Gold Standard). Large = involving >= 10 targets.

6.3.1 Autosomal deletions

First, I investigated the autosomal deletions identified in this cohort of patients. I prioritised the rare large variants that affected likely loss-of-function intolerant genes (pLI score > 0.9), similarly as done in the previous chapter, and assessed their potential role in male infertility. Table 6.1 shows the 5 deletions prioritised with this method.

Patient	Genomic location (GRCh37)	Size	Probes	Genes	Genes with a pLI score > 0.9
380	chr5:10235247-10261933	27 kb	13	*ATPSCKMT - CCT5*	*CCT5*
1141	chr10:115965927-116021130	55 kb	19	*TDRD1 - VWA2*	*TDRD1*
1790	chr7:5347661-5521638	174 kb	27	*FBXL18 - LOC100129484 - TNRC18*	*TNRC18*
1453	chr7:32526797-35058284	2.5 Mb	111	*AVL9 - BBS9 - BMPER - DPY19L1 - DPY19L1P1 - DPY19L1P2 - FKBP9 - FLJ20712 - KBTBD2 - LINC00997 - LSM5 - MIR548N - MIR550A2 - MIR550B2 - NPSR1 - NPSR1-AS1 - NT5C3A - RP9 - RP9P - ZNRF2P1*	*KBTBD2*
53	chr16:15457440-17564729	2.1 Mb	156	*ABCC1 - ABCC6 - BMERB1 - CEP20 - LOC102723692 - MARF1 - MIR484 - MIR6506 - MPV17L - MYH11 - NDE1 - NOMO3 - NPIPA5 - XYLT1*	*XYLT1*

Table 6.1. Proband, genomic coordinates, size, and genes involved in the rare large deletions encompassing at least one gene with a pLI score > 0.9 identified in the 142 patients.

Two of these deletions were larger than 2 Mb (2.5 Mb in sample 1453 and 2.1 Mb in sample 53) (Figure 6.2 and 6.3). The region deleted on chromosome 7 of patient 1453 has not been reported as entirely deleted in either the DGV Gold Standard or the GnomAD-SV database, while the deletion of the area affected on chromosome 16 of proband 53 has been reported in 3 samples of the two population databases. The high pLI score genes involved in these two large CNVs were *KBTBD2* (pLI score = 1) and *XYLT1* (pLI score = 0.92). In population databases, the first has never been reported as completely deleted, while the second gene has been listed as deleted only in the 3 large deletions just mentioned. *XYLT1* is involved in bone development (Pönighaus *et al.*, 2007), while there is not a known function in humans for *KBTBD2*. According to the Protein Atlas database, both the genes have low tissue specificity for their protein expression. Given this information, we could not associate the deletion of these two high pLI score genes to the phenotype of the patients.

Except for the deletion in patient 1141, we could not clearly associate any of the other deletions with male infertility. The deletion in patient 1141 on chromosome 10 and removed 55 kb (Figure 6.4). It removed a large coding region of the *TDRD1* gene, which has 25 exons and a pLI score of 1. The region involved has not been listed as deleted in the population databases, and only two single exon losses have been reported in the *TDRD1* gene. The expression of this gene is high at both protein and RNA levels in the human testis. In the discussion of this chapter, I explore the information available in the literature that suggests that the impairment of *TDRD1* might play a role in the origin of patient 1141's infertility.

Finally, it is worth mentioning that a 1.8 Mb rare deletion was identified on chromosome 15 of proband 227 (Figure 6.5). The CNV encompassed 21 genes; however, none of them had a pLI score > 0.9. The region involved has not been reported as deleted in the samples of the DGV Gold Standard or GnomAD-SV database. This deletion and the two described above were the only ones identified in the cohort with a size > 1 Mb.

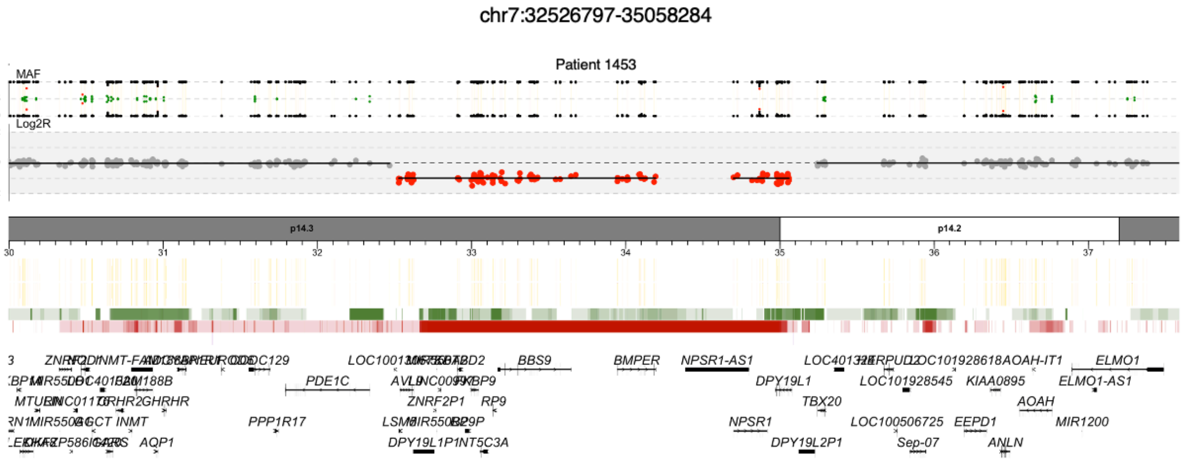


Figure 6.2. CNV plot of a 2.5 Mb large deletion identified in patient 1453. The deletion removed a single copy of 19 genes. The MAF track shows absence of heterozygous SNPs (MAF = 0.5) within the deleted region, which indicates loss of heterozygosity.

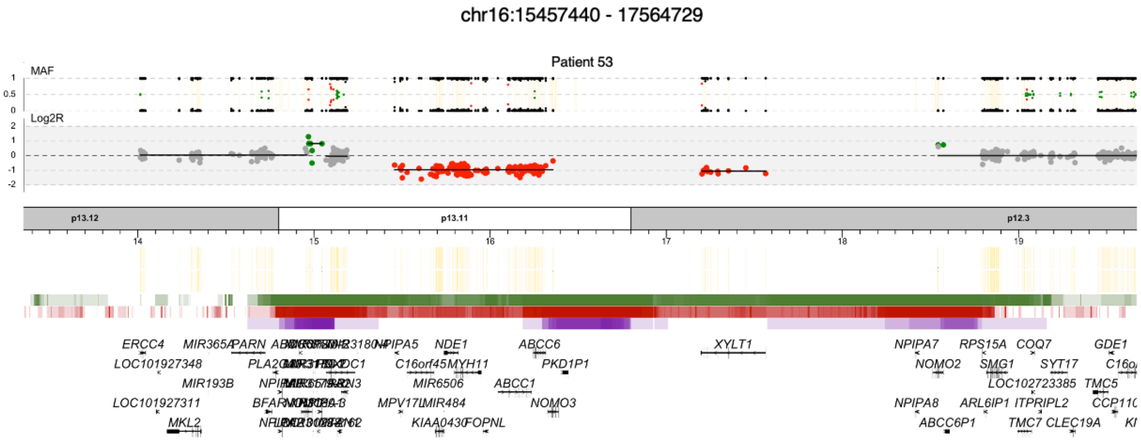


Figure 6.3. CNV plot of a 2.1 Mb large deletion identified in patient 53. The deletion removed a single copy of 14 genes.

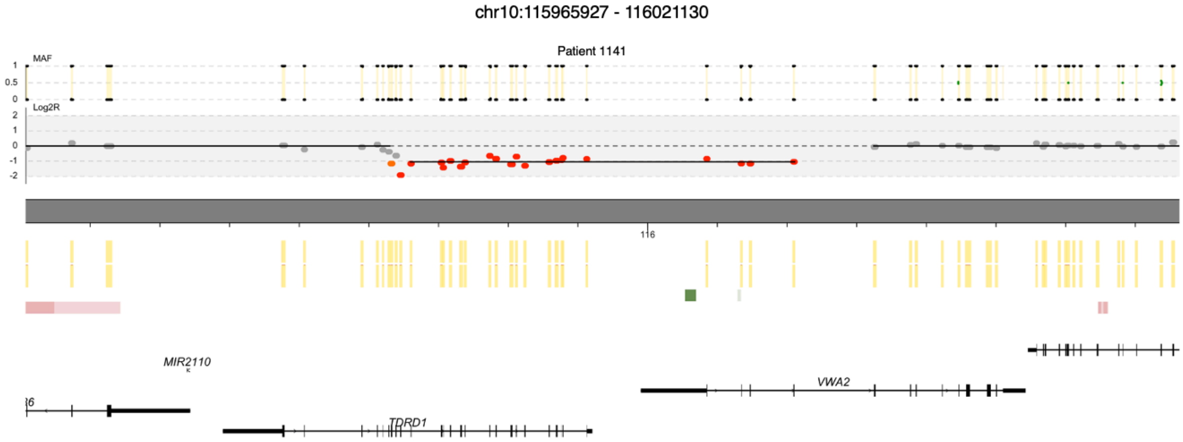


Figure 6.4. CNV plot of a rare large deletion identified in patient 1141. The deletion involved 2 genes. The MAF track shows absence of heterozygous SNPs (MAF = 0.5) within the deleted region, which indicates loss of heterozygosity.

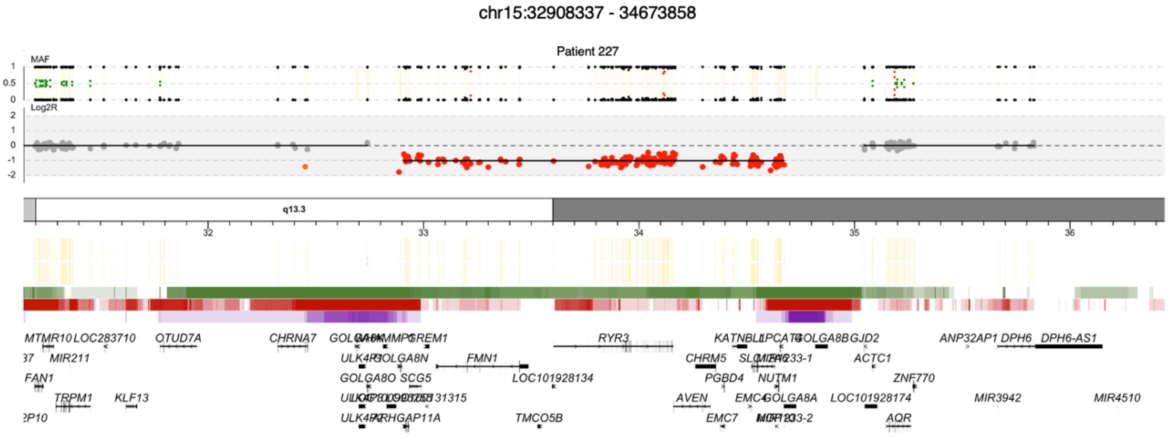


Figure 6.5. CNV plot of a 1.8 Mb large deletion identified in patient 227. The deletion removed a single copy of 21 genes. No genes involved in this deletion had a pLI score > 0.9. The MAF track shows absence of heterozygous SNPs (MAF = 0.5) within the deleted region, which indicates loss of heterozygosity.

6.3.2 Autosomal duplications

Duplications are more difficult to interpret compared to the deletions, as mentioned in other chapters. This is complicated further for a singleton cohort of patients, where we do not know the parent of origin of the variants or whether the gains occurred *de novo*. As previously done, I selected large and rare autosomal duplications that encompassed at least one gene with a pLI score > 0.9. These duplications might impair the function of these genes producing additional copies or might disrupt the gene sequence with the position of their breakpoints. In total 9 duplications with these features were identified (Table 6.2).

Patient	Genomic location (GRCh37)	Size	Probes	Genes	Genes with a pLI score > 0.9	Genes with a pLI score > 0.9 disrupted by the CNV breakpoints
23	chr10:12708659-12940755	232 kb	10	*CAMK1D - CCDC3 - MIR548Q*	*CAMK1D*	*CAMK1D - CCDC3*
220	chr22:20918661-20941067	22 kb	12	*MED15*	*MED15*	*MED15*
760	chr16:74485878-74504047	18 kb	12	*GLG1*	*GLG1*	*GLG1*
963	chr1:201915132-201984527	69 kb	26	*ELF3 - ELF3-AS1 - LMOD1 - MIR6740 - RNPEP - SNORA70H - TIMM17A*	*ELF3*	*LMOD1*
1860	chr1:27057567-27212700	155 kb	31	*ARID1A - GPN2 - PIGV - SFN - ZDHHC18*	*ARID1A*	*ARID1A - GPN2*
1832	chr2:241621572-241706514	85 kb	34	*AQP12A - AQP12B - KIF1A - LOC285191*	*KIF1A*	*KIF1A*
227	chr1:228547222-228879563	332 kb	48	*BTNL10 - DUSP5P1 - H2AW - H2BU1 - H3-4 - MIR4666A - MIR6742 - OBSCN - RHOU - RNA5S1 - RNA5S10 - RNA5S11 - RNA5S12 - RNA5S13 - RNA5S14 - RNA5S15 - RNA5S16 - RNA5S17 - RNA5S2 - RNA5S3 - RNA5S4 - RNA5S5 - RNA5S6 - RNA5S7 - RNA5S8 - RNA5S9 - RNF187 - TRIM11 - TRIM17*	*TRIM11*	*OBSCN*
1043	chr4:151682859-152070789	388 kb	52	*LRBA - RPS3A - SH3D19 - SNORD73A - SNORD73B*	*RPS3A*	*LRBA - SH3D19*
1617	chr16:28841103-28998290	157 kb	102	*ATP2A1 - ATP2A1-AS1 - ATXN2L - CD19 - LAT - MIR4517 - MIR4721 - NFATC2IP - RABEP2 - SH2B1 - SPNS1 - TUFM*	*ATXN2L - CD19 - SH2B1*	-

Table 6.2. Proband, genomic coordinates, size, and genes involved in the rare large duplications encompassing at least one gene with a pLI score > 0.9 identified in the 142 patients.

In 4 of the duplications prioritised (in patients 963, 227, 1043 and 1617) the genes disrupted by the breakpoints were not the same as the ones with a high pLI score, while for the other 5, the genes coincided.

The duplication found on chromosome 10 in patient 23 disrupted the sequence of the *CAMK1D* gene (pLI score = 1). A maternally inherited duplication of this gene was also found in proband 1697 (Figure 6.6). The exome enrichment kit used for samples 23 and 1697 did not include a target for the first exon of the gene, as can be seen in Figure 6.6. For this reason, we could establish whether the gene was duplicated in proband 1697 in its entirety. However, we could determine that *CAMK1D* was partially duplicated in patient 23, and that one of the breakpoints of the duplication disrupted the coding sequence of this gene. This is demonstrated by the fact that one target of *CAMK1D* in this patient has a log2R = 0 indicating the presence of two alleles (Figure 6.6). The coding region of *CAMK1D* has never been listed as duplicated or deleted in the population databases examined. *CAMK1D* functions are related to dendritic growth in the hippocampal neurons and granulocytes calcium-mediated function regulation (Verploegen *et al.*, 2005; Kamata *et al.*, 2007). Its protein expression has low tissue specificity, according to the Protein Atlas database. Based on the literature, we could not associate the impairment of this gene with the patient's phenotype.

The only gene that could be linked with spermatogenesis and testis functions was *SH2B1* (pLI score = 0.97). This gene was entirely duplicated in patient 1617 as part of a larger 157 kb in size duplication (Figure 6.7). KO mice for this gene have shown male and female infertility as well as other age-dependent problems such as insulin resistance and glucose intolerance (Ohtsuka et al., 2002; Duan et al., 2004). In humans, the gene has a medium expression in the testis, according to the Protein Atlas database. In the GnomAD-SV database and the DGV Gold Standard, the region duplicated in patient 1617 has been listed as deleted 5 times and 7 times as entirely duplicated. Despite the role of the gene in mice fertility, we could not link this variant to the patient's infertility since it is unclear what the effect of a complete duplication of *SH2B1* may be.

Given the information from the CNV analysis and the gene information publicly available, we could not associate these duplications with the patients' disease, and their consequences are unknown at present.

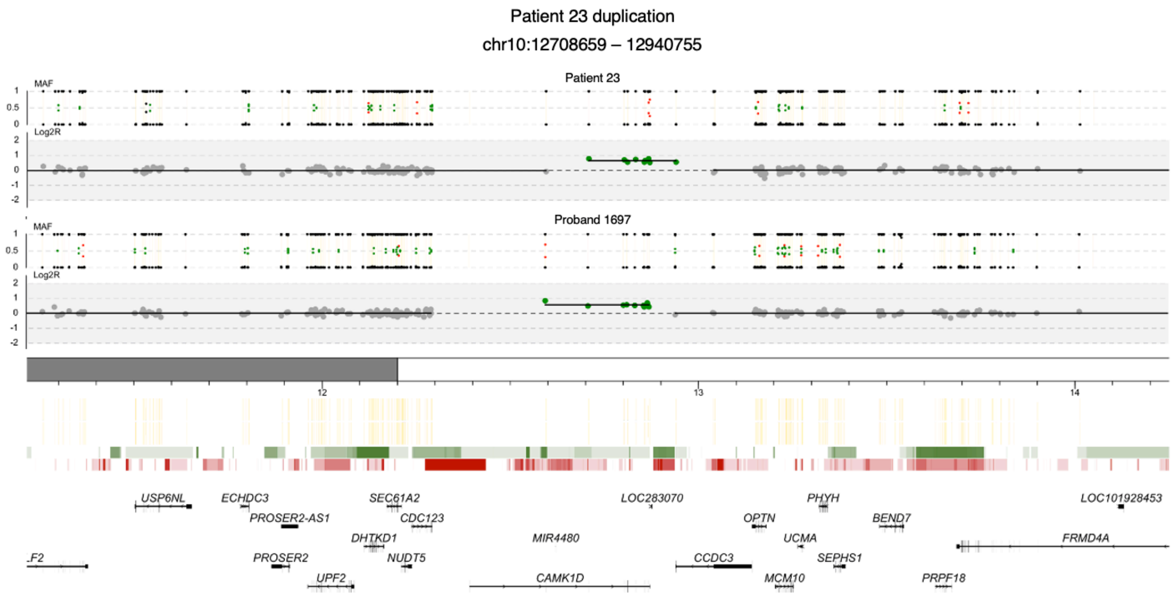


Figure 6.6. CNV plots of a rare large duplication identified in patient 23 and proband 1687. The one in patient 23 had a breakpoint within the coding region of the gene, and at least one exon of the gene was not involved in the gain. The exome enrichment kit used to sequence the DNA of the two patients did not cover the first exon of *CAMK1D*. Therefore, we do not know its copy-number status.

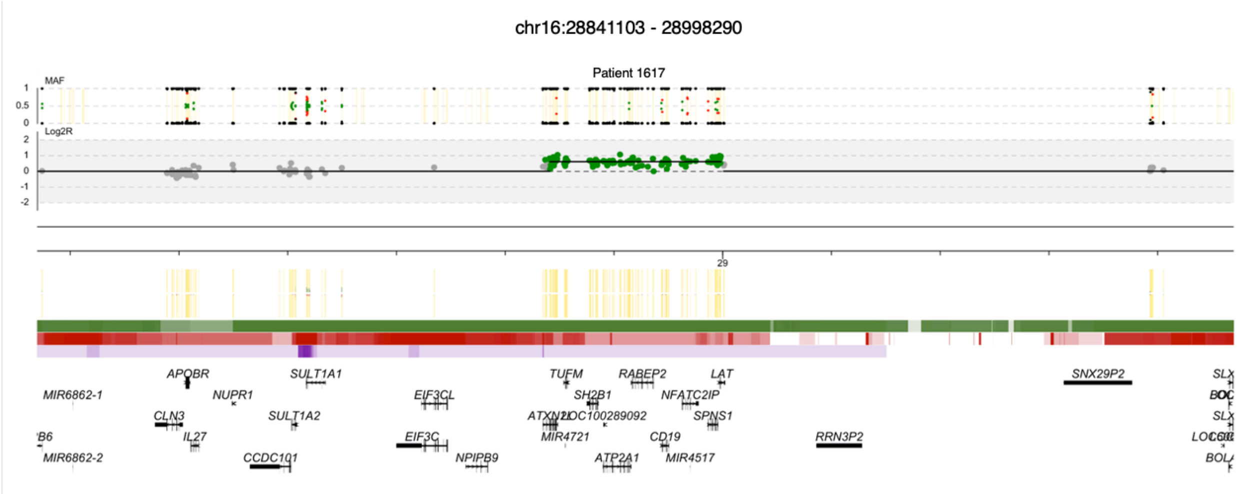


Figure 6.7. CNV plot of a rare large duplication identified in patient 1617. None of the 12 genes encompassed was disrupted by the breakpoint.

6.3.3 CNVs on the sex chromosomes

On the sex chromosomes of the 142 patients of the cohort, only 9 duplications were detected. Amongst these, I selected those that were rare and large. A total of 5 duplications were prioritised (Table 6.3), 4 on chromosome X and one on chromosome Y.

Patient	Genomic location (GRCh37)	Size	Probes	Genes	Genes disrupted by the CNV breakpoints
20	chrX:2707609-2799330	92 kb	12	*GYG2, XG*	-
215	chrX:2707609-2799330	92 kb	12	*GYG2, XG*	-
53	chrX:2707609-2799330	92 kb	12	*GYG2, XG*	-
6	chrX: 6968261-7894236	926 kb	15	*MIR4767, PUDP, STS, VCX, PNPLA4*	-
1635	chrY:14821305-14969662	148 kb	41	*USP9Y*	*USP9Y*

Table 6.3. Proband, genomic coordinates, size, and genes involved in the rare large duplications on the sex chromosomes identified in the 142 patients. 4 were located on chromosome X, of which three had recurrent breakpoints. One occurred on chromosome Y.

Three duplications, found in patients 20, 215 and 53, had the same breakpoints and consequently the same size (i.e., 92 kb and 12 sequencing probes). These variants involved the *XG* gene partially and the *GYG2* gene in its entirety. The CNV plots of the 3 duplications (Figure 6.8) show that the targets for the first exons of the *XG* gene are absent. A similar CNV was identified on chromosome X of proband 1636 from the trio cohort (Figure 6.9). This duplication involving *GYG2* and partially *XG* had an unclear inheritance since the same CNV was found in father 1636, but not in mother 1636, who should have transmitted the X chromosome. These duplications might be a technical artefact considering that they are at the boundaries of a highly complex repeat-rich region or might mask a more complex structural rearrangement that cannot be characterised by short-read sequencing. In the discussion, I explain the characteristics of this complex region.

The duplication found in patient 6 was 926 kb in size and overlapped with the maternally inherited duplication identified in proband 368 described in the previous chapter (Figure 6.10). The two duplications might have a similar size. However, since in patient 6, the sequencing targets for *VCX3A*, *VCX2*, and *MIR651* genes were not present, we could not establish their copy-number status. Also, the *VCX* gene in patient 6 had log2R = 0, which suggested the presence of only one copy. Nevertheless, considering that *VCX* is a multi-copy gene and the presence of SNPs with a MAF close to 0.5 (which represents heterozygous SNPs) within the *VCX* genomic region, it is possible that the number of copies of the gene in patient 6 is higher than one (Figure 6.10). Duplications in this region, as large as the one found in proband 368 and patient 6, have not been reported in the population databases examined. The possible role of duplications in this genomic region in the disease aetiology has been discussed extensively in the discussion of chapter 5 and is summarised in the discussion of this chapter.

The last duplication was identified in patient 1635. It was the only one detected on chromosome Y, and it was 148 kb large. The *USP9Y* gene was almost entirely duplicated, and only one target was not involved (Figure 6.11). No variants involving the coding region of this gene have been reported in the DGV Gold Standard or GnomAD-SV database. The Protein Atlas database does not contain protein expression data for *USP9Y*, while the RNA expression data shows low tissue specificity. This gene is part of the AZFa region. In the discussion, I debate the role of this gene in male infertility, based on the literature information available.

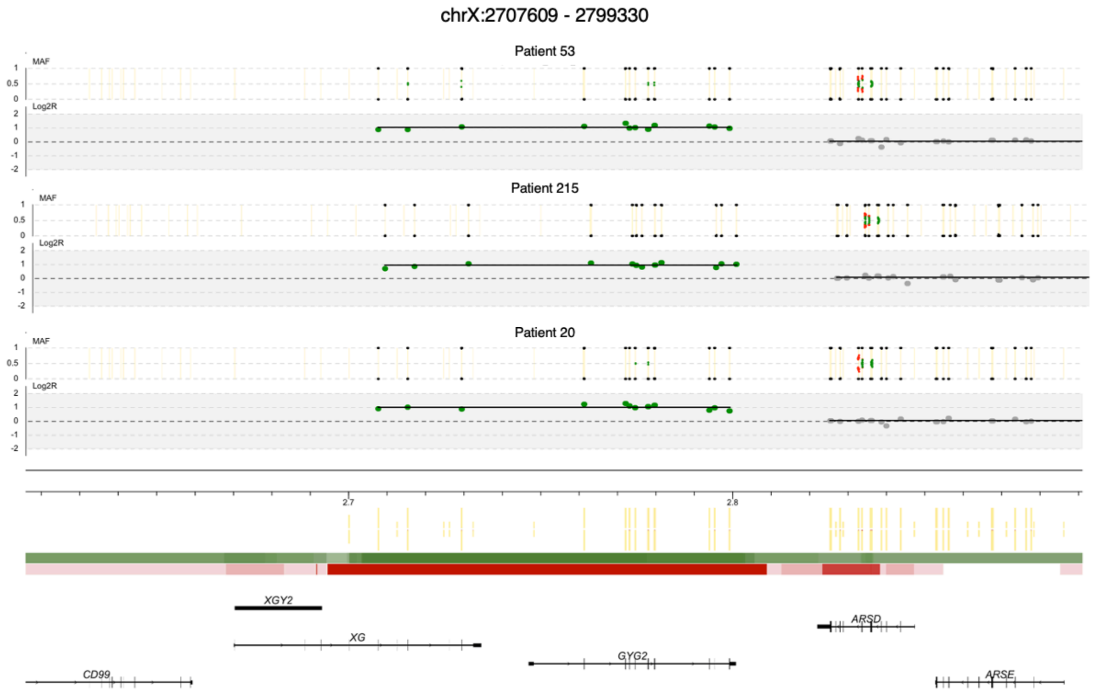


Figure 6.8. CNV plots of 3 identical rare large X-linked duplications identified in patient 53, 215, and 20. The duplications partially encompassed the *XG* gene. However, as visible from the CNV plots, there were no informative sequencing targets covering the first exons of *XG*. The same scenario was present for the gene *CD99* and continued for the entire telomeric region of the short arm of chromosome X.

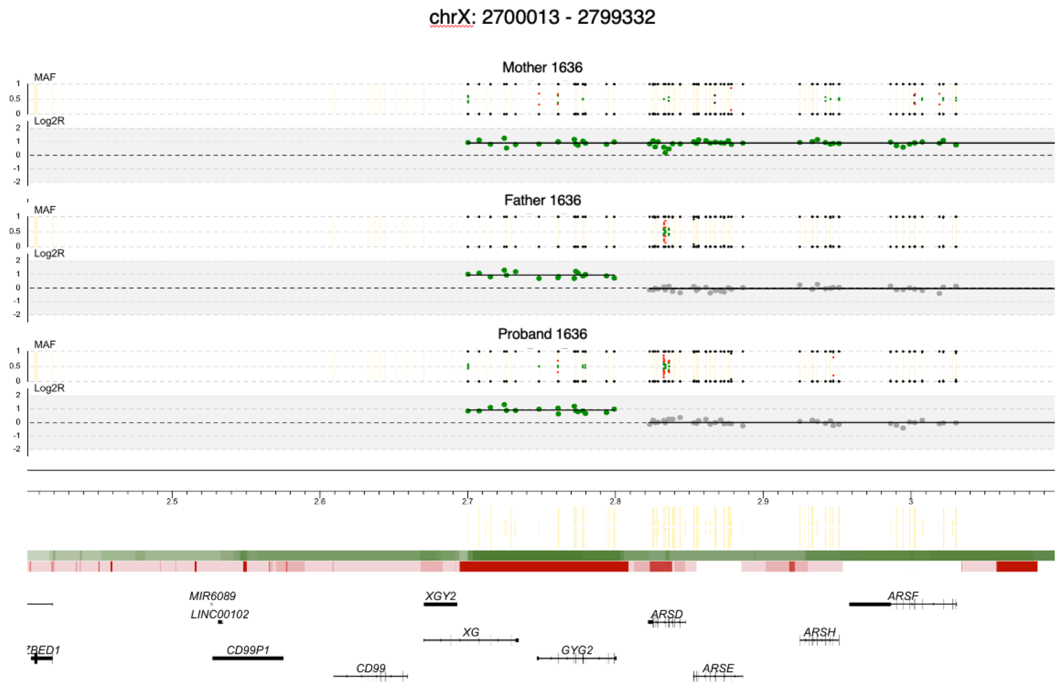


Figure 6.9. CNV plot of a rare large X-linked duplication identified in proband 1636 and overlapping the duplications found in patients 53, 215, and 20. The same duplication was present in father 1636, but it was not detected in mother 1636, who should have transmitted the chromosome X to the proband.

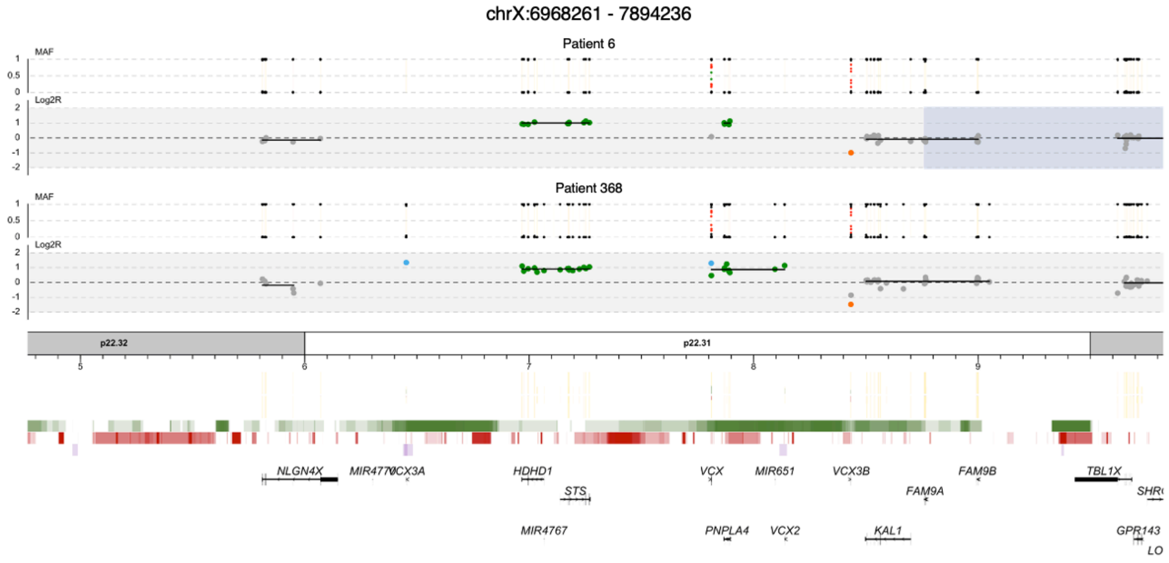


Figure 6.10. CNV plots of two overlapping rare large X-linked duplications identified in patient 6 and proband 368. In patient 6 there were no targets for the genes *VCX3A*, *MIR651* and *VCX2*.

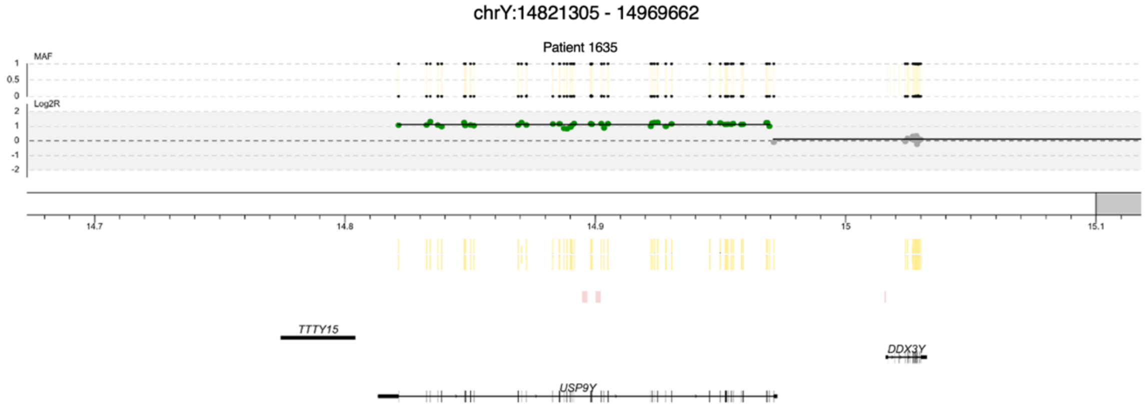


Figure 6.11. CNV plot of a rare large Y-linked duplication identified in patient 1635. The gene was almost entirely duplicated, and only one target was not involved. The gene is the first coding gene on the long arm of chromosome Y after the centromere.

6.3.4 CNVs overlapping novel dominant candidate male infertility genes

The cohort of patients presented in this chapter was affected by azoospermia or severe oligozoospermia, as was the cohort of trios investigated in chapters 4 and 5. To identify genes recurrently mutated in patients, I looked for variants involving the genes affected by the *de novo* CNVs identified in the trios (*QSER1, DEPDC7, TCP11L1, CSTF3, CSTF3-AS1, HIPK3, KIAA1549L, NXT2)* or by *de novo* point mutations classified as possibly causative of the disease or of unknown significance identified by my colleagues in the same cohort (98 *de novo* mutations in 96 different genes) (see Appendix B). Also, I looked for variants affecting the novel candidate dominant male infertility genes identified by the analysis of maternally inherited CNVs (*AKT1* and *PSME4*) in chapter 5. Amongst all the CNVs identified in the 142 patients, there was no variant that overlapped with any of the genes mentioned.

6.4 Discussion

In this chapter, I analysed the CNVs detected in a cohort of 142 patients affected by azoospermia or severe oligozoospermia, for which parental samples were not available.

The average number of total CNVs per sample identified in this cohort (i.e., 10) was comparable to the average number identified in the trio probands (i.e., 11) (see chapter 4), and 667 deletions and 702 duplications were detected in total. The rare large CNVs encompassing a high pLI score gene were 5 deletions and 12 duplications in total in these patients.

Amongst these CNVs, the 55 kb large deletion on chromosome 10 of patient 1141 might play a role in male infertility. The *TDRD1* gene with a pLI score of 1 was deleted in this patient. According to the Protein Atlas database, *TDRD1* has high expression in the human testis exclusively. Its expression has been shown to be lower in testicular biopsies of patients with maturation arrests in comparison to patients with obstructive azoospermia (Babakhanzadeh *et al.*, 2020). Also, a case-control study associated two *TDRD1* polymorphisms to reduced risk of spermatogenic impairment in the Han Chinese population (Zhu *et al.*, 2016), suggesting that the gene is involved in spermatogenesis. In mice, the *TDRD1* gene plays an essential role in preserving the germline integrity, regulating the movement of transposable elements during meiosis (Chuma *et al.*, 2006; Reuter *et al.*, 2009; Vagin *et al.*, 2009), and it is required for male fertility (Chuma *et al.*, 2006). Homozygous mutations in other two genes from the TDRD family, *TDRD9* and *TDRD7*, have been recently associated with non-obstructive azoospermia (Arafat *et al.*, 2017; Tan *et al.*, 2019). This suggests that the TDRD gene family may be enriched in genes involved in spermatogenesis. For this reason, we hypothesise that the *TDRD1* deletion in patient 114 may be possibly pathogenic for male fertility. However, *TDRD1* exact function in humans needs to be clarified by further investigations.

Interestingly, we identified 3 deletions larger than 1.5 Mb in size in the patients of this cohort, something not seen in the trios, and uncommon in the general population. The 3 deletions could not be associated with severe male. One deletion, in patient 227, did not encompass any gene with a pLI score > 0.9, while the other two, in patients 1453 and 53, involved each one likely haploinsufficient gene, *KBTBD2* (pLI score = 1) and *XYTL1* (pLI score = 0.92), respectively. In the literature, these two genes have not been reported as associated with spermatogenesis or testis function. *KBTBD2* has no known function in humans, while *XYLT1* is involved in bone development (Pönighaus *et al.*, 2007). Two articles in 2014 reported recessive mutations in the *XYLT1* gene associated with a short stature syndrome (Bui *et al.*, 2014; Schreml *et al.*, 2014). These findings and the 3 deletions of the *XYLT1* gene reported in the population databases suggest that the gene is probably tolerant to loss-of-function mutations, even though it has a high pLI score. We cannot exclude, though, that the 3 large deletions may play a role in the infertility of the respective patients, but it is unlikely that they act in a dominant fashion. Additional mutations in the coding regions of the genes on the remaining allele might alter the spermatogenesis. For instance, the deletion found in patient 1453 involved the gene *DPY19L1*, whose function is unknown. This gene is a paralog of *DY19L2*, a gene that causes globozoospermia when impaired by homozygous variants (Harbuz *et al.*, 2011; Koscinski *et al.*, 2011; Elinati *et al.*, 2012), and might be as well involved in sperm production. Also, the deletion found in patient 53 affected the gene *CEP20*, coding for a centrosomal protein involved in cilia biogenesis (Sedjaï *et al.*, 2010). A STRING analysis shows that this gene interacts with *CEP135* (Figure 6.12), a gene involved in centriole biogenesis and carrying homozygous mutations in patients with multiple morphological abnormalities of the sperm flagella (Y. W. Sha *et al.*, 2017). Thus, *CEP20* might also be involved in the production of the sperm tail. Although recessive mutations are less likely to explain male infertility cases in outbred cohorts compared to consanguineous families, follow-up studies should aim to detect a combination of heterozygous CNVs and loss-of-function mutations and identify novel recessive disease candidates. This goes beyond the scope of this thesis, which is focused on the identification of novel dominant candidate male infertility genes.

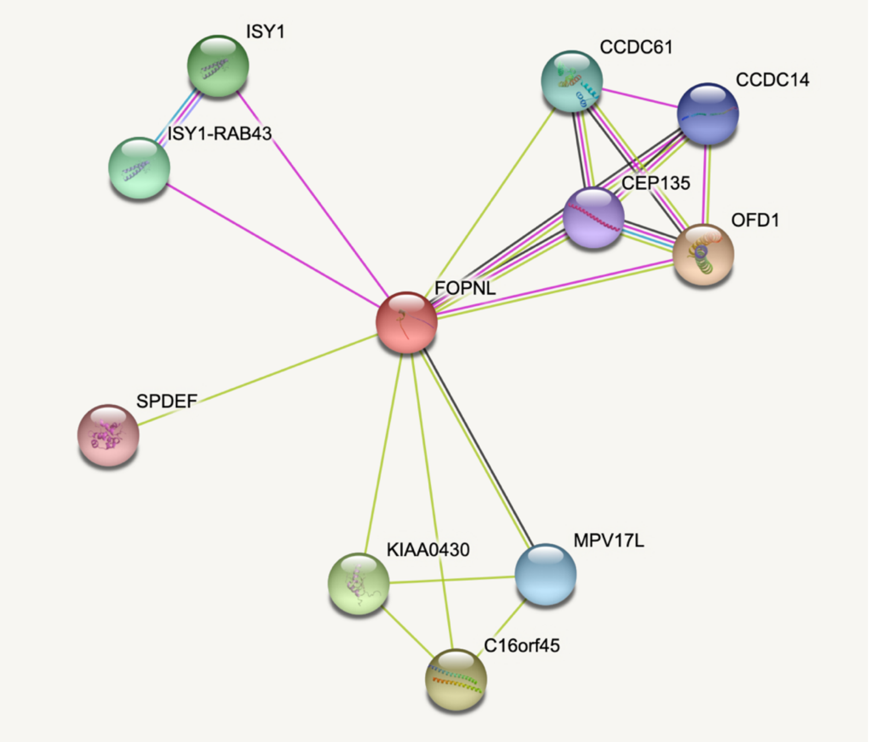


Figure 6.12. STRING analysis for *CEP20* (*FOPNL*). The magenta lines indicate known experimentally determined interactions between the proteins (e.g., between *CEP20* and *CEP135*). Black lines and yellow lines indicate information regarding co-expression and information derived from text mining, respectively.

From the information provided by the CNV analysis and the literature data publicly available, we could not associate any of the rare large duplications encompassing a high pLI score gene to severe male infertility. We cannot exclude a role of these CNVs in male infertility, but the information provided by the WES data, which do not include the number of additional copies, or their position and orientation, did not allow an accurate assessment of the likely pathogenicity of these variants.

In the patients of this cohort, only 5 rare and large duplications on the sex chromosomes were found. Notably, the sex chromosomes did not carry any deletion. This suggests intolerance of the sex chromosomes towards deletions, as seen in the trio cohort, where only one loss was detected in a larger cohort of patients. This is expected considering that males have only one copy of the X and Y chromosome, and a deletion would eliminate the unique copies of the genes involved. Also, all the patients underwent AZF deletions screening test before participating in our study, which decreased the probability of identifying undetected losses on the Y.

Three rare large duplications on chromosome X had the same size and breakpoints position in different patients (patients 20, 215 and 53). These gains partially duplicated the gene *XG* and the entire gene *GYG2*. The copy-state status of the first exons of *XG*, as well as that of the entire upstream region of chromosome X, were not available. This is due to the fact that the first 3 exons of *XG* are part of the PAR1 region, a 2.6 Mb segment containing 24 genes that shares homology with the PAR1 region on chromosome Y (Helena Mangs and Morris, 2007). For this reason, the short reads generated by WES cannot be mapped accurately to either of these regions and consequently, they could not be reliably used for the CNV analysis. PAR1 regions are required for male fertility in humans and mice since PAR1 and PAR2 regions, located on the telomeres of X and Y chromosomes, pair and recombine during meiosis, acting similarly to the autosomes (Gabriel-Robez *et al.*, 1990; Burgoyne *et al.*, 1992; Mohandas *et al.*, 1992; Helena Mangs and Morris, 2007). A similar duplication was found in proband 1636 from the trio cohort. The same duplication was found in the respective father but not in the mother, who should have transmitted the X chromosome to the proband. These four duplications might be technical artefacts possibly caused by the poor mapping of the short reads at the boundary of the PAR1 region or might mask a more complex genomic rearrangement involving the telomeres. WES data cannot provide further information, and additional experiments, ideally with optical mapping or long-read sequencing technologies, are required to characterise this region in the patients. Lastly, it is worth mentioning that patient 53, other than being part of the group of patients carrying these duplications on the X chromosome, also carries the 2.1 Mb deletion involving *XYLT1* described earlier in this chapter.

The chromosome X duplication in patient 6 affected the same region duplicated in mother and proband 368 (described in the previous chapter). Both duplications encompassed the multi-copy gene *VCX*, expressed in male germ cells during spermatogenesis (Lahn and Page, 2000; Zou *et al.*, 2003; Van Esch *et al.*, 2005). The possible role of the *VCX* gene in male infertility has been discussed extensively in the discussion of the previous chapter. In summary, the consequences of variants in this region of chromosome X (Xp22.31) have been debated for more than 10 years (Li *et al.*, 2010), but only recently in relation to male infertility. Ji *et al.* in 2016 observed an increase of *VCX* duplications in infertile men compared to controls and *in vitro* experiments showed that copy-number gains of *VCX* lead to apoptosis (Ji *et al.*, 2016). Zhuang *et al.* recently reported a duplication involving *VCX* in a fertile father (Zhuang *et al.*, 2019). None of these studies characterised the exact number of *VCX* copies in the individuals analysed. Further studies should be conducted to determine whether the number of copies of *VCX* influences male fertility and if the increase in the number of copies correlates with a decrease in fertility.

The only CNV detected on chromosome Y was a duplication that encompassed the *USP9Y* gene in patient 1635. This gene is part of the AZFa region, a locus known to be involved in male infertility (Krausz and Riera-Escamilla, 2018). This gene has been associated with spermatogenesis impairment for several years (Sun *et al.*, 1999; Foresta *et al.*, 2000; Lee *et al.*, 2003), but more recent studies have questioned this assumption. For instance, Krausz *et al.* reported, in 2006, the natural transmission of two *USP9Y* deletions in two independent families and suggested a "fine tuner" role of this gene for spermatogenesis (Krausz *et al.*, 2006). Subsequently, in 2009, Luddi *et al.* reported another case of complete deletion of the gene in a normozoospermic man as well as in his brother and father and proposed the gene as dispensable for male fertility (Luddi *et al.*, 2009). Other articles supporting this thesis have been subsequently published (Tyler-Smith and Krausz, 2009; Alksere *et al.*, 2019). This suggests that the well-known AZF regions require a detailed characterisation that establishes which genes are essential for male fertility and which are dispensable. In patient 1635, the CNV data showed that the last exon of the gene is not duplicated; therefore, the breakpoint of the gain disrupted its sequence. *USP9Y* is the first gene downstream of the centromere, and the copy-number status of the upstream region is unknown. As the partial duplications of *XG* described above, this gain might indicate a more complex structural variant that cannot be characterised by WES. However, we know that the AZF regions do not carry large deletions, as all the patients of the study were previously tested for AZF deletions prior to inclusion in the study. If the duplication involves only the *USP9Y* gene, it is unlikely to play a role in the patient's infertility.

Lastly, I looked for CNVs that encompassed genes affected by *de novo* mutation of unclear significance or classified as possibly causative that were identified in the trio cohort by my colleagues, as well as the genes involved in the *de novo* CNVs. Additionally, I looked for CNVs affecting the candidate male infertility genes identified with the analysis of maternally inherited CNVs. No CNVs covered these genes, which is consistent with isolated severe male infertility being highly heterogeneous. Larger studies are required to find recurrently mutated genes in male infertility patients. Our research group is part of the International Male Infertility Genomics Consortium (IMIGC: www.imigc.org), which includes several groups working on the genetics of male infertility in three different continents. This type of initiative is essential to access larger cohort of patients, find recurrently mutated genes and design projects that investigate different aspects of the disease. This will help to tackle the high genetic heterogeneity of the disease and will facilitate the discovery of novel disease genes.

6.5 Conclusions

In this study, I analysed the CNV burden in a cohort of 142 patients affected by severe idiopathic male infertility. For these patients, parental samples were not available. Therefore, we could not establish the inheritance of the genetic variants identified. Despite this limitation, we were able to identify 3 rare deletions larger than 1.5 Mb of unknown significance and one rare large possibly pathogenic deletion encompassing a high pLI score gene that might be involved in spermatogenesis. The analysis of the CNVs on the sex chromosomes revealed another duplication in Xp22.31, a region that had already been found to be duplicated in a proband from the trio cohort and containing at least one gene involved in male fertility. This study confirms that WES is a valid method to study the CNVs in cohorts of patients and highlights the value of investigating cohorts of singletons when parental samples are not available.

Chapter 7. CNVs in Patients With Idiopathic Qualitative Sperm Defects

7.1 Introduction

Hitherto I investigated the role of rare CNVs in cohorts of patients affected by quantitative defects of the sperm production. Male infertility, however, can also be characterised by several forms of qualitative defects of the sperm morphology or function. Asthenoteratozoospermia is a pathogenic male infertility phenotype characterised by both reduced sperm motility and morphological defects of the spermatozoa (Esteves *et al.*, 2018). Multiple Morphological Abnormalities of the Flagella (MMAF) are often the underlying morphological cause of asthenoteratozoospermia, and this denomination was first proposed by Ben Khelifa *et al.* in 2014 to describe the absent, short, bent, coiled and/or irregular-calibre flagella of patients with the disease. (Ben Khelifa *et al.*, 2014). The authors of that study described homozygous *DNAH1* variants in 5 unrelated infertile men with asthenoteratozoospermia. *DNAH1* is required for normal biogenesis of the axoneme, a core component of the sperm flagellum (Ben Khelifa *et al.*, 2014). In the following 5 years, a total of 18 autosomal genes associated with asthenoteratozoospermia with MMAF were identified, and homozygous mutations in these genes explained between 30 to 60% of the cases in different patients' cohorts (Coutton *et al.*, 2019; Liu *et al.*, 2019; Touré *et al.*, 2020). In 2020, 4 additional WES-based papers revealed bi-allelic variants in 4 novel MMAF genes: *DZIP1*, *DNAH8*, *MAATS1* and *CFAP58* (He *et al.*, 2020; Liu *et al.*, 2020; Lv *et al.*, 2020; Martinez *et al.*, 2020), increasing the total number of disease genes to 22. These findings demonstrate that genes involved in the development of the sperm tail are extremely important for fertility and pathogenic variants affecting them are often the cause of the disease in patients with asthenoteratozoospermia. Remarkably, X-linked genes associated with asthenoteratozoospermia with MMAF have not been identified. This is unexpected considering that this chromosome is enriched in genes involved in spermatogenesis (Vockel *et al.*, 2019) and that other X-linked genes have been associated with different male infertility phenotypes, as discussed in the previous chapters.

Most of the mutations found in genes associated with asthenoteratozoospermia with MMAF are homozygous SNVs. However, Tang *et al.* 2017 found also a 3.3 kb heterozygous deletion in combination with a deleterious missense mutation impairing the function of the *CFAP43* gene (Tang *et al.*, 2017). Notably, this research group decided to perform an additional microarray test to detect CNVs, despite having sequenced the DNA of the patients with WES. This highlights the general distrust towards the sensitivity of CNV detection from WES data, which is the explanation that Tang *et al.* provided. In the previous chapters, I demonstrated the ability of WES to detect rare CNVs and even *de novo* variants of potential clinical interests in male infertility patients. Similarly, in 2018, Kherraf *et al.* demonstrated the usefulness of performing CNV detection from WES data of patients with asthenoteratozoospermia with MMAF. Using this method, they identified an 8.4 kb homozygous deletion in *WDR66* (also named *CFAP251*) in 7 patients with the disease, part of a cohort of 78 affected individuals (Kherraf *et al.*, 2018).

In this chapter, I use the G4BP CNV detection workflow in combination with the WES data of 24 asthenoteratozoospermia patients to identify CNVs of potential clinical interest.

7.2 Aims

In this chapter, I aimed to:

- Assess the potential pathogenicity of rare large autosomal and X- and Y-linked CNVs identified in the 24 patients with asthenoteratozoospermia.
- Identify CNVs that encompass known MMAF genes.
- Identify novel candidate male infertility genes impaired in asthenoteratozoospermia cases.

7.3 Results

The WES data of 24 patients from Australia affected with asthenoteratozoospermia was analysed with the G4BP CNV tool to identify possibly deleterious CNVs. In total, 155 CNVs were identified and 8 CNVs per sample on average. A total of 83 deletions and 64 duplications were identified (Figure 7.1). When prioritising rare (present in < 1% of the samples of the DGV Gold Standard) CNVs, the number of duplications and deletions were similar (i.e., 15 losses and 16 gains), and the number of rare and large (comprising >= 10 probes) variants was equal (3 per CNV type).

On the sex chromosomes, only X-linked CNVs were identified. Specifically, 3 rare deletions and 3 rare duplications were found, but only 1 loss was rare and large (Figure 7.1).

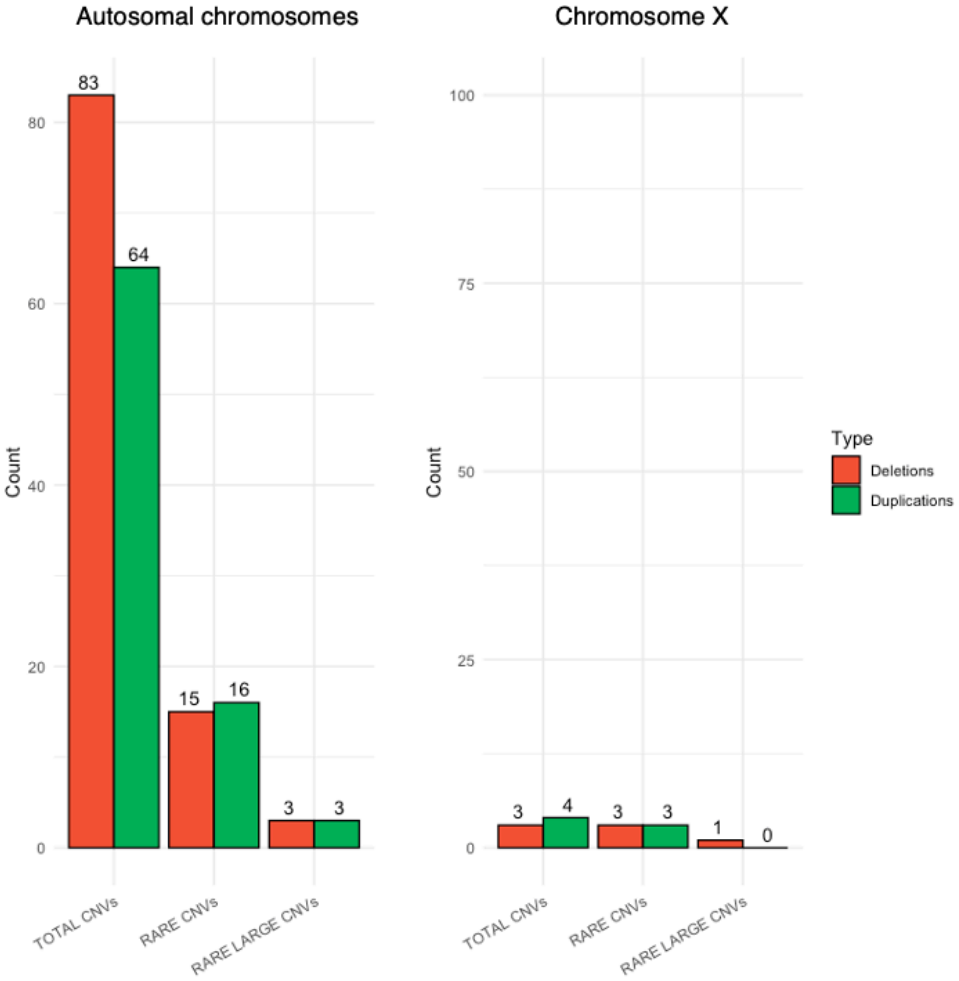


Figure 7.1. Number of CNVs identified in 24 patients with asthenoteratozoospermia. The number of total deletions was higher that the number of duplications. Instead, the number of rare deletions and duplications was similar, and the number of rare large ones was equal. No CNV on the Y chromosome were detected. On the X chromosome, the number of deletions and duplications was similar, and only one rare large deletion was identified.

7.3.1 Rare and large CNVs

The number of rare large CNVs identified on the autosomes and the sex chromosomes of the 24 patients was 7 (Table 7.1). This category of CNVs is more likely to be deleterious than the rest of the CNVs since they are rare in the general population and encompass a larger fraction of the coding region. These CNVs included 4 losses, of which one was on the X chromosome, and 3 duplications. Amongst the variants detected on the autosomes, no homozygous deletions were identified, and none encompassed a gene likely to be haploinsufficient (pLI score > 0.9).

Patient	Genomic location (GRCh37)	Type	Size	Probes	Genes	Genes with a pLI score > 0.9
1903	chr2:97749416-97858866	Deletion	109 kb	34	*ANKRD36 - FAHD2B*	-
2093	chr12:130827061-130856219	Deletion	29 kb	20	*PIWIL1*	-
1903	chr15:43924321-43991363	Deletion	67 kb	14	*CATSPER2 - CKMT1A - PPIP5K1P1-CATSPER2*	-
2603	chrX:34147944-37312950	Deletion	3.2 Mb	46	*CFAP47 - FAM47A - FAM47B - FAM47C - FTHL18 - LOC101928627 - MAGEB16 - PRRG1 - TMEM47*	*CFAP47*
2321	chr19:6890416-7083831	Duplication	193 kb	27	*ADGRE1 - ADGRE4P - FLJ25758 - MBD3L2 - MBD3L2B - MBD3L3 - MBD3L4 - MBD3L5 - ZNF557*	-
2649	chr3:136573227-136729141	Duplication	156 kb	12	*IL20RB - NCK1 - NCK1-DT - SLC35G2*	-
2163	chr9:39085657-39178433	Duplication	93 kb	12	*CNTNAP3*	-

Table 7.1. Proband, genomic coordinates, type of CNV, size, and genes involved in the rare large CNVs identified in the 24 patients.

The literature information revealed that 2 CNVs affected at least one gene with known function in spermatogenesis or male fertility.

One of these CNVs was a 29 kb large autosomal deletion on chromosome 12 in patient 2093 encompassing *PIWIL1* (Figure 7.2). This gene is exclusively expressed in the testis in humans, according to the Protein Atlas database. In mice, it is essential for regular spermatogenesis as it plays an important role in piRNA production (Kuramochi-Miyagawa *et al.*, 2004). The consequences of *PIWIL1* heterozygous loss-of-function mutations are still debated (Gou *et al.*, 2017; M. S. Oud *et al.*, 2021) (see discussion). Only two losses, deleting the untranslated region (UTR) and the promotor of the gene, respectively, have been described in the population databases examined (see Chapter 2 for methods). Its pLI score is equal to 0, indicating that the gene is unlikely intolerant to loss-of-function variants.

The second CNV encompassing a gene with a function in spermatogenesis was a 67 kb large deletion on chromosome 15 of patient 1903 (Figure 7.3). It removed, amongst others, a single copy of *CATSPER2*. In humans, this gene is highly expressed at the RNA level in the testis, according to the Protein Atlas database. *CATSPER2* is involved in physiological responses crucial for successful fertilisation, such as sperm hyperactivation, a process that increases sperm motility and allows the sperm to penetrate the zona pellucida of the oocyte (Lishko, Botchkina and Kirichok, 2011). The same function has been described in mice, where *CATSPER2*-null specimens are infertile despite a normal sperm count (Quill *et al.*, 2003). A total of 603 deletions involving the coding region of *CATSPER2* have been listed in population databases. As expected from such a high number of variants in the general population, the pLI score of *CATSPER2* is equal to 0.

Other than the two described deletions, the other CNVs did not comprise genes known to be relevant for male fertility according to the literature. It is worth mentioning, though, that a 3.2 Mb large hemizygous deletion was identified on chromosome X of patient 2603 (Figure 7.4). The loss removed 9 genes. There is no well-established function for any of these genes, but *TMEM47* is thought to regulate the cell epithelial cell junctions in vertebrates (Dong and Simske, 2016). Amongst the deleted genes, *CFAP47* is the only gene with a pLI score above 0.9 (i.e., pLI score = 0.99), which suggests high intolerance to loss-of-function variants. The Protein Atlas database only includes the RNA expression data for this gene. *CFAP47* RNA has been found in several tissues including brain, testis, and fallopian tube. *CFAP47* is part of the Cilia and Flagella-associated protein family. In this group of genes, there are 8 others found mutated in patients with asthenoteratozoospermia with MMAF, for example, *CFAP43* and *CFAP44* (Tang *et al.*, 2017; Y.-W. Sha *et al.*, 2017). The region encompassed by this CNV has not been reported as entirely deleted in the individuals of the two population databases examined, nor deletions of the high pLI score gene *CFAP47* have been described.

This deletion was experimentally validated by a PCR assay performed by Dr Bilal Alobaidi. The PCR primers were designed to amplify 3 genomic loci within the predicted deleted region and 3 in the flanking genes (Figure 7.5). The PCR assay produced no PCR product for the region within the CNV breakpoints, while it amplified the regions in the flanking genes. These results confirmed the CNV and excluded the deletion of the flanking genes.

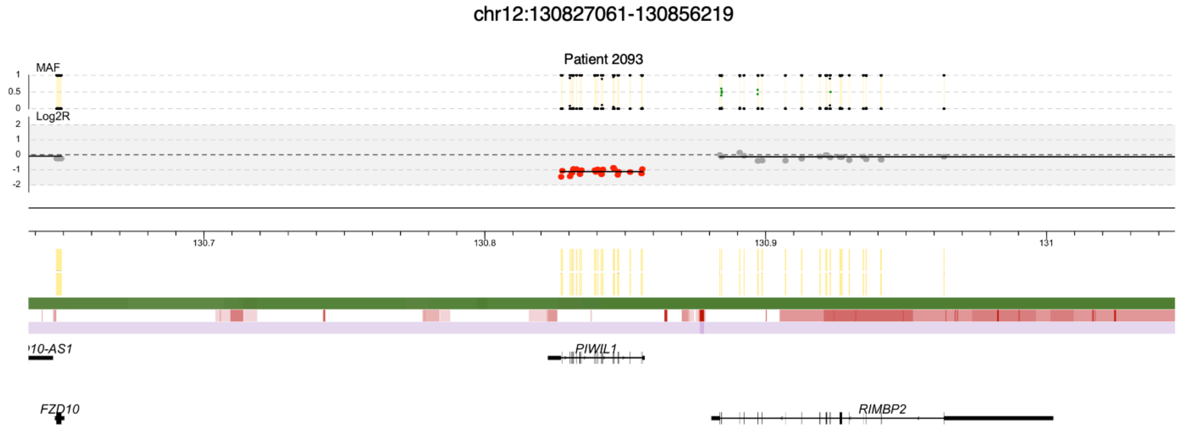


Figure 7.2. CNV plot of a rare large deletion identified in patient 2093. It removed a single copy of *PIWIL1* completely. The absence of heterozygous SNPs (MAF = 0.5) in the MAF tracks suggests loss of heterozygosity.

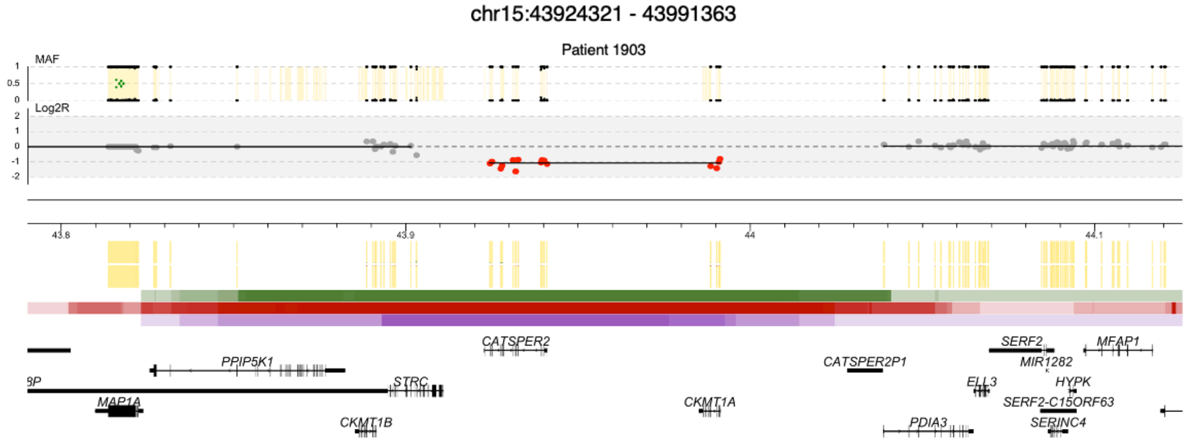


Figure 7.3. CNV plot of a rare large deletion identified in patient 1903. It removed completely a single copy of 2 protein-coding genes (and one pseudogene – not represented). The absence of heterozygous SNPs (MAF = 0.5) in the MAF tracks suggests loss of heterozygosity.

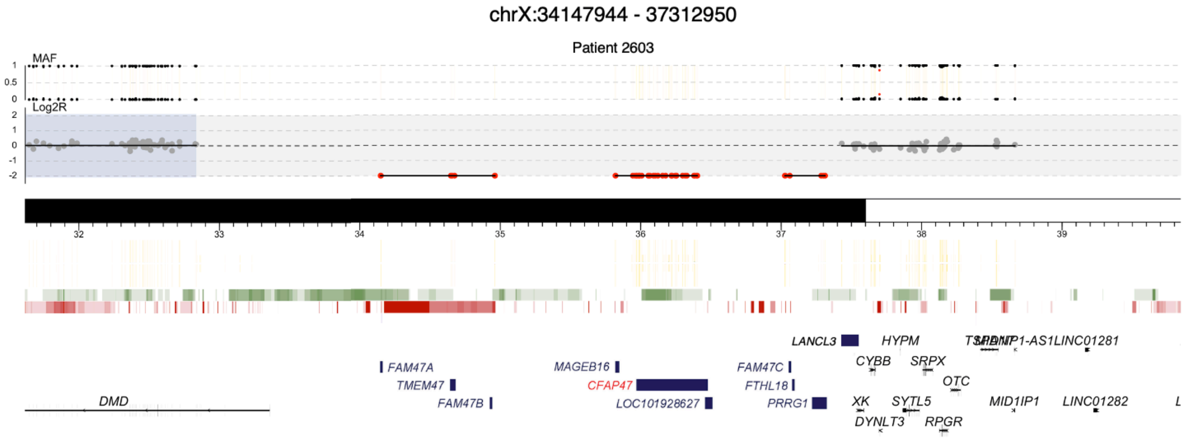


Figure 7.4. CNV plot of a rare large X-linked deletion identified in patient 2603. It removed the unique copy of 9 genes completely. The complete absence of SNPs in the MAF track suggests the loss of the unique allele.

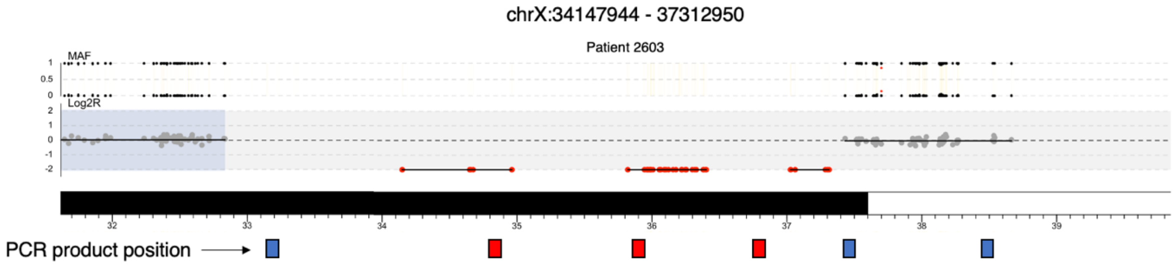


Figure 7.5. Position of the genomic regions included in the PCR assay performed to validate the X-linked deletion identified in patient 2603. The blue rectangles indicate the regions that were amplified, while the red ones indicate the absence of PCR product.

7.3.2 Additional patients carrying CFAP47 variants

After the identification of the hemizygous deletion in patient 2603, our group started a collaboration with Feng Zhang's group at Fudan University in Shanghai, China. This research group, with other colleagues in China and France, had identified likely pathogenic missense variants in the *CFAP47* gene in 3 patients affected by asthenoteratozoospermia with MMAF. They found decreased levels of *CFAP47* mRNA in the sperm of these men, compared to fertile controls with a real-time quantitative PCR. Mice with a *CFAP47* hemizygous frameshift mutation (*CFAP47-*/Y) were generated. This mutation was predicted to produce a premature termination of the protein product, and real-time quantitative PCR confirmed significantly reduced levels of *CFAP47* mRNA in the spermatozoa of the mutated mice compared to controls. The mutated mice were infertile and showed reduced sperm motility and MMAF. In addition, our colleagues suggested that *CFAP47* interacts with *CFAP65*, a gene known to be associated with the asthenoteratozoospermia with MMAF (Zhang *et al.*, 2019; W. Li *et al.*, 2020). In fertile men, *CFAP65* protein is localised at the base of the flagella and at the equatorial segment of the sperm head, while in the sperm of men with likely pathogenic variants in *CFAP47*, it was found in the sperm acrosome. Also, abnormal localisation of *CFAP65* was not found in other patients with a mutation in the known disease genes *DNAH8*, *SPEF2* and *CFAP58*.

These findings suggested an important role of the *CFAP47* gene for normal sperm production, possibly through its interaction with *CFAP65*. The impairment of *CFAP47* is likely the cause of the pathogenic phenotype in patient 2603 as well as in the 3 patients described by our colleagues. Also, despite patient 2603 being classified as affected by asthenoteratozoospermia patient without specific tail abnormalities characterisation, these results suggest that he is likely to suffer from MMAF, which were not tested during his initial recruitment. Patient 2603 and 2 patients from Zhang's cohort and their respective partners were able to conceive following ICSI.

7.3.3 CNVs overlapping known MMAF disease genes

Since novel MMAF genes were often identified in men classified as asthenoteratozoospermic, I looked for CNVs encompassing any of the know 22 MMAF genes (*AK7, ARMC2, CEP135, CFAP43, CFAP44, CFAP58, CFAP65, CFAP69, CFAP70, CFAP91, CFAP251, DNAH1, DNAH2, DNAH6, DNAH8, DNAH17, DZIP1, FSIP2, QRICH2, SPEF2, TTC21A* and *TTC29*) (Liu *et al.*, 2020; Lv *et al.*, 2020; Martinez *et al.*, 2020; Touré *et al.*, 2020). No CNVs involving these genes were identified.

7.4 Discussion

In this study, the WES data of 24 patients affected by asthenoteratozoospermia was analysed to identify possible disease-causing CNVs. It has been demonstrated that bi-allelic variants in genes important for sperm tail development are often found in asthenoteratozoospermia patients (Touré *et al.*, 2020). These mutations cause abnormalities of the sperm flagella, a characteristic defined as MMAF syndrome, and can be the underlying genetic causes of the asthenoteratozoospermia phenotype.

In our cohort of 24 patients, a total of 83 autosomal deletions and 64 autosomal duplications were detected. In this case, the number of losses exceeded the number of duplications by 19 CNV events. Interestingly, in the cohort of 142 patients with quantitative defects of the sperm production analysed in the previous chapter, the total number of duplications on the autosomes was instead higher than the number of deletions. Moreover, despite this cohort comprised 118 more patients than the asthenoteratozoospermia group, no deletions were found on the X chromosome, while 3 losses in total were detected in the 24 asthenoteratozoospermia patients. This suggests that patients with asthenoteratozoospermia might be more likely to carry genomic losses than patients with azoospermia and severe oligozoospermia.

To detect likely pathogenic CNVs, I prioritised rare (present in < 1% of the samples of the DGV Gold Standard) and large (comprising >= 10 probes) CNVs. These CNVs are more likely to be pathogenic as they are rare in the general population and encompass a larger fraction of the coding region compared to the other CNVs. In this category, there were 3 deletions and 3 duplications on the autosomes and a single deletion on the X chromosome. Amongst these, 2 losses encompassed 2 different genes with a role in spermatogenesis, while the CNV on the X chromosome removed a gene from the CFAP family, which includes 8 genes found mutated in patients with asthenoteratozoospermia with MMAF in the literature (Touré *et al.*, 2020).

The deletion found in patient 2093 completely removed a single copy of the *PIWIL1* gene. *PIWIL1* is expressed exclusively in the testis in humans, and it is essential for normal spermatogenesis in mice (Kuramochi-Miyagawa *et al.*, 2004). In 2017, Guo *et al.* reported 3 heterozygous variants, of which 2 *de novo* and 1 maternally inherited, in a specific region of this gene, called D-box, in 3 azoospermic patients from a cohort of 413 affected individuals (Gou *et al.*, 2017). The authors suggested that these variants lead to male infertility acting in a dominant fashion. These results have been recently challenged by Oud *et al.*, who claimed that the Sanger sequencing results validating these variants were not clear and possibly misinterpreted (Oud *et al.*, 2021). In addition, Oud *et al.* did not find any variant in the same region of *PIWIL1* in 1950 patients with azoospermia and 790 patients with severe oligozoospermia (Oud *et al.*, 2021), which is not consistent with the variant frequency proposed by Guo *et al*. Also, they did not identify a significant difference in the number of non-synonymous *PIWIL1* variants between their cohort of patients and 3347 fertile controls. The pLI score of *PIWIL1* is 0, indicating that the gene is unlikely to show haploinsufficiency when only one allele is lost. This conclusion is also supported by Oud *et al.*, who reported 2 loss-of-function variants in the gene in two fertile men (Oud *et al.*, 2021). In patient 2093, we did not identify a loss-of-function variant that could impair the remaining copy of *PIWIL1*. Therefore, the *PIWIL1* heterozygous deletion alone is unlikely to cause a male infertility phenotype. Other pathogenic variants affecting non-coding regulatory regions of *PIWIL1*, not sequenced with WES, might contribute to the loss of the gene function in patient 2093. Nevertheless, the role of *PIWIL1* in male infertility and the consequences of its impairments still need to be clarified.

The heterozygous deletion found in patient 1903 deleted one copy of the coding genes *CATSPER2* and *CKMT1A*. The genomic region where these genes are located on chromosome 15 (15q15.3) has been previously associated with a deafness-infertility syndrome (Avidan *et al.*, 2003). Avidan *et al.*, in 2003, described, for the first time, a bi-allelic deletion involving, amongst others, *CATSPER2* and *STRC*, which cosegregated with deafness and asthenoteratozoospermia in a French family. Later, other studies reported bi-allelic deletions of the same genes in other families and supported the hypothesis of *CATSPER2* losses causing infertility and *STRC* deletions causing sensorineural hearing loss (Zhang *et al.*, 2007; Hoppman *et al.*, 2013). The expression data of the two genes is consistent with the phenotype of the patients described, as *CATSPER2* is expressed in the sperm, while *STRC* is expressed in the inner ear (Zhang *et al.*, 2007). *CATSPER2*-null mice are infertile with a normal sperm count, and it is thought that the absence of the gene causes the lack of sperm hyperactivation and consequent inability to fertilise the oocyte (Quill *et al.*, 2003). Patient 1903 did not report hearing impairment, and the *STRC* gene was not involved in the deletion detected. Also, patient 1903 deletion was heterozygous, and additional loss-of-function variants in *CATSPER2* were not identified, while only bi-allelic pathogenic variants of this gene have been reported in infertile patients. The 603 deletions of the *CATSPER2* coding region reported in population databases and the pLI score of *CATSPER2* equal to 0 support the hypothesis that the heterozygous deletion identified in patient 1903 is unlikely to cause the disease alone. A second loss-of-function variant not detected on the remaining allele of *CATSPER2* or on the gene's promoter in trans with the heterozygous deletion could explain the infertility of patient 1903. Another possibility, although speculation, is that a homozygous deletion of *CATSPER2* and *CKTM1A* was not detected in the WES data due to the limitation of short-read sequencing. Avidan *et al.* and Zhang *et al.* described a second copy of *CATSPER2* and *CKTM1A* (the second copy is now referred to as *CKTM1B*) on chromosome 15, which have both > 98% sequence identity with the first copies of the genes (Avidan *et al.*, 2003; Zhang *et al.*, 2007). It is possible that some sequencing reads generated from the sequencing of the second copies of *CATSPER2* and *CKTM1* were mapped erroneously against *CATSPER2* and *CKTM1* original genes due to the high homology of their sequences, simulating the presence of one allele in the WES data and masking a homozygous deletion of *CASTPER2* (original gene) and *CKTM1A*. This should be further tested with a dedicated assay that amplifies specifically the original copy of *CATSPER2* (and *CKTM1A*) gene in patient 1903. Interestingly, the presence of an additional copy of *CATSPER2* and those of the flanking genes *CKTM1*, *STRC* and *KIAA0377* might be the cause of the recurrent deletions in 15q15.3 locus via non-allelic homologous recombination, as Avidan *et al.* and Zhang *et al.* have previously suggested (Avidan *et al.*, 2003; Zhang *et al.*, 2007).

Lastly, I described a 3.2 Mb large hemizygous deletion in patient 2603. This loss removed 9 genes with unknown functions, including the *CFAP47* gene, which was the only high pLI score gene affected. Deletions larger than 1 Mb on the sex chromosomes have not been detected in the male infertility patients described in the previous chapters (325 patients in total). Also, such deletion has not been reported in population databases, even in female individuals where a second copy of the chromosome X might compensate for the deleterious effects of the CNV. The data from our colleagues showed the disruption of *CFAP47* in other 3 patients with asthenoteratozoospermia with MMAF. Their experiments on a mouse model and protein expression demonstrated that impairment of the *CFAP47* gene reduces the *CFAP47* expression in the sperm drastically and modifies the location of the protein encoded by *CFAP65*, a known MMAF gene (Zhang *et al.*, 2019; W. Li *et al.*, 2020). Jointly, these findings demonstrated that *CFAP47* is necessary for normal spermatogenesis and its impairment cause asthenoteratozoospermia with MMAF, possibly disrupting *CFAP47-CFAP65* interaction. The results of this study were published earlier this year in the American Journal of Human Genetics (Liu *et al.*, 2021). It is unclear what are the consequences of the loss of the other genes involved in the deletion in patient 2603, since their function is not known, and they were not deleted in the asthenoteratozoospermia patients analysed by our colleagues.

Importantly, 3 of these men harbouring variants in *CFAP47* and their partners conceived successfully after ICSI. In these cases, a mutated allele is transmitted to 100% of female offspring, who will be carriers, while 100% of male offspring receives the chromosome X from the mother. Notably, instead, men with bi-allelic mutation in other autosomal MMAF genes must transmit a mutated allele to all the offspring, who will have 50% chance of transmitting it to the next generation. This understanding of the genetic model of the disease is very important to improve genetic counselling for male infertility patients.

Overall, based on the study of the WES data from a cohort of 24 patients with asthenoteratozoospermia, this investigation revealed variants in two loci of interest and, in combination with experimental data from our colleagues, a novel MMAF gene. While for *PIWIL1* and *CATSPER2*, additional experiments are required to clarify their role in the disease of the respective patients, we identified the cause of the disease for patient 2603: a pathogenic deletion of *CFAP47*, the first X-linked gene associated with asthenoteratozoospermia with MMAF.

7.5 Conclusions

In this study, I investigated the role of CNVs in 24 patients affected by asthenoteratozoospermia, a phenotype characterised by reduced sperm motility, reduced progressive motility, and morphological defects of the spermatozoa. The analysis revealed 2 rare large heterozygous deletions involving 2 genes with known involvement in spermatogenesis. The role of one of the impaired genes in male infertility is unclear (*PIWIL1*), while for the other (*CATSPER2*), only homozygous pathogenic variants have been described in patients before. The 2 deletions are unlikely to impair the spermatogenesis alone and other undetected genetic variants might contribute to the aetiology of the disease in the 2 patients. In addition, a very large deletion of unknown significance was detected on chromosome X of a third patient. Our data and those from our colleagues eventually demonstrated that this CNV was pathogenic and one of the genes involved (*CFAP47*) is a novel disease gene associated with asthenoteratozoospermia with MMAF. It is the first X-linked gene associated with this male infertility phenotype. The findings on *CFAP47* will improve the genetic diagnosis of patients with this male infertility phenotype and the genetic counselling as ICSI has been successful for patients carrying *CFAP47* deleterious mutations and their partners.

Chapter 8. General Discussion

8.1 WES Is an Effective Tool to Study the Genetics of Male Infertility

Throughout this thesis, I demonstrated that WES is able to detect CNVs of potential clinical interest in male infertility patients using both saliva-derived and blood-derived DNA, as well as two different exome enrichment kits. In the field of male infertility, NGS techniques are substituting Sanger sequencing in gene discovery and replication studies (Oud *et al.*, 2019). This reflects the advantages of techniques such as WES, which allows the analysis of SNVs and CNVs in all the genes with a single genetic test. As an example, between the end of 2019 and the first months of 2021, WES greatly improved our understanding of the genetic of asthenoteratozoospermia with MMAF, allowing the discovery of 5 novel disease genes (He *et al.*, 2020; Liu *et al.*, 2020; Lv *et al.*, 2020; Martinez *et al.*, 2020; Liu *et al.*, 2021). Nowadays, very large cohorts of infertile patients are available. Two examples are the cohorts collected by the GEMINI and the MERGE studies, which respectively include 1200 azoospermic patients and 1000 infertile men who underwent WES. At present, these datasets are used predominantly for SNV studies, and so far, no CNV studies have been reported making use of these large cohorts. CNV studies in such large groups of patients would be very interesting. They could, for example, be used to assess the frequency of the likely pathogenic CNVs identified in this thesis and to look for deleterious variants in the candidate disease genes revealed without additional costs. Also, the GEMINI cohort includes 2000 fertile controls that could be screened to test whether the likely pathogenic variants identified in our cohorts are absent in a large group of fertile men. The information and updates regarding these two studies are available on the IMIGC website (http://www.imigc.org/). Compared to WES, WGS has several advantages (Meynert *et al.*, 2014). It allows a better characterisation of all structural variants and the detection of non-coding variants (Zhao *et al.*, 2013). However, due to its current costs, very few studies have adopted this technique to study male infertility (Dong *et al.*, 2015; Gershoni *et al.*, 2017), and no data from large cohorts are available at present. A a pilot study using WGS is currently ongoing for all the patient-parent trios studied in this thesis. When the costs of WGS decreases further, this method will likely be more commonly used.

8.2 Novel Candidate Azoospermia and Severe Oligozoospermia Genes

In my PhD projects, I used a novel CNV detection tool developed for WES data to identify likely pathogenic CNVs in different male infertility cohorts. Initially, I studied the role of dominant CNVs in a cohort of azoospermia and severe oligozoospermia patient-parent trios. Next, I identified and studied the CNVs present in a patient-only cohort with similar phenotypes and in a cohort of asthenoteratozoospermia patients.

First, the analysis of the *de novo* CNVs (chapter 4), performed on the cohort of 183 patient-parent trios, revealed 2 *de novo* deletions in 2 different probands. One of them partially overlapped with a deletion of unknown parental origin previously reported in an azoospermic man (Seabra *et al.*, 2014). The overlapping region encompasses a total of 7 genes, of which 2 are predicted to be intolerant to loss-of-function variants: *CSTF3* and *QSER1*. *CSTF3* has a high expression in the testis, and it is involved in pre-mRNA processing (Grozdanov *et al.*, 2018). The second *de novo* deletion identified in the trio cohort removed the only copy of the *NXT2* gene on chromosome X in a proband. This gene has a high RNA expression in the testis and plays a role in the mRNA nuclear export (Herold *et al.*, 2000). Pre-mRNA alternative splicing is thought to be one of the main mechanisms that regulate spermatogenesis and testis development in mammals (Song *et al.*, 2020), and my colleagues identified other novel candidate disease genes, impaired by a *de novo* mutation of pathogenic nature, that are involved in this process. The two *de novo* deletions described might be involved in the origin of male infertility in the respective probands, and they are the first likely pathogenic *de novo* CNVs identified outside the chromosome Y in patients with idiopathic quantitative male infertility.

Secondly, the analysis of the inherited CNVs (chapter 5), performed on the same cohort of trios, revealed 2 maternally inherited deletions that might play a role in the impaired spermatogenesis of the patients. One of them deleted, amongst others, one copy of *AKT1* (pLI score = 0.98). This gene has been described as a prosurvival factor that, when phosphorylated, prevents the mature spermatozoa from undergoing a default apoptotic pathway in humans (Koppers *et al.*, 2011) and is essential for normal spermatogenesis in mice (Kim, Omurtag and Moley, 2012). A second deletion removed, amongst others, one copy of *PSME4* (pLI score = 1). This gene is involved in the histone exchange occurring during spermatogenesis in mammals (Qian *et al.*, 2013). In mice, its loss leads to impaired male, but not female, fertility (Khor *et al.*, 2006). These two CNVs were the only maternally inherited deletions that encompassed at least one constrained gene in the trios, and we propose them as likely pathogenic for male fertility. These inherited deletions provide the first evidence that CNVs escaping negative selection in females may play an important role in the aetiology of severe male infertility.

In addition to these possibly causative CNVs, the analysis of the inherited CNVs in the trios and the analysis of the CNVs in the patient-only cohort (chapter 6) revealed genomic loci of interest for further investigations.

A maternally inherited deletion removed the *SAGE1* gene on chromosome X of one proband. This gene does not have a known function, but its protein is only found in the spermatogonia (Uhlén *et al.*, 2015).

A large region in the Xp22.31 locus was duplicated in 2 patients. One of these duplications was maternally inherited, while the origin of the other is unknown since parental samples were not available. Considering that the chromosome X is maternally inherited in males, we assume that, in that case, the variant is either maternally inherited or *de novo*. The 2 duplications encompassed, amongst others, the *VCX* gene. *VCX* duplications have been found more frequently in infertile men than controls and *in vitro* experiments suggested that the upregulation of the gene might lead to cell apoptosis (Ji *et al.*, 2016). The consequences of variants in this genomic region have been debated for a long time (Li *et al.*, 2010), but since further studies on the consequences of additional copies of *VCX* in male infertility patients have not been performed yet, the role of *VCX* is still not fully understood.

In the patient-only cohort, 3 very large autosomal deletions (> 1 Mb) were identified. Autosomal deletions of such size were not detected in the trio cohort, and they are rarely found in the general population. Two of these CNVs encompassed each a constrained gene (*KBTBD2* and *XYLT1*), but their function is either unknown or not related to spermatogenesis. It is possible that these deletions are in combination with other undetected pathogenic variants that impair the other genes on the remaining allele.

Finally, in the patient-only cohort, a deletion affecting the high pLI score (1) gene *TDRD1* was identified. This gene is required for male fertility in mice, where it is essential for preserving the germline integrity (Chuma *et al.*, 2006; Reuter *et al.*, 2009; Vagin *et al.*, 2009). Bi-allelic variants in 2 genes from the TDRD family (*TDRD9* and *TDRD7*) have been previously associated with non-obstructive azoospermia (Arafat *et al.*, 2017; Tan *et al.*, 2019). *TDRD1* function in humans is still unclear, but it might have a function similar to that described in mice and be important for male infertility.

These analyses demonstrated the advantages of studying dominant variants in male infertility patient-parent trios. I identified, for the first time, *de novo* (outside the chromosome Y) and maternally inherited CNVs with a possible role in male infertility as well as several novel candidate dominant genes for male infertility. We hope that this approach will be more commonly used in the future as it is likely to improve our understanding of the role of dominant variants in quantitative male infertility. The analysis of patient-only cohorts, when parental samples are not available, should not be neglected since, as demonstrated here and by other studies discussed in this thesis, it can reveal loci of potential clinical interest for further studies. However, *de novo* and maternally inherited variants cannot be investigated with this approach. The study of these male infertility cohorts confirmed the high genetic heterogeneity of quantitative male infertility as the possible causative variants identified were detected only in one patient.

8.3 The Role of Dominant CNVs in Azoospermia and Severe Oligozoospermia

In the introduction of this thesis, I compared the role of *de novo* variants in ID to that hypothesised for severe male infertility. Based on our trio cohort, the rate of *de novo* CNVs per generation in azoospermia and oligozoospermia patients was ~ 0.01 (2 out of 183 trios). Microarray-based studies have demonstrated that in patients with ID, the rate of *de novo* CNVs per generation is ~ 0.1 (Vissers, Gilissen and Veltman, 2016). Such a large difference between the rates estimated for male infertility and ID patients might have different explanations. First, it is possible that, due to the limitation of CNV detection from WES, in our cohort, other *de novo* CNVs remained undetected. Secondly, compared to ID, where a highly penetrant pathogenic *de novo* CNV would cause the disease phenotype in both sexes, male infertility might be caused by variants that do not affect female fertility. This means that reduced selective pressure on CNVs transmitted through the female line might reduce the proportion of male infertility cases explained by *de novo* CNVs. Thirdly, it might be that *de novo* CNVs, especially those affecting many genes, more often result in a severe form of IDs rather than in severe forms of isolated male infertility. Lastly, it is important to remember that patients with AZF deletions, which generally occur *de novo*, were excluded from our study. Therefore, our *de novo* CNV rate in male infertility patients is likely underestimated.

Large deleterious CNVs are a recognised cause of severe IDs (Gilissen *et al.*, 2014). In 2011, a study investigated the burden of very large CNVs (> 1Mb) in a neurodevelopmental disorders cohort comprising different ID phenotypes (Girirajan *et al.*, 2011). The authors showed that very large CNVs were more often found in ID patients compared to patients with autism. Similarly, these variants were more frequent in ID patients with multiple congenital anomalies compared to ID alone. This is not surprising considering that the most severe phenotypes are often caused by the loss of several genes (Girirajan *et al.*, 2011). In my CNV analyses, I prioritised rare and large CNVs in men with isolated male infertility as these criteria are often used to prioritise pathogenic CNVs in other diseases (Leppa *et al.*, 2016; Tsuchida *et al.*, 2018; Viñas-Jornet *et al.*, 2018). While this method was useful to select the most obvious candidate causative variants, it is possible that smaller CNVs contribute more to the origin of infertility. The men included in our studies do not have developmental disorders or other major diseases apart from impaired spermatogenesis, and the deletion or duplication of a single gene might be sufficient to impair their sperm production. Unfortunately, the frequency of the CNVs is inversely correlated to their genomic size, making the prioritisation of smaller CNVs for follow-up much more challenging. For these reasons, it is essential to (1) continue to work with patient-parent trios and focus on *de novo* and maternally inherited CNVs; (2) continue to expand these CNV studies to large cohorts in order to identify recurrently affected genes; and (3) integrate the CNV studies with SNV studies as this will allow us to identify genes affected by both types of mutations (see the example in the next paragraph).

8.4 The Genetic Model of Asthenoteratozoospermia With MMAF

If cohorts of patients without parental samples are of limited value for the identification of novel candidate disease genes for azoospermia and severe oligozoospermia, this might not be true for other male infertility phenotypes. In Chapter 7, I identified an X-linked deletion involving the gene *CFAP47* in a patient with asthenoteratozoospermia. An international collaboration with other research groups in China, Australia and France revealed that *CFAP47* was hemizygously mutated in other 3 patients with asthenoteratozoospermia and MMAF. Experimental validations performed by our colleagues confirmed that this gene is the first X-linked gene associated with this male infertility phenotype (Liu *et al.*, 2021).

A recent study that reviewed the first 18 genes associated with asthenoteratozoospermia with MMAF reported that 30 to 60% of patients in cohorts of men with this phenotype have pathogenic variants in known disease genes (Touré *et al.*, 2020). We recently confirmed this report analysing the exome of 21 men with asthenoteratozoospermia. In this cohort, we identified pathogenic or likely pathogenic mutations in known MMAF-associated genes in 48% of the individuals and found predicted pathogenic mutations in novel candidate genes in other 33% of the patients (Oud *et al.*, 2021). This study and the results presented in chapter 7 suggests that contrarily to azoospermia and severe oligozoospermia, asthenoteratozoospermia with MMAF has a smaller mutational target. One reason could be that the development of the sperm tail in humans may be delegated to a small number of genes and to specific gene families, contrarily to the entire spermatogenesis process, which involves numerous genes. In fact, ~57% of the 23 known MMAF-associated genes are part of only two gene families: the Cilia and Flagella Associated protein (CFAP) and the Dynein Axonemal Heavy Chain protein (DNAH) families. Patient-only cohorts might be then sufficient to identify a genetic diagnosis in these infertile men and to reveal novel disease genes associated with asthenoteratozoospermia with MMAF since the current genetic model proposed for the disease is autosomal recessive as well as X-linked, and the total number of disease genes is expected to be much lower compared quantitative forms of male infertility.

8.5 How to Improve Male Infertility Diagnostics?

The findings of our exploratory studies on the genetics of azoospermia and severe oligozoospermia will not immediately change the diagnostics for male infertility patients. The field first needs to (1) further study the contribution of dominant variants to the genetics of the disease, (2) understand the exact function of the candidate genes in humans, and (3) find unrelated patients with the same impaired gene. Once our knowledge of the disease genes is more accurate, we will be able to test the presence of pathogenic variants in these genes routinely and provide a genetic diagnosis to infertile men with severe quantitative male infertility.

Currently, diagnostic practice only recommends chromosome Y deletions analysis and karyotyping to patients with azoospermia and oligozoospermia and *CFTR* mutation screening to patients with obstructive azoospermia (Jungwirth A.D.T. *et al.*, 2018). Considering our current knowledge of the genetic of asthenoteratozoospermia, it is surprising that there are no genetic tests advised for patients with this phenotype. These infertile men would greatly benefit from a genetic test that looks for pathogenic variants in the current known disease genes. For instance, three out of four patients harbouring a pathogenic variant in *CFAP47* described in Chapter 7 previously achieved, with their partner, successful fertilisation through ICSI treatment (Liu *et al.*, 2021). We now know that their pathogenic mutations on chromosome X will not be transmitted to male offspring and that females will be heterozygous carriers. Also, we know that men with pathogenic autosomal homozygous variants undergoing ICSI will transmit one pathogenic allele to the offspring regardless of their sex. Thus, providing a genetic diagnosis to men with asthenoteratozoospermia would massively improve genetic counselling and would allow them to make an informed decision on treatments.

Diagnostic implementation of WES and WGS already revolutionised the field of Mendelian disease and cancer (Meyerson, Gabriel and Getz, 2010; Nakagawa *et al.*, 2015; Bamshad, Nickerson and Chong, 2019). The introduction of these methods in the diagnostic practice for all male infertility patients would have several advantages. First, we would be able to test the presence of chromosomal abnormalities, chromosome Y deletions and *CFTR* mutations with a single test in patients with azoospermia or oligozoospermia. Also, the same test could be used for patients with asthenoteratozoospermia and other male infertility phenotypes to test the presence of pathogenic variants in known disease genes. Secondly, the systematic collection of WES data from large cohorts of patients, and parents when available, would greatly benefit the genetic research, as it would facilitate the study of recessive and dominant variants in all the genes and the identification of new genes confidently linked to male infertility. While pathogenic SNVs might be the primary focus of diagnostic WES and WGS, these methods would achieve a higher diagnostic yield compared to other genetic tests as they also allow an accurate investigation of the CNVs and other structural variants.

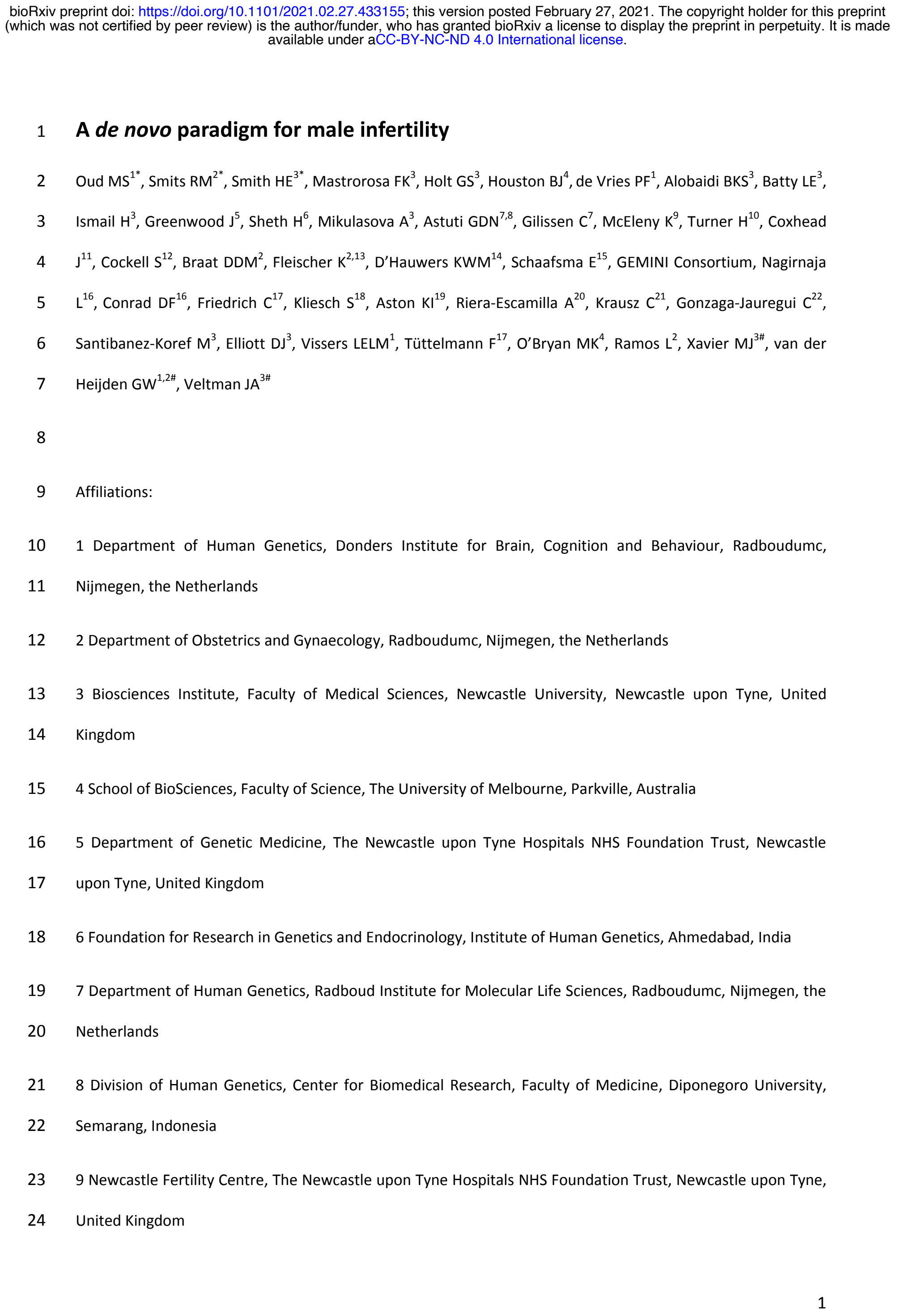
8.6 Concluding Remarks

Around 10-15% of couples in developed and developing countries are infertile, and half of the cases are due to male factors. A limited number of genes have been, so far, associated with isolated male infertility and most of the known disease genes are recessive. Dominant variants are more likely to explain cases in the general population, but their contribution has not been investigated yet in a systematic manner. The work included in this thesis is pioneering for the field of male infertility genetics. It represents the first attempt to demonstrate that that *de novo* (outside the chromosome Y) and maternally inherited CNVs might be an unstudied cause of azoospermia and severe oligozoospermia. The chapters of this thesis highlight the advantages of collecting and studying patient-parent trios in male infertility. They also provide a list of novel candidate dominant male infertility genes and loci of potential clinical interest. The investigation of different male infertility cohorts shows the convenience of using WES for gene discovery and CNV analysis. With this approach, we were able to also identify a novel disease gene associated with asthenoteratozoospermia with MMAF. Future studies should focus on the collection of large cohorts of trios and patient-only cohorts when parental samples are not available. Also, they should take advantage of genomics techniques such as WES and WGS, which allow the identification of point mutations and structural variants with an unbiased approach. Finally, it is important to remark the crucial role of collaborations amongst research groups. This is essential for accessing larger datasets and tackle the large mutational target of severe quantitative forms of male infertility. Moreover, the diverse expertise of different research groups would allow conducting complementary studies on different aspects of the disease, such as the genetic variants interpretation, the function of candidate genes in humans, the consequences of the impairment of these genes in an animal model, and the clinical aspects of the patients' recruitment and phenotype classification.

Chapter 9. Appendices

9.1 Appendix A: Research Papers

During my PhD, I have been listed as co-author in 3 research papers.



A *de novo* paradigm for male infertility

Oud MS[1*], Smits RM[2*], Smith HE[3*], Mastrorosa FK[3], Holt GS[3], Houston BJ[4], de Vries PF[1], Alobaidi BKS[3], Batty LE[3], Ismail H[3], Greenwood J[5], Sheth H[6], Mikulasova A[3], Astuti GDN[7,8], Gilissen C[7], McEleny K[9], Turner H[10], Coxhead J[11], Cockell S[12], Braat DDM[2], Fleischer K[2,13], D'Hauwers KWM[14], Schaafsma E[15], GEMINI Consortium, Nagirnaja L[16], Conrad DF[16], Friedrich C[17], Kliesch S[18], Aston KI[19], Riera-Escamilla A[20], Krausz C[21], Gonzaga-Jauregui C[22], Santibanez-Koref M[3], Elliott DJ[3], Vissers LELM[1], Tüttelmann F[17], O'Bryan MK[4], Ramos L[2], Xavier MJ[3#], van der Heijden GW[1,2#], Veltman JA[3#]

Affiliations:

1 Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboudumc, Nijmegen, the Netherlands

2 Department of Obstetrics and Gynaecology, Radboudumc, Nijmegen, the Netherlands

3 Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

4 School of BioSciences, Faculty of Science, The University of Melbourne, Parkville, Australia

5 Department of Genetic Medicine, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

6 Foundation for Research in Genetics and Endocrinology, Institute of Human Genetics, Ahmedabad, India

7 Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboudumc, Nijmegen, the Netherlands

8 Division of Human Genetics, Center for Biomedical Research, Faculty of Medicine, Diponegoro University, Semarang, Indonesia

9 Newcastle Fertility Centre, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

10 Department of Cellular Pathology, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

11 Genomics Core Facility, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

12 Bioinformatics Support Unit, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

13 TFP Center of Reproductive Medicine Düsseldorf, Germany

14 Department of Urology, Radboudumc, Nijmegen, the Netherlands

15 Department of Pathology, Radboudumc, Nijmegen, the Netherlands

16 Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, Oregon, United States of America

17 Institute of Reproductive Genetics, University of Münster, Münster, Germany

18 Centre of Reproductive Medicine and Andrology, Department of Clinical and Surgical Andrology, University Hospital Münster, Münster, Germany

19 Department of Surgery, Division of Urology, University of Utah School of Medicine, Salt Lake City, Utah, United States of America

20 Andrology Department, Fundació Puigvert, Universitat Autònoma de Barcelona, Instituto de Investigaciones Biomédicas Sant Pau (IIB-Sant Pau), Barcelona, Catalonia, Spain.

21 Department of Biomedical, Experimental and Clinical Sciences “Mario Serio”, University of Florence, Florence, Italy

22 Regeneron Genetics Center, Tarrytown, New York, United States of America

* Joint first authors

# Joint last authors

Introduction

De novo mutations (DNMs) are known to play a prominent role in sporadic disorders with reduced fitness[1]. We hypothesize that DNMs play an important role in male infertility and explain a significant fraction of the genetic causes of this understudied disorder. To test this hypothesis, we performed trio-based exome-sequencing in a unique cohort of 185 infertile males and their unaffected parents. Following a systematic analysis, 29 of 145 rare protein altering DNMs were classified as possibly causative of the male infertility phenotype. We observed a significant enrichment of Loss-of-Function (LoF) DNMs in LoF-intolerant genes (p-value=1.00x10-5) as well as predicted pathogenic missense DNMs in missense-intolerant genes (p-value=5.01x10-4). One DNM gene identified, RBM5, is an essential regulator of male germ cell pre-mRNA splicing[2]. In a follow-up study, 5 rare pathogenic missense mutations affecting this gene were observed in a cohort of 2,279 infertile patients, with no such mutations found in a cohort of 5,784 fertile men (p-value=0.009). Our results provide the first evidence for the role of DNMs in severe male infertility and point to many new candidate genes affecting fertility.

Main

Male infertility contributes to approximately half of all cases of infertility and affects 7% of the male population. For the majority of these men the cause remains unexplained[3]. Despite a clear role for genetic causes in male infertility, there is a distinct lack of diagnostically relevant genes and at least 40% of all cases are classified as idiopathic[3–6]. Previous studies in other conditions with reproductive lethality, such as neurodevelopmental disorders, have demonstrated an important role for *de novo* mutations (DNMs) in their etiology[1]. In line with this, recurrent *de novo* chromosomal abnormalities play an important role in male infertility. Both azoospermia Factor (AZF) deletions on the Y chromosome as well as an additional X chromosome, resulting in Klinefelter syndrome, occur *de novo.* Collectively, these *de novo* events explaining up to 25% of all cases of non-obstructive azoospermia (NOA)[3,6]. Interestingly, in 1999 a DNM in the Y-chromosomal gene USP9Y was reported in a man with azoospermia[7]. Until now, however, a systematic analysis of the role of DNMs in male infertility had not been attempted. This is partly explained by a lack of basic research in male reproductive health in general[6,8], but also by the practical challenges of collecting parental samples for this disorder, which is typically diagnosed in adults.

In this study, we investigated the role of DNMs in 185 unexplained cases of oligozoospermia (<5 million sperm cells/ml; n=74) and azoospermia (n=111) by performing whole exome sequencing (WES) in all patients and their parents (see Supplementary Figure 1 and 2, Supplementary notes and tables for details on methods and clinical description). In total, we identified and validated 192 rare DNMs, including 145 protein altering DNMs. All *de novo* point mutations were autosomal, except for one on chromosome X, and all occurred in different genes (Supplementary Table 1). Two *de novo* copy number variations (CNVs) were also identified affecting a total of 7 genes (Supplementary Figure 3).

None of the 145-protein altering DNMs occurred in a gene already known for its involvement in autosomal dominant human male infertility. This is not unexpected as only 4 autosomal dominant genes have so far been linked to isolated male infertility in humans[5,9]. Broadly speaking, across genetic disorders, dominantly acting disease genes are usually intolerant to loss-of-function (LoF) mutations, as represented by a high pLI score[10]. The median pLI score of genes with a LoF DNM (n=17) in our cohort of male infertility cases was significantly higher than that of genes with 181 LoF DNMs identified in a cohort of 1,941 control cases from denovo-db v1.6.1[11] (pLI male infertility=0.80, pLI controls=$3.75x10^{-5}$, p-value=$1.00x10^{-5}$) (Figure 1). This observation indicates that LoF DNMs likely play an important role in male infertility, similar to what is known for developmental disorders and severe intellectual disability[12,13]. As an example, a heterozygous likely pathogenic frameshift DNM was observed in the LoF intolerant gene *GREB1L* (pLI=1) of Proband_076. Homozygous *Greb1L* knock-out mice appear to be embryonic lethal, however, typical male infertility phenotypic features such as abnormal fetal testis morphology and decreased fetal testis volume are observed[14]. Interestingly, this patient has a reduced testis volume and severe oligospermia (Supplementary Notes Table 1). Nonsense and missense mutations in *GREB1L* in humans are known to cause renal agenesis[15] (OMIM: 617805), not known to be present in our patient. Of note, all previously reported damaging mutations in *GREB1L* causing renal agenesis are either maternally inherited or occurred *de novo.* This led the authors of one of these renal agenesis studies to speculate that disruption to *GREB1L* could cause infertility in males[14]. 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[16]. We also assessed the damaging effects of the two *de novo* CNVs 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 6 genes in total. This deletion partially overlapped with a deletion reported in 2014 in a patient with cryptorchidism and NOA[17]. Two genes affected in both patients, *QSER1* and *CSTF3*, are extremely 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 poylyadenylation[18].

To systematically evaluate and predict the likelihood of these DNMs causing male infertility and identify novel candidate disease genes, we assessed the predicted pathogenicity of all DNMs using three prediction methods based on SIFT[19], MutationTaster[20] and PolyPhen2[21]. Using this approach, 84/145 protein altering DNM were predicted to be pathogenic, while the remaining 61 were predicted to be benign. To further analyse the impact of the variants on the genes affected, we looked at the missense Z-score of all 122 genes affected by a missense variant, which indicates the tolerance of genes to missense mutations[22]. Our data highlights a significantly higher missense Z-score in genes affected by a missense DNM predicted as pathogenic (n=63) when compared to genes affected by predicted benign (n=59) missense DNMs (p-value=5.01x10^{-4}, Figure 2, Supplementary Figure 4). Furthermore, using the STRING database[23], we found 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 x 10^{-2}, Figure 3). 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, Supplementary Figure 5). These two findings suggest that (1) the predicted pathogenic missense DNMs detected in our study affect genes sensitive to missense mutations, and (2) the proteins affected by predicted pathogenic DNMs share common biological functions.

The STRING network analysis also highlighted a central module of interconnected proteins with a significant enrichment of genes required for mRNA splicing (Supplementary Figure 6). The genes *U2AF2*, *HNRNPL*, *CDC5L*, *CWC27* and *RBM5* all contain predicted pathogenic DNMs and likely interact at a protein level during the mRNA splicing process. Pre-mRNA splicing allows gene functions to be expanded by creating alternative splice variants of gene products and is highly elaborated within the testis[24]. One of these genes, *RBM5* has been previously highlighted as an essential regulator of haploid male germ cell pre-mRNA splicing and male fertility[2]. Mice with a homozygous ENU-induced allele point mutation in *RBM5* present with azoospermia and germ cell development arrest at round spermatids. Whilst in mice a homozygous mutation in *RBM5* is required to cause azoospermia, this may not be the case in humans as is well-documented for other genes[25], including the recently reported male infertility gene *SYCP2*[9]. Of note, *RBM5* is a tumour suppressor in the lung[26], with reduced expression affecting RNA splicing in patients with non-small cell lung cancer[27]. *HNRNPL* is another splicing factor affected by a possible pathogenic DNM in our study. One study implicated a role for *HNRNPL* in patients with Sertoli cell only phenotype[28]. The remaining three mRNA splicing genes have not yet been implicated in human male infertility. However, mRNA for all three is expressed at medium to high levels in human germ cells and all are widely expressed during spermatogenesis[29]. 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[30], as is CWC27[31]. U2AF2 plays a role in pre-mRNA splicing and 3'-end processing[32]. Interestingly, *CSTF3,* one of the genes affected by a *de novo* CNV in Proband_066, affects the same mRNA pathway[17].

Whilst DNMs most often cause dominant disease, they can contribute to recessive disease, usually in combination with an inherited variant on the trans allele. This was observed in Proband_060, who carried a DNM on the paternal allele, in trans with a maternally inherited variant in Testis and Ovary Specific PAZ Domain Containing 1 (*TOPAZ1*) (Supplementary Figure 7). *TOPAZ1* is a germ-cell specific gene which is highly conserved in vertebrates[33]. Studies in mice revealed that *Topaz1* plays a crucial role in spermatocyte, but not oocyte progression through meiosis[34]. In men, *TOPAZ1* is expressed in germ cells in both sexes[29,35,36]. Analysis of the testicular biopsy of this patient revealed a germ cell arrest in early spermiogenesis (Figure 4).

In addition to all systematic analyses described above, we evaluated the function of all DNM genes to give each a final pathogenicity classification (Table 1, details in Material & Methods). Of all 145 DNMs, 29 affected genes linked to male reproduction and were classified as possibly causative. For replication purposes, unfortunately no other trio-based exome data are available for male infertility, although we note that a pilot study including 13 trios was recently published[37]. While this precluded a genuine replication study, we were able to study these candidate genes in exome datasets of infertile men (n=2,279), in collaboration with members of the International Male Infertility Genomics Consortium and the Geisinger Regeneron DiscovEHR collaboration[38]. The 33 candidate genes selected for this analysis include the 29 genes mentioned above and 4 additional LoF intolerant genes carrying LoF DNMs with an 'unclear' final pathogenicity classification. For comparison, we included an exome dataset from a cohort of 11,587 fertile men and women from Radboudumc.

In the additional infertile cohorts, we identified only 2 LoF mutations in our DNM LoF intolerant genes (Supplementary table 2). Next, we looked for an enrichment of rare predicted pathogenic missense mutations in these cohorts (Table 2). A burden test revealed a significant enrichment in the number of such missense mutations present in infertile men compared to fertile men in the *RBM5* gene (adjusted p-value=0.009). In this gene, 5 infertile men were found to carry a distinct rare pathogenic missense mutation, in addition to the proband with a *de novo* missense mutation (Supplementary figure 8, Supplementary table 3). 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[2], is highly intolerant to missense mutations (missense Z-score 4.17).

Given the predicted impact of these DNMs on spermatogenesis, we were interested in studying the parental origin of DNMs in our trio-cohort. We were able to phase 29% of all our DNMs using a combination of short-read WES and targeted long-read sequencing (Supplementary Table 4). In agreement with literature[39–42], 72% of all DNMs occurred on the paternal allele. Interestingly, phasing of 8 likely causative DNMs showed that 6 of these were of paternal origin (75%). This suggests that DNMs with a deleterious effect on the future germline can escape negative selection in the paternal germline. This may be possible because the DNM occurred after the developmental window in which the gene is active, or the DNM may have affected a gene in the gamete's genome that is critical for somatic cells supporting the (future) germline. Transmission of pathogenic DNMs may also be facilitated by the fact that from spermatogonia onwards, male germ cells form cysts and share mRNAs and proteins[43]. As such, the interconnectedness of male germ cells, which is essential for their survival[44], could mask detrimental effects of DNMs occurring during spermatogenesis.

In 2010, we published a pilot study pointing to a *de novo* paradigm for mental retardation[45] (now more appropriately termed developmental delay or intellectual disability). This work contributed to the widespread implementation of patient-parent WES studies in research and diagnostics for neurodevelopmental disorders[46], 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 sequencing in male infertility. 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.

Data access

Raw and processed exome sequencing data of our 185 patient-parent trios is available under controlled access and requires a Data Transfer Agreement from the European Genome-Phenome Archive (EGA) repository: EGAS00001004945.

Acknowledgements

We are grateful for the participation of all patients and their parents in this study. We thank Laurens van de Wiel (Radboudumc), Sebastian Judd-Mole (Monash University), Arron Scott and Bryan Hepworth (Newcastle University) for technical support, and Margot J Wyrwoll (University of Münster) for help with handling MERGE samples and data. This project was funded by The Netherlands Organisation for Scientific Research (918-15-667) to JAV as well as an Investigator Award in Science from the Wellcome Trust (209451) to JAV, a grant from the Catherine van Tussenbroek Foundation to MSO, a UUKi Rutherford Fund Fellowship awarded to BJH and the German Research Foundation Clinical Research Unit 'Male Germ Cells' (DFG, CRU326) to CF and FT. This project was also supported in part by funding from the Australian National Health and Medical Research Council (APP1120356) to MKOB, by grants from the National Institutes of Health of the United States of America (R01HD078641 to D.F.C and K.I.A, P50HD096723 to D.F.C.) and from the Biotechnology and Biological Sciences Research Council (BB/S008039/1) to DJE.

Author contributions

This study was designed by MSO, LELMV, LR and JAV. RMS, JG, HT and GWvdH provided all clinical data and performed the TESE histology and cytology analysis under supervision of LR, DDMB, ES, KF, KDH and KM. JC performed the exome sequencing with support from BA, and bioinformatics support was provided by MJX, GA, CG and SC. Sanger sequencing was performed by PFdV, HI, HES, LEB and BKSA. MSO and HES performed the SNV analyses with support from MJX, FKM performed CNV analysis with support from AM and MSK, and GSH and LEB performed the phasing. DJE, HS, BJH and MKOB provided support on the functional interpretation of mutations. DFC, LN, CF, SK, FT, KIA, ARE, CK, and CG-J were involved in the replication study. The first draft of the manuscript was prepared by MSO, HES, RMS, MJX, GWvdH, and JAV. All authors contributed to the final manuscript.

References

1. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
2. O'Bryan, M. K. *et al.* RBM5 Is a Male Germ Cell Splicing Factor and Is Required for Spermatid Differentiation and Male Fertility. *PLoS Genet.* **9**, e1003628 (2013).
3. Krausz, C. & Riera-Escamilla, A. Genetics of male infertility. *Nat. Rev. Urol.* **15**, 369–384 (2018).
4. Tüttelmann, F., Ruckert, C. & Röpke, A. Disorders of spermatogenesis. *medizinische Genet.* **30**, 12–20 (2018).
5. Oud, M. S. *et al.* A systematic review and standardized clinical validity assessment of male infertility genes. *Hum. Reprod.* **34**, 932–941 (2019).
6. Kasak, L. & Laan, M. Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives. *Hum. Genet.* **140**, 135–154 (2021).
7. Sun, C. *et al.* An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. *Nat. Genet.* **23**, 429–432 (1999).
8. De Jonge, C. & Barratt, C. L. R. The present crisis in male reproductive health: an urgent need for a political, social, and research roadmap. *Andrology* **7**, 762–768 (2019).
9. Schilit, S. L. P. *et al.* SYCP2 Translocation-Mediated Dysregulation and Frameshift Variants Cause Human Male Infertility. *Am. J. Hum. Genet.* **106**, 41–57 (2020).
10. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. denovo-db, Seattle, WA (denovo-db.gs.washington.edu) [Aug 2020].
12. Gu, Y. *et al.* Three intellectual disability-associated de novo mutations in MECP2 identified by trio-WES analysis. *BMC Med. Genet.* **21**, 99 (2020).
13. Fritzen, D. *et al.* De novo FBXO11 mutations are associated with intellectual disability and behavioural anomalies. *Hum. Genet.* **137**, 401–411 (2018).
14. De Tomasi, L. *et al.* Mutations in GREB1L Cause Bilateral Kidney Agenesis in Humans and Mice. *Am. J. Hum. Genet.* **101**, 803–814 (2017).
15. Brophy, P. D. *et al.* A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal

Agenesis Gene in Humans. *Genetics* **207**, 215–228 (2017).

16. Alhathal, N. *et al.* A genomics approach to male infertility. *Genet. Med.* **22**, 1967–1975 (2020).

17. Seabra, C. M. *et al.* A novel Alu-mediated microdeletion at 11p13 removes WT1 in a patient with cryptorchidism and azoospermia. *Reprod. Biomed. Online* **29**, 388–391 (2014).

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

19. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

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

21. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

22. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).

23. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).

24. Song, H., Wang, L., Chen, D. & Li, F. The Function of Pre-mRNA Alternative Splicing in Mammal Spermatogenesis. *Int. J. Biol. Sci.* **16**, 38–48 (2020).

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

26. Jamsai, D. *et al.* In vivo evidence that RBM5 is a tumour suppressor in the lung. *Sci. Rep.* **7**, 16323 (2017).

27. Liang, H. *et al.* Differential Expression of RBM5, EGFR and KRAS mRNA and protein in non-small cell lung cancer tissues. *J. Exp. Clin. Cancer Res.* **31**, 36 (2012).

28. Li, J. *et al.* HnRNPL as a key factor in spermatogenesis: Lesson from functional proteomic studies of azoospermia patients with sertoli cell only syndrome. *J. Proteomics* **75**, 2879–2891 (2012).

29. Wang, M. *et al.* Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis. *Cell Stem Cell* **23**, 599-614.e4 (2018).

30. Ajuh, P. Functional analysis of the human CDC5L complex and identification of its components by mass

spectrometry. *EMBO J.* **19**, 6569–6581 (2000).

31. Brea-Fernández, A. J. *et al.* Expanding the clinical and molecular spectrum of the CWC27-related spliceosomopathy. *J. Hum. Genet.* **64**, 1133–1136 (2019).

32. Millevoi, S. *et al.* An interaction between U2AF 65 and CF Im links the splicing and 3' end processing machineries. *EMBO J.* **25**, 4854–4864 (2006).

33. Baillet, A. *et al.* TOPAZ1, a Novel Germ Cell-Specific Expressed Gene Conserved during Evolution across Vertebrates. *PLoS One* **6**, e26950 (2011).

34. Luangpraseuth-Prosper, A. *et al.* TOPAZ1, a germ cell specific factor, is essential for male meiotic progression. *Dev. Biol.* **406**, 158–171 (2015).

35. Guo, F. *et al.* The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161**, 1437–1452 (2015).

36. Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 858-873.e4 (2017).

37. Hodžić, A. *et al.* De novo mutations in idiopathic male infertility—A pilot study. *Andrology* **9**, 212–220 (2021).

38. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science (80-.).* **354**, aaf6814 (2016).

39. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).

40. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).

41. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).

42. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).

43. Braun, R. E., Behringer, R. R., Peschon, J. J., Brinster, R. L. & Palmiter, R. D. Genetically haploid spermatids are phenotypically diploid. *Nature* **337**, 373–376 (1989).

44. Greenbaum, M. P., Iwamori, T., Buchold, G. M. & Matzuk, M. M. Germ Cell Intercellular Bridges. *Cold Spring Harb. Perspect. Biol.* **3**, a005850–a005850 (2011).

45. Vissers, L. E. L. M. *et al.* A de novo paradigm for mental retardation. *Nat. Genet.* **42**, 1109–12 (2010).

46. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17**, 9–18 (2016).

Figures and Tables

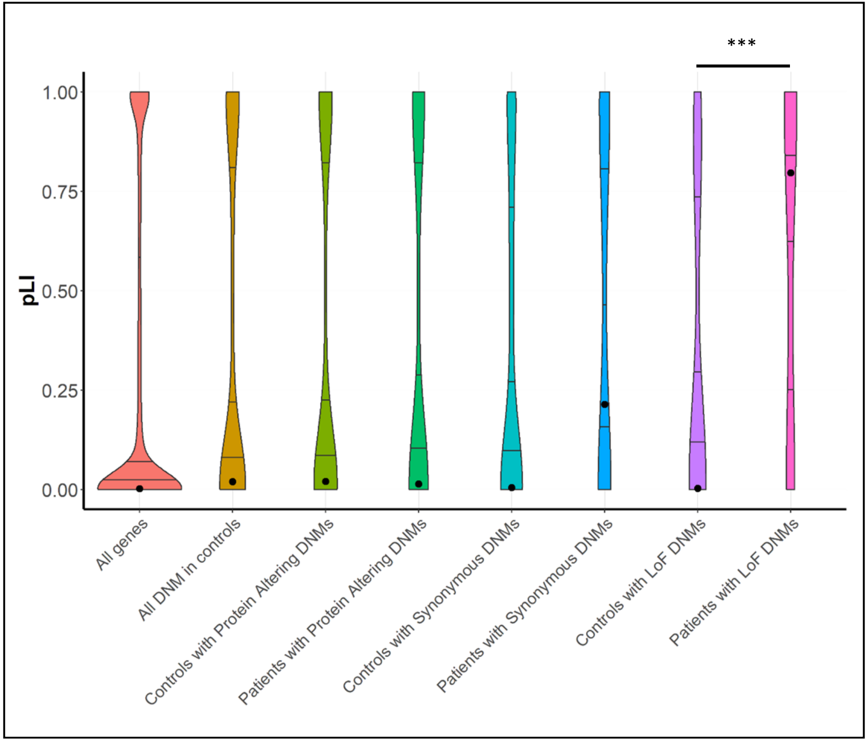
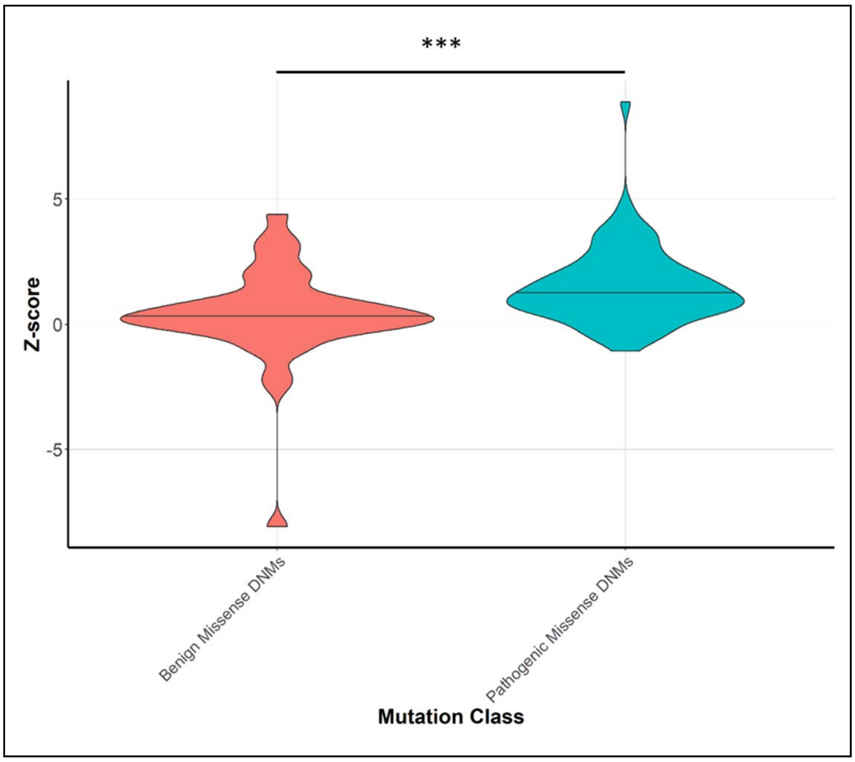
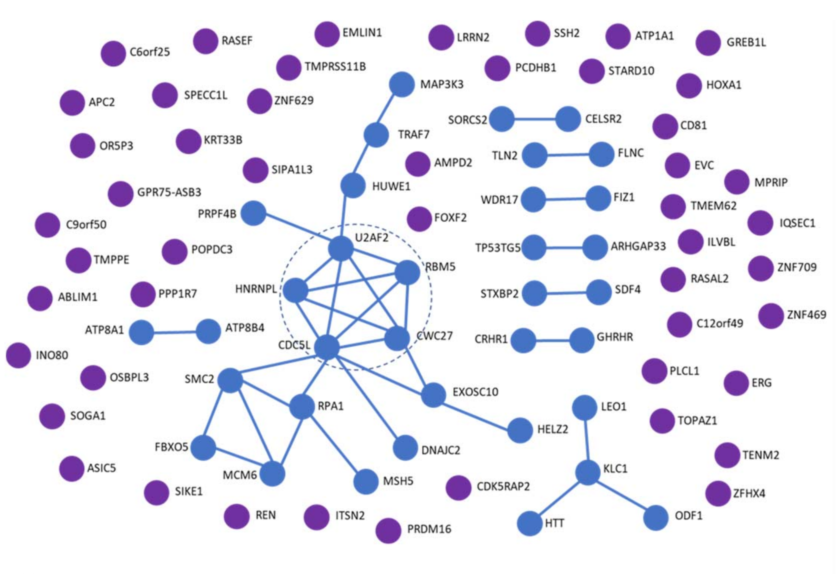


Figure 1: Analysis of the intolerance to loss-of-function variation for DNM genes. Violin plots represent the distribution of the pLI scores of all genes in gnomAD, all genes affected by DNMs and all LoF DNM in this study and in a control population (http://denovo-db.gs.washington.edu/denovo-db/). The observed median pLI score is displayed for each category as a black circle. The closer the pLI score is to 1, the more intolerant to LoF variation a gene is[10]. Comparison between LoF DNMs in our study and control populations shows a significance difference (p-value=1.00x10^{-5}).



358 **Figure 2: Intolerance to missense variants for genes with a DNM.** Violin plots show the distribution of Z-scores of genes
359 containing a missense DNM in our cohort, where an enrichment can be observed for predicated pathogenic DNMs in genes
360 more intolerant to missense mutations based on their mean z-score with a p-value of 5.01×10^{-4}.



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 3: Protein-protein interactions predicted for proteins encoded by damaging DNM genes. A protein-protein interaction analysis was performed for all 84 genes containing a DNM scored as damaging using the STRING tool[23]. A significantly larger number of interactions is observed between our damaging DNM genes than is expected for a similar sized dataset of randomly selected genes (PPI enrichment p-value 2.35 x 10^{-2}) with the number of expected edges being 25 and the observed being 36. The central module of the main interaction network within the figure contains 5 genes which are all involved in the process of mRNA splicing (Supplementary figure 6)

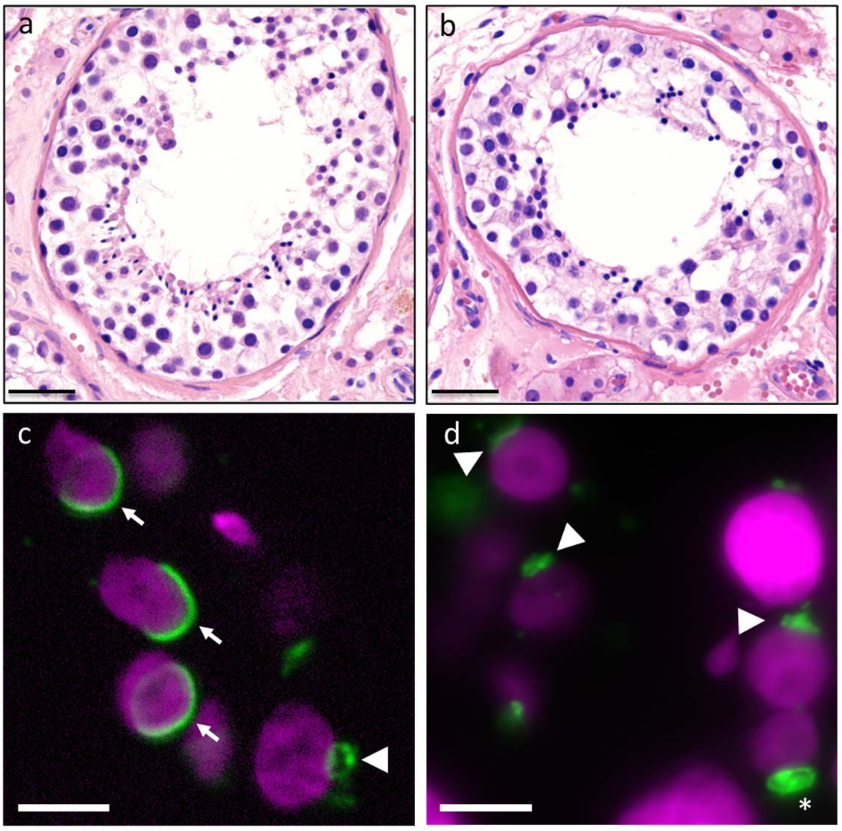


Figure 4: Description of control and *TOPAZ1* proband testis histology and aberrant acrosome formation: (a,b): H&E stainings of **(a)** control and **(b)** Proband_060 with DNM in *TOPAZ1* gene. The epithelium of the seminiferous tubules in the *TOPAZ1* proband show reduced numbers of germ cells and an absence of elongating spermatids. **(c,d):** immunofluorescent labelling of DNA (magenta) and the acrosome (green) in control sections **(c)** and *TOPAZ1* proband sections **(d)**. **(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. Size bar in a, b: 40 µm, c, d: 5 µm.

Table 1: *De novo* mutation classification summary.

	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

A total of 192 rare DNMs were classified based on pathogenicity scores as well as functional data into 4 categories, ‘Possibly causative’, ‘Unclear’, ‘Unlikely Causative’ and ‘Not causative’.

Table 2: Rare potentially pathogenic missense mutations in exome data from various cohorts of infertile men and fertile control cohorts.

Gene	Missense Z-score	NIJ/NCL Cohort of Patient-Parent Trios (n=185)	NIJ/NCL Cohort of Infertile Men (Singleton) (n=145)	MERGE Cohort of Infertile Men (n=887)	GEMINI Cohort of NOA Men (n=926)	Geisinger-Regeneron DiscovEHR Cohort of Infertile Men (n=88)	Italian Cohort of NOA Men (n=48)	Total Infertile Men (n=2,279)	Fertile Dutch Men (n=5,784)	Fertile Dutch Women (n=5,803)	Burden test Infertile vs Fertile Men (Bonf)	Burden test Fertile Men vs Women (Bonf)
ABLIM1	1.62	1	1	1	1	1	0	5	1	1	0.15	1
ATP1A1	6.22	0	0	0	1	0	0	1	0	1	1	1
CDC5L	2.78	1	1	1	3	0	0	6	2	4	0.15	1
CDK5RAP2	-0.37	1	0	1	1	0	0	3	5	5	1	1
HUWE1	8.87	1	0	2	0	0	0	3	0	0	0.41	1
INO80	3.53	1	0	1	0	0	0	2	3	3	1	1
MAP3K3	2.04	1	0	2	0	0	0	3	1	2	1	1
MCM6	1.07	1	1	1	3	0	0	6	4	8	0.64	1
PPP1R7	1.86	0	0	0	1	0	0	1	1	1	1	1
QSER1	1.34	0	1	1	0	0	0	2	8	1	1	0.38
RASAL2	1.40	0	1	1	2	1	0	5	25	13	1	0.94
RBM5	4.17	1	2	2	0	1	0	6	0	2	0.009	1
RPA1	1.22	1	0	0	1	0	0	2	3	3	1	1
SDF4	0.53	1	0	0	0	0	1	2	1	1	1	1
SOGA1	2.27	1	0	1	1	0	0	3	15	5	1	0.47
STARD10	1.34	1	0	2	0	0	0	3	4	5	1	1
TENM2	3.30	1	0	2	2	0	2	7	16	16	1	1
ZFHX4	1.01	0	0	3	3	0	0	6	14	8	1	1

The genes included in this analysis were among the strongest candidate genes affected by a DNM (either missense or LoF mutation). The missense Z-score is included here to indicate a relative (in)tolerance to missense mutation[22]. For the original NIJ/NCL discovery cohort, only the missense DNMs are included in this Table (7 of these genes were affected by a LoF DNM). A burden test was done to compare the total number of predicted pathogenic missense mutations observed in the infertile vs. fertile men, as well as between fertile men and fertile women (Fisher's Exact test, adjusted for multiple testing following Bonferroni correction).

ARTICLE

Deleterious variants in X-linked *CFAP47* induce asthenoteratozoospermia and primary male infertility

Chunyu Liu,[1,2,20] Chaofeng Tu,[3,4,5,20] Lingbo Wang,[1,2,6,20] Huan Wu,[7,8,9,20] Brendan J. Houston,[10,11] Francesco K. Mastrorosa,[12] Wen Zhang,[13] Ying Shen,[14] Jiaxiong Wang,[15,16] Shixiong Tian,[1,2] Lanlan Meng,[4] Jiangshan Cong,[1,2] Shenmin Yang,[15,16] Yiwen Jiang,[1] Shuyan Tang,[1,2] Yuyan Zeng,[1] Mingrong Lv,[7,8,9] Ge Lin,[3,4,5] Jinsong Li,[6] Hexige Saiyin,[1] Xiaojin He,[7,8,9] Li Jin,[1] Aminata Touré,[17] Pierre F. Ray,[17,18] Joris A. Veltman,[12] Qinghua Shi,[19,21] Moira K. O'Bryan,[10,11,21] Yunxia Cao,[7,8,9,21] Yue-Qiu Tan,[3,4,5,21,*] and Feng Zhang[1,2,21,*]

Summary

Asthenoteratozoospermia characterized by multiple morphological abnormalities of the flagella (MMAF) has been identified as a sub-type of male infertility. Recent progress has identified several MMAF-associated genes with an autosomal recessive inheritance in human affected individuals, but the etiology in approximately 40% of affected individuals remains unknown. Here, we conducted whole-exome sequencing (WES) and identified hemizygous missense variants in the X-linked *CFAP47* in three unrelated Chinese individuals with MMAF. These three *CFAP47* variants were absent in human control population genome databases and were predicted to be deleterious by multiple bioinformatic tools. *CFAP47* encodes a cilia- and flagella-associated protein that is highly expressed in testis. Immunoblotting and immunofluorescence assays revealed obviously reduced levels of CFAP47 in spermatozoa from all three men harboring deleterious missense variants of *CFAP47*. Furthermore, WES data from an additional cohort of severe asthenoteratozoospermic men originating from Australia permitted the identification of a hemizygous Xp21.1 deletion removing the entire *CFAP47* gene. All men harboring hemizygous *CFAP47* variants displayed typical MMAF phenotypes. We also generated a *Cfap47*-mutated mouse model, the adult males of which were sterile and presented with reduced sperm motility and abnormal flagellar morphology and movement. However, fertility could be rescued by the use of intra-cytoplasmic sperm injections (ICSIs). Altogether, our experimental observations in humans and mice demonstrate that hemizygous mutations in *CFAP47* can induce X-linked MMAF and asthenoteratozoospermia, for which good ICSI prognosis is suggested. These findings will provide important guidance for genetic counseling and assisted reproduction treatments.

Introduction

Infertility is a common medical condition affecting an estimated 15% of couples worldwide.[1] Males account for approximately 50% of infertile individuals, and the potential factors for male infertility are complex and diverse.[2] Asthenoteratozoospermia is the clinical descriptor used to describe men who produce sperm with abnormal morphology and reduced motility and encompasses a significant proportion of male infertility.[3]

Multiple morphological abnormalities of the flagella (MMAF) is a subtype of asthenoteratozoospermia proposed in 2014 and is characterized by the presence in the ejaculate of sperm with a combination of absent, short, coiled, bent, and/or irregular-caliber flagella.[4] Autosomal recessive inheritance has been suggested for many forms of human MMAF,

[1]Obstetrics and Gynecology Hospital, NHC Key Laboratory of Reproduction Regulation (Shanghai Institute of Planned Parenthood Research), State Key Laboratory of Genetic Engineering at School of Life Sciences, Fudan University, Shanghai 200011, China; [2]Shanghai Key Laboratory of Female Reproductive Endocrine Related Diseases, Shanghai 200011, China; [3]Institute of Reproductive and Stem Cell Engineering, School of Basic Medical Science, Central South University, Changsha 410000, China; [4]Reproductive and Genetic Hospital of CITIC-Xiangya, Changsha 410000, China; [5]Clinical Research Center for Reproduction and Genetics in Hunan Province, Changsha 410000, China; [6]State Key Laboratory of Cell Biology, Shanghai Key Laboratory of Molecular Andrology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China; [7]Reproductive Medicine Center, Department of Obstetrics and Gynecology, The First Affiliated Hospital of Anhui Medical University, Hefei 230022, China; [8]NHC Key Laboratory of Study on Abnormal Gametes and Reproductive Tract, Anhui Medical University, Hefei 230032, China; [9]Key Laboratory of Population Health Across Life Cycle, Anhui Medical University, Ministry of Education of the People's Republic of China, Hefei 230032, China; [10]School of Biological Sciences, Monash University, Clayton, VIC 3800, Australia; [11]School of BioSciences, The University of Melbourne, Parkville, VIC 3010, Australia; [12]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, NE2 4HH Newcastle upon Tyne, UK; [13]Fudan University Pudong Medical Center, Shanghai Key Laboratory of Medical Epigenetics, Institutes of Biomedical Sciences, Department of Systems Biology for Medicine, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China; [14]Department of Obstetrics/Gynecology, Key Laboratory of Obstetric, Gynecologic and Pediatric Diseases and Birth Defects of Ministry of Education, West China Second University Hospital, Sichuan University, Chengdu 610041, China; [15]State Key Laboratory of Reproductive Medicine, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou 215002, China; [16]Suzhou Municipal Hospital, Suzhou 215002, China; [17]Team Genetics Epigenetics and Therapies of Infertility, Institute for Advance Biosciences, Grenoble Alpes University, INSERM U1209, Centre National de la Recherche Scientifique UMR 5309, Grenoble 38000, France; [18]UM de genetique de l'infertilite et de diagnostic pre-implantatoire, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble 38000, France; [19]The First Affiliated Hospital of USTC, Hefei National Laboratory for Physical Sciences at Microscale, CAS Key Laboratory of Innate Immunity and Chronic Disease, School of Basic Medical Sciences, Division of Life Sciences and Medicine, CAS Center for Excellence in Molecular Cell Science, Collaborative Innovation Center of Genetics and Development, University of Science and Technology of China, Hefei 230027, China

[20]These authors contributed equally to this work

[21]These authors contributed equally to this work

*Correspondence: tanyueqiu@csu.edu.cn (Y.-Q.T.), zhangfeng@fudan.edu.cn (F.Z.)

https://doi.org/10.1016/j.ajhg.2021.01.002.

and to date, 22 autosomal MMAF-associated genes have been identified: *AK7* (MIM: 615364), *ARMC2* (MIM: 618424), *CEP135* (MIM: 611423), *CFAP43* (MIM: 617558), *CFAP44* (MIM: 617559), *CFAP58* (MIM: 619129), *CFAP65* (MIM: 614270), *CFAP69* (MIM: 617949), *CFAP70* (MIM: 618661), *CFAP91* (MIM: 609910), *CFAP251* (MIM: 618146), *DNAH1* (MIM: 603332), *DNAH2* (MIM: 603333), *DNAH6* (MIM: 603336), *DNAH8* (MIM: 603337), *DNAH17* (MIM: 610063), *DZIP1* (MIM: 608671), *FSIP2* (MIM: 618153), *QRICH2* (MIM: 618304), *SPEF2* (MIM: 610172), *TTC21A* (MIM: 611430), and *TTC29* (MIM: 618735).[5–9] Yet there are still many individuals with MMAF that cannot be causally diagnosed, indicating the potential involvement of other genetic factors.

Sex chromosomes display important roles in sex determination and fertility. Previous studies indicated that many genes specifically or preferentially expressed in the testis are enriched on sex chromosomes.[10,11] Thus, deleterious mutations in these genes may have a direct phenotypic effect on male fertility because of the lack of the second compensatory X or Y allele.[12] Although the roles of Y chromosomal microdeletions in male infertility have been well understood,[13] only a few X-linked genes have been found to be associated with infertile phenotypes. For example, hemizygous *TEX11* (MIM: 300311) mutations were shown to cause meiotic arrest and azoospermia and hemizygous *ADGRG2* (MIM: 300572) mutations led to obstructive azoospermia in a large Pakistani family.[14,15]

In this study, we identified hemizygous deleterious variants of X-linked *CFAP47* (also known as *CXorf22*) in four unrelated men displaying asthenoteratozoospermia. Furthermore, an X-linked *Cfap47*-mutated mouse model was generated with CRISPR-Cas9 technology, and the adult males with a hemizygous *Cfap47* mutation were sterile and also displayed reduced sperm motility and abnormal flagella. Intra-cytoplasmic sperm injection (ICSI) treatment led to successful fertilization using the spermatozoa from *Cfap47*-mutated male mice and three of the four men harboring hemizygous *CFAP47* variants. All these findings indicate that X-linked *CFAP47* mutations are an important pathogenic factor of MMAF and asthenoteratozoospermia.

Material and methods

Human subjects

The initial two Chinese cohorts were composed of 243 MMAF-affected Chinese men recruited from the Reproductive and Genetic Hospital of CITIC-Xiangya (Changsha, China) and 88 MMAF-affected individuals recruited from the First Affiliated Hospital of Anhui Medical University (Hefei, China). The third cohort comprised 24 individuals with asthenoteratozoospermia enrolled in Australia. The clinical phenotypes of the affected individuals are described in the supplemental information (see Supplemental note). The study regarding the cohorts was approved by the institutional review boards at all the participating institutes, and signed informed consents were obtained from all subjects participating in the study.

Whole-exome sequencing

Genomic DNA was isolated from peripheral blood samples of human subjects via a DNeasy Blood and Tissue Kit (QIAGEN, 51106). Whole-exome sequencing (WES) analysis was performed on MMAF-affected subjects as previously described.[16,17] Briefly, we used 1 μg of genomic DNA to enrich the human exome by using the Agilent SureSelect Human All Exon V6 Kit or Twist Bioscience's Human Core Exome Kit and sequenced this on the HiSeq 2000 or NovaSeq 6000 sequencing platforms (Illumina). The obtained data were mapped to the human genome reference assembly (GRCh37/hg19) by the Burrows-Wheeler Aligner (BWA) software,[18] and PCR duplicates were marked and removed via the Picard software. Then ANNOVAR software was used for functional annotation with information from a variety of databases and bioinformatic tools, including OMIM, Gene Ontology, SIFT, PolyPhen-2, 1000 Genomes Project, and gnomAD.[19–23] Deleterious missense variants were predicted simultaneously via SIFT, PolyPhen-2, CADD, and/or M-CAP. Sanger sequencing was conducted for variant verification with the primers listed in Table S1.

Structural modeling for CFAP47 and its mutants

The mutants Ser1742Gly and Ile2385Asn of CFAP47 are located near the Pfam or CDD motifs, hence their effect on protein structure could be modeled with homology models. The structures with homology domains near the mutants Ser1742Gly and Ile2385Asn were modeled by SWISS-MODEL. The putative homology model for the proline rich peptide (Tyr2884-Gln2905) was also built for analysis on the basis of similar proline rich templates searched by SWISS-MODEL. Furthermore, the molecular dynamics (MD) simulation of this peptide was performed by UCSF Chimera with default settings. The major clusters of simulated structures were compared with original model (one cluster comparison was displayed). In addition, we also used the online tool HOPE[24] to further predict the potential effects of *CFAP47* missense variants on the structure of CFAP47.

Detection and validation of genomic copy number variations

Copy number variation (CNV) calling was performed with a custom GATK4-based pipeline. This workflow exploits the GATK4 sequence read-depth normalization[25] and a custom R-based segmentation and visualization.[26] The detected CNVs were annotated with AnnotSV.[27] All the CNVs present in more than 1% of the samples of the Database of Genomic Variations (DGV) were excluded. The remaining rare deletions and duplications were individually inspected through the genomic profiles and detailed Log$_2$Ratio plots generated by the workflow. Manual inspection of these plots allowed for distinguishing possible CNVs from background noise, and only CNVs involving more than two exons were considered for further validation.

The detected deletion was validated with a PCR assay. We designed three primer pairs to amplify the region encompassed by the deletion, while we designed three other primer pairs to amplify the closest gene outside the predicted breakpoints (Table S2 and Figure S1). The assay was conducted on the MMAF-affected subject MA2603 II-1 and male and female control individuals.

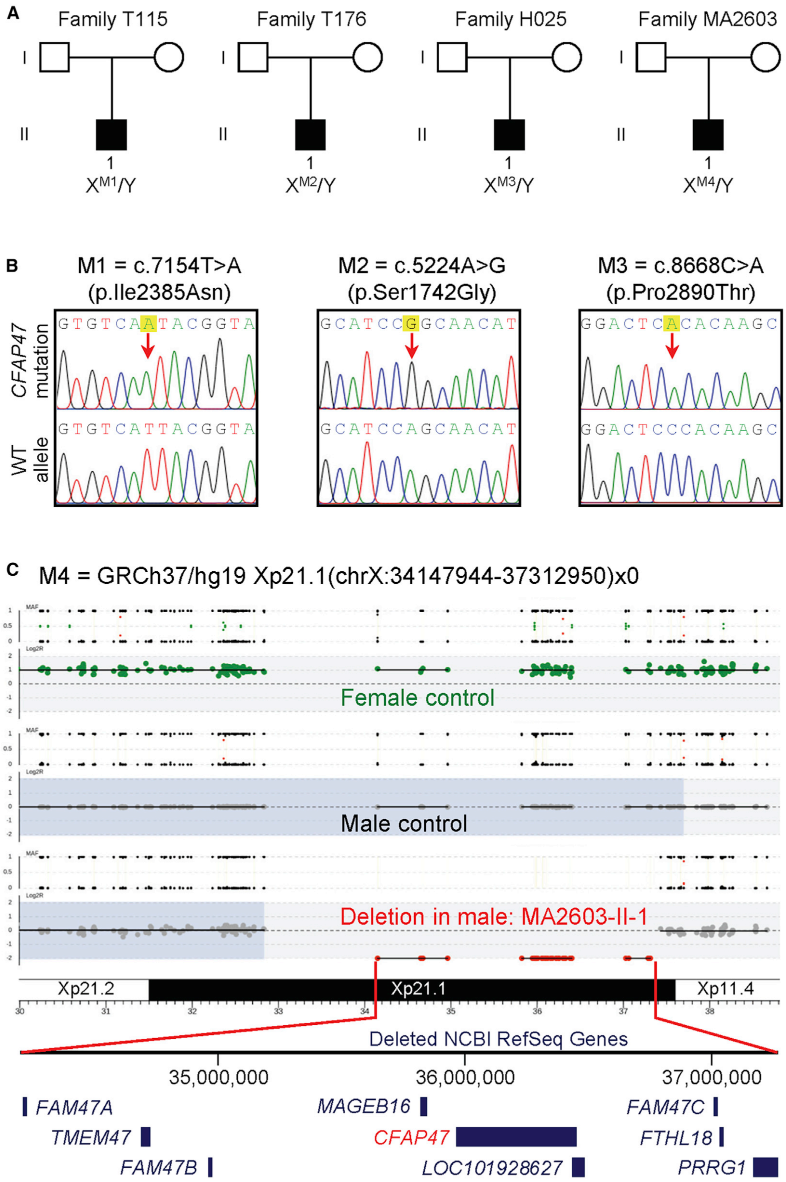


Figure 1. Identification of hemizygous variants of X-linked *CFAP47* in men with asthenoteratozoospermia

(A) Pedigrees of four families affected by hemizygous *CFAP47* variants (M1–M4). Black filled squares indicate the male individuals with asthenoteratozoospermia.

(B) Sanger sequencing confirmed hemizygous *CFAP47* missense variants (M1–M3) in subjects T115 II-1, T176 II-1, and H025 II-1, respectively. The positions of variants are indicated by red arrows. WT, wild type.

(C) An approximately 3.2-Mb Xp21.1 deletion affecting *CFAP47* (M4) in subject MA2603 II-1. This hemizygous deletion removed the entire *CFAP47* gene copy.

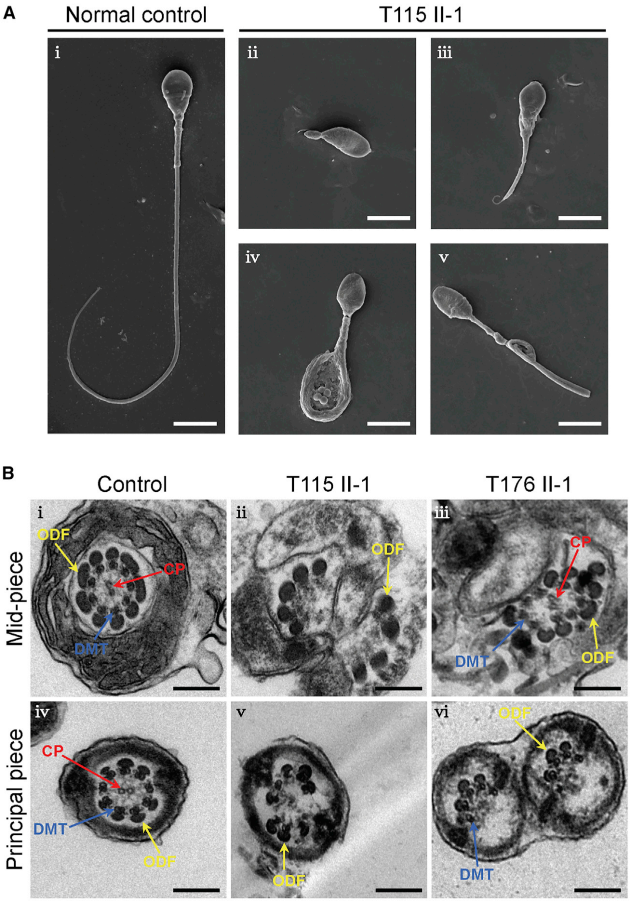


Figure 2. Sperm morphology and ultrastructure analyses for men harboring hemizygous *CFAP47* variants
(A) SEM analysis of the spermatozoa from a male control individual and men harboring hemizygous *CFAP47* variants. (i) Normal morphology of the spermatozoon from a healthy control male. (ii–v) Most spermatozoa from men harboring hemizygous *CFAP47* variants displayed typical MMAF phenotypes, including absent (ii), short (iii), coiled (iv), and bent flagella (v). The data of subject T115 II-1 were shown as an example. Scale bars: 5 μm.
(B) TEM analysis of the spermatozoa from a male control individual and men harboring hemizygous *CFAP47* variants. Cross-sections of the midpiece (i) and principal piece (iv) of the sperm flagella from a male control individual displayed typical "9+2" microtubule structure: the central pair of microtubules (CP; red arrows) and nine pairs of peripheral microtubule doublets (DMTs; blue arrows) surrounded by nine outer dense fibers (ODFs; yellow arrows). The organized mitochondrial sheath and fibrous sheath are also observed. Cross-sections of the spermatozoa from men harboring hemizygous *CFAP47* variants revealed various axonemal anomalies, including misarranged ODFs (ii, iii) and missing DMTs and/or the CP (v and vi). Scale bars: 200 nm.

Semen parameter analysis

Semen samples of human subjects were collected through masturbation after 2–7 days of sexual abstinence and analyzed in the source laboratories as part of the routine biological examination according to the 5th World Health Organization (WHO) guidelines. The morphology of the sperm cells was assessed with hematoxylin and eosin (H&E) staining and scanning electron microscopy (SEM). The morphological abnormalities of sperm flagella were classified into five categories: absent, short, bent, coiled flagella, and flagella of irregular caliber.[4] We examined at least 200 spermatozoa for each subject to evaluate the percentages of morphologically abnormal spermatozoa.

For sperm morphology and motility analyses of the mouse model, spermatozoa were extracted from the caput, corpus, and cauda epididymides through dissection of adult male mice and diluted in 1 mL human tubal fluid (HTF; Millipore, Cat. # MR-070-D) for 15 min at 37°C. Sperm morphology was analyzed by H&E staining and/or SEM, and sperm motility was further assessed with the spermatozoa from cauda epididymides by a computer-assisted sperm analysis (CASA) system.

Electron microscopy evaluation

For electron microscopy evaluation, semen samples were prepared as previously described.[28] In brief, for SEM assay, sperm specimens were deposited on poly-L-lysine-coated coverslips, fixed in 2.5% glutaraldehyde, washed in 0.1 mol/L phosphate buffer, and post-fixed in osmic acid. The specimens were then progressively dehydrated with ethanol and isoamyl acetate gradient, then dried with a CO_2 critical-point dryer (Eiko HCP-2, Hitachi). Next, the specimens were mounted on aluminum stubs, sputter coated by use of an ionic sprayer meter (Eiko E-1020, Hitachi), and analyzed via SEM (Stereoscan 260) under an accelerating voltage of 20 kV.

For transmission electron microscopy (TEM), semen samples were rinsed and immersed routinely and then were progressively dehydrated with graded ethanol (50%, 70%, 90%, and 100%) and 100% acetone, followed by infiltration with 1:1 acetone and SPI-Chem resin overnight at 37°C. After being embedded in Epon 812, the specimens were sliced with ultra-microtome, stained with uranyl acetate and lead citrate, and observed and photographed via TEM (TECNAI-10, Philips) with an accelerating voltage of 80 kV. For TEM analysis of mouse sperm, cauda epididymis samples were prepared as described previously.[16]

Table 1. Hemizygous deleterious *CFAP47* variants identified in Chinese MMAF-affected men

***CFAP47* variant**	**M1**	**M2**	**M3**
cDNA alteration	c.7154T>A	c.5224A>G	c.8668C>A
Variant allele	hemizygous	hemizygous	hemizygous
Protein alteration	p.Ile2385Asn	p.Ser1742Gly	p.Pro2890Thr
Variant type	missense	missense	missense
Allele frequency in human population			
1000 Genomes Project	0	0	0
East Asians in gnomAD	0	0	0
All individuals in gnomAD	0	0	0
Function prediction			
SIFT	damaging	damaging	damaging
PolyPhen-2	damaging	N/A	damaging
M-CAP	N/A	damaging	damaging
CADD	8.6	23.4	23.1

NCBI reference sequence number of *CFAP47* is GenBank: NM_001304548.2. Variants with CADD values greater than 4 are considered to be deleterious. N/A, not available.

Mouse model generation

Cfap47-mutated mice were generated with CRISPR-Cas9 technology. Cas9 and sgRNA were prepared as previously described.[29] The CRISPR-Cas9 reagents were directly injected into zygotes of C57BL/6 mice. After injection, the zygotes were further cultured in KSOM medium (Millipore, Cat. # MR-106-D) at 37°C under 5% CO_2 to reach the 2-cell stage, followed by embryo transfer into oviducts of female pseudopregnant Institute of Cancer Research (ICR) mice at 0.5 days post-coitum (dpc). We used PCR assay and Sanger sequencing to identify the frameshift mutation in founder mice (Table S3). Adult mice (aged 7 weeks or older) were used in this study. All animal experiments were carried out in accordance with the recommendations of the US National Institutes of Health's Guide for the Care and Use of Laboratory Animals. The study was approved by the animal ethics committee at the School of Life Sciences of Fudan University.

Real-time quantitative PCR and reverse-transcription PCR

For real-time quantitative PCR, total RNAs of human spermatozoa and mouse testes were extracted with the Allprep DNA/RNA/Protein Mini Kit (QIAGEN). Approximately 1 μg of obtained RNA was converted into cDNA with HiScript II Q RT SuperMix for quantitative PCR (Vazyme). The obtained cDNA was individually diluted 5-fold to be used as templates for the subsequent real-time quantitative PCR with AceQ quantitative PCR SYBR Green Master Mix (Vazyme) on a CFX Connect Real-Time PCR Detection System. *GAPDH*/*Gapdh* was used as an internal control, and primers for real-time quantitative PCR are listed in Table S4.

For reverse-transcription PCR (RT-PCR), total RNAs of various tissues of adult C57BL/6N mice were extracted and reverse transcribed as described above. RT-PCR was performed with 10 ng of cDNA, and *Hprt* was used as an internal control (Figure S2 and Table S4).

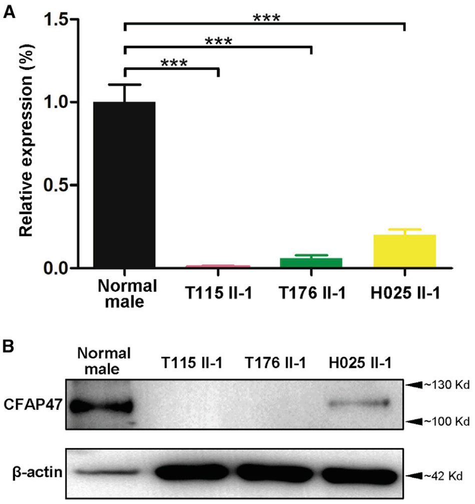


Figure 3. Expression analysis of *CFAP47* mRNA and CFAP47 in the spermatozoa from a male control individual and men harboring hemizygous *CFAP47* variants

(A) Real-time quantitative PCR analysis indicated that the abundance of *CFAP47* mRNA was dramatically reduced in the sperm from men harboring hemizygous *CFAP47* variants when compared to that of a control male. Data represent the means ± standard error of measurement of three independent experiments. Two-tailed Student's paired or unpaired t tests were used as appropriate (***$p < 0.001$).

(B) Immunoblotting assay revealed that CFAP47 was dramatically reduced or nearly absent in the spermatozoa from men harboring *CFAP47* mutations. β-actin was used as a loading control.

Immunoblotting

The proteins of human sperm cells were extracted via Minute Total Protein Extraction Kit for Animal Cultured Cells and Tissues (Invent) then denatured at 95°C for 10 min. The denatured proteins were separated by 10% sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred onto polyvinylidene difluoride (PVDF) membrane (Millipore). Membranes were blocked in 5% non-fat milk for 1 h at room temperature before incubation overnight at 4°C with the following primary antibodies: rabbit polyclonal anti-CXorf22 (i.e., anti-CFAP47; GTX80633, GeneTex, 1:1000) and HRP-conjugated beta actin (HRP-60008, Proteintech, 1:2000). The membranes were then washed in TBST (Tris-buffered saline with Tween-20) three times and incubated with HRP-conjugated anti-Rabbit IgG antibody (Abmart, M21002) at a 1:2500 dilution in blocking solution for 1 h at room temperature. We used the Chemistar High-sig ECL Western Blotting Substrate (Tanon) to detect immunoreactive protein bands by Tanon 5200.

Immunoprecipitations

For immunoprecipitations, proteins were extracted from the sperm cells of a control man and the testes of adult wild-type mice and incubated with 5 μg of a customized CFAP65 antibody (Abclonal, China, specifically binding the amino acids 1401–1635 of mouse Cfap65) overnight at 4°C. Next, 50 μL of Protein

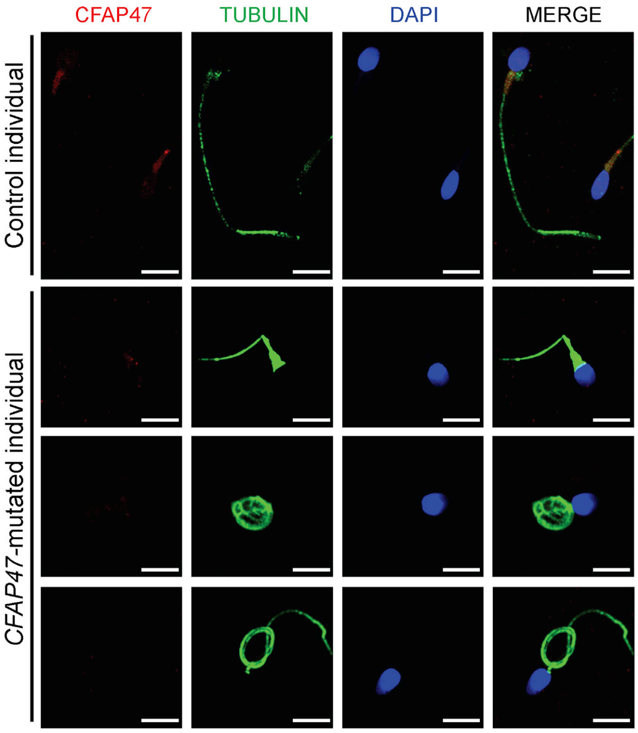


Figure 4. Immunofluorescence staining of CFAP47 in the spermatozoa from a male control individual and men harboring hemizygous *CFAP47* variants
Sperm cells were stained with anti-CFAP47 (red) and anti-α-tubulin (green) antibodies. DNA was counterstained with DAPI (4′,6-diamidino-2-phenylindole) as a marker of the cell nucleus. CFAP47 staining is concentrated at the base of sperm flagella from the control individual, but the signal was almost absent in the sperm flagella from men harboring hemizygous *CFAP47* variants. Scale bars: 5 μm.

A/G Magnetic Beads (Pierce Biotechnology, Rockford, IL, USA) was added to each incubation sample for 1 h at room temperature on a rotating mixer. The beads were then washed three times with the manufacturer's immunoprecipitation lysis and wash buffer. Finally, the co-immunoprecipitated proteins were analyzed by immunoblotting with CFAP65 (diluted at 1:1000) and CXorf22 (i.e., CFAP47; diluted at 1:1000) antibodies.

Immunofluorescence analysis

For immunofluorescence localization of proteins, sperm cells obtained from human MMAF-affected individuals and control subjects were washed in phosphate-buffered saline (PBS), fixed in 4% paraformaldehyde for 30 min at room temperature, and coated on the slides treated with 0.1% poly-L-lysine precoated slides (Thermo Fisher). Non-specific antibody binding was blocked in 10% donkey serum for 1 h at room temperature, and sperm were incubated overnight at 4°C with the following primary antibodies: rabbit polyclonal anti-CXorf22 (GTX80633, GeneTex, 1:100), anti-CFAP65 (CSB-PA757963LA01HU, CUSABIO, 1:100), anti-SPAG16 (PA5-57995, Invitrogen, 1:100), and monoclonal mouse anti-α-tubulin (T9026, Sigma, 1:500). Next, the slides were washed with PBS with 0.1% (v/v) Tween-20 before 1 h incubation at room temperature with the highly cross-absorbed secondary antibodies Alexa Fluor 488 anti-Mouse IgG (34106ES60, Yeasen, 1:1000) and Cy3-conjugated AffiniPure Goat Anti-Rabbit IgG (111-165-003, Jackson, 1:4000). The images were captured with a confocal microscope (Zeiss LSM 880).

In vitro fertilization and ICSI in mice

In vitro fertilization (IVF) and ICSI analyses in mice were conducted as previously described.[17] In brief, the wild-type female mice were superovulated by injection of 5–7.5 IU of pregnant mare serum gonadotropin (PMSG), followed by 5–7.5 IU of human chorionic gonadotropin (hCG) 48 h later. For IVF, sperm samples collected from mouse cauda epididymides were added into the HTF drop. Next, cumulus-intact oocytes collected from superovulated female mice were transferred into the sperm-containing HTF drop. After incubation for 5–6 h, mouse embryos were washed in HTF and transferred into KSOM medium (Millipore, Cat. # MR-106-D) for further culture (37°C, 5% CO_2). We then evaluated fertilization rates by recording the numbers of two-cell embryos and late-stage blastocysts at 20 h and 91 h later, respectively. For ICSI, mouse sperm heads were separated from sperm tails and injected into mouse oocytes obtained from superovulated females by a Piezo driven pipette as previously described.[30] Then the injected oocytes were cultured in KSOM medium at 37°C under 5% CO_2. Two-cell embryos and blastocysts were counted 20 h and 96 h later, respectively.

Results

Identification of rare and hemizygous *CFAP47* variants in men with asthenoteratozoospermia

In this study, we combined SNV and CNV calling from WES data to identify potential candidate genes associated with MMAF. Hemizygous missense variants of X-linked *CFAP47* were initially identified in three Chinese MMAF-affected individuals from unrelated families in the Chinese cohorts: c.7154T>A (p.Ile2385Asn) in subject T115 II-1, c.5224A>G (p.Ser1742Gly) in subject T176 II-1, and c.8668C>A (p.Pro2890Thr) in subject H025 II-1 (Figures 1A, 1B, and 2 and Table 1). These hemizygous *CFAP47* variants were absent in the human population genome

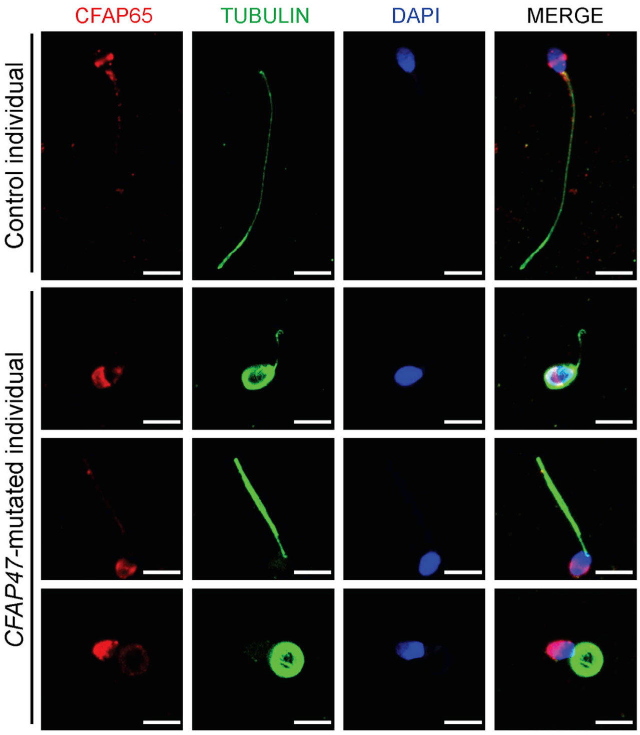


Figure 5. Abnormal CFAP65 location in the spermatozoa from men harboring hemizygous *CFAP47* variants
Sperm cells were stained with anti-CFAP65 (red) and anti-α-tubulin (green) antibodies. DNA was counterstained with DAPI as a marker of the cell nucleus. CFAP65 staining was observed mainly at the equatorial segment of sperm head and the base of flagella in control spermatozoa, but the signal was only diffusely clustered in the acrosome of spermatozoa from men harboring hemizygous *CFAP47* variants. Scale bars: 5 μm.

datasets, including 1000 Genomes Project and gnomAD (v2.1.1 with 141,456 samples) (Table 1). All these *CFAP47* missense variants were predicted to be deleterious through utilization of the PolyPhen-2, SIFT, CADD, and M-CAP tools (Table 1).

CFAP47 (formerly known as *CXorf22*, GenBank: NM_001304548.2) is located on the human chromosome X and highly expressed in the testis. *CFAP47* encodes a cilia- and flagella-associated protein. The residues in CFAP47 affected by *CFAP47* missense variants p.Ile2385Asn (subject T115 II-1), p.Ser1742Gly (subject T176 II-1), and p.Pro2890Thr (subject H025 II-1) are all highly conserved across species (Figure S3). Further analysis of protein structural modeling via online bioinformatic tools revealed the severe effects of these amino acid-substituting mutations on the structure and/or stability of CFAP47. These included changes in the hydrophobicity in CFAP47 mutant p.Ile2385Asn, the structure flexibility in CFAP47 mutant p.Ser1742Gly, and backbone flexibility or residue charge in CFAP47 mutant p.Pro2890Thr. These findings indicated the potential contribution of these *CFAP47* missense variants to MMAF (Figure S4 and Table S5).

Furthermore, CNV calling from WES data permitted the identification of a hemizygous Xp21.1 deletion eliminating the entire *CFAP47* in subject MA2603 II-1 from the Australian cohort (Figure 1C). PCR analysis with primer pairs for the region encompassed by the deletion and outside the predicted breakpoints further confirmed the deletion segment (Figure S1). No genic or exonic deletions of *CFAP47* were reported in either DGV (14,316 samples) or gnomAD-SV (v2.1 with 10,847 samples). These findings further suggest the involvement of deleterious *CFAP47* variants in MMAF across human populations.

Deficiency of CFAP47 and its association with CFAP65

To investigate the pathogenicity of the *CFAP47* variants identified in this study, we analyzed the expression of *CFAP47* mRNA and protein in the sperm samples from a fertile control individual and the men harboring hemizygous *CFAP47* variants. Real-time quantitative PCR assays suggested that the abundances of *CFAP47* mRNA were significantly reduced in the spermatozoa from men harboring hemizygous *CFAP47* variants (Figure 3A). Consistently, as shown by immunoblotting and immunostaining assays, the signal of CFAP47 was dramatically reduced, or absent, in the spermatozoa from men harboring hemizygous *CFAP47* variants (Figures 3B and 4). Thus, MMAF phenotypes described in this study were most likely caused by hemizygous deleterious variants in *CFAP47*.

Previous studies have reported that the deficiency of another cilia- and flagella-associated protein, CFAP65, can cause male infertility with MMAF.[28,31–33] Furthermore, STRING analysis indicates that CFAP47 may be highly connected with CFAP65 (Figure S5). To investigate the potential association between these two proteins, we performed immunostaining assay by using a commercial antibody against CFAP65 on the spermatozoa from men harboring hemizygous *CFAP47* variants. We observed that CFAP65 immunostaining was localized mainly at the equatorial segment of sperm head and the base of flagella in normal

Table 2. Semen characteristics and sperm morphology in men harboring hemizygous *CFAP47* variants

Subject	T115 II-1	T176 II-1	H025 II-1	MA2603 II-1	Reference limits
Semen parameter					
Semen volume (mL)	4.0	0.9	5.2	5.6	1.5[a]
Sperm concentration (10^6/mL)	2.5	3.0	29.5	0.5	15.0[a]
Motility (%)	18.9	14.3	23.3	10.0	40.0[a]
Progressive motility (%)	7.1	8.1	18.5	5.0	32.0[a]
Sperm morphology					
Absent flagella (%)	37.5	11.0	22.0	N/A	5.0[b]
Short flagella (%)	34.0	17.0	14.0	N/A	1.0[b]
Coiled flagella (%)	18.5	42.0	50.5	N/A	17.0[b]
Angulation (%)	3.5	5.0	2.0	N/A	13.0[b]
Irregular caliber (%)	6.5	6.5	3.5	N/A	2.0[b]

N/A, not available.
[a] Reference limits according to the WHO standards.[35]
[b] Reference limits according to the distribution range of morphologically normal spermatozoa observed in 926 fertile individuals.[36]

spermatozoa from control individuals, whereas the CFAP65 signal was only diffusely clustered in the acrosome of spermatozoa from MMAF-affected individuals harboring hemizygous *CFAP47* variants (Figure 5). Importantly, the abnormal location of CFAP65 was not observed in sperm cells from MMAF-affected individuals with other genetic defects such as bi-allelic mutations in *DNAH8*, *SPEF2*, or *CFAP58* (Figure S6). In addition, sperm cells from men harboring bi-allelic *CFAP65* variants displayed a significantly reduced CFAP47 staining when compared with the obvious CFAP47 signal in the mid-piece of flagella in normal spermatozoa (Figure S7). Consistently, obvious CFAP47/Cfap47 signals were also observed when we immunoprecipitated CFAP65/Cfap65 from human spermatozoa or mouse testis lysates and immunoblotted with CFAP47/Cfap47 antibodies (Figure S8). These findings collectively confirm a potential interaction and/or inter-regulation between CFAP65 and CFAP47 during spermiogenesis.

Asthenoteratozoospermia phenotypes in men harboring hemizygous *CFAP47* variants

Semen parameters of men harboring hemizygous *CFAP47* variants were acquired from the source laboratories during routine examination of the individuals according to WHO guidelines.[34] A semen analysis indicated the severely reduced sperm motility in all of four men harboring hemizygous *CFAP47* variants (Table 2). Furthermore, the rates of sperm progressive motility were dramatically decreased in men harboring hemizygous *CFAP47* variants (Table 2). Sperm morphological study was conducted by H&E staining and SEM. In comparison to long and thin tails of the sperm obtained from a fertile control man, the spermatozoa from men harboring hemizygous *CFAP47* variants displayed frequently abnormal flagella, including absent, short, coiled flagella, and irregular caliber (Figure 2A and Table 2).

We further conducted TEM to investigate sperm flagellar ultrastructure in the two Chinese men harboring hemizygous *CFAP47* variants. In contrast to typical "9+2" axoneme microtubule structure in the sperm flagella from a control specimen, the sperm flagella of men harboring hemizygous *CFAP47* variants displayed a variety of ultrastructural defects, including disorganization of outer dense fibers, and absence of peripheral or central microtubules at the midpiece and principal piece of sperm flagella (Figure 2B). In addition, the deficiency of SPAG16 (a component of core axoneme complex) was revealed, further suggesting the defect of core axoneme in the sperm flagella from men harboring hemizygous *CFAP47* variants (Figure S9).

Asthenoteratozoospermia phenotypes in *Cfap47*-mutated male mice

As shown in Figure S3, CFAP47 is conserved among mammalian species. Our RT-PCR assays using various mouse tissues indicated the preferential expression of mouse *Cfap47* (also located on the X chromosome) in the testis (Figure S2). To further investigate the role of mouse CFAP47 in sperm flagellar formation, we constructed *Cfap47*-mutated mice by using the CRISPR-Cas9 system (Figure S10A). Sanger sequencing of *Cfap47*-mutated male mice (*Cfap47*$^-$/Y) confirmed the presence of a hemizygous frameshift mutation (c.2559insT), which was predicted to cause premature translational termination (p.Gly855Profs*8) (Figure S10B). Real-time quantitative PCR assays showed that the abundance of *Cfap47* mRNA was significantly reduced in the spermatozoa from *Cfap47*$^-$/Y male mice when compared with wild-type male (*Cfap47*$^+$/Y) controls (Figure S11A). Consistently, immunostaining analysis performed with CFAP47 antibody revealed that CFAP47 staining was mainly located at the mid-piece of sperm flagella from wild-type male mice, but CFAP47 signal was almost absent in the spermatozoa from *Cfap47*-mutated male mice (Figure S11B).

Sperm morphology and parameters, together with flagellar ultra-structure, were also investigated in *Cfap47*-mutated male mice. Notably, diminished and abnormal sperm movements (a "folding" motion) were observed in *Cfap47*-mutated mice (Videos S1 and S2). CASA analyses indicated that both total sperm motility and progressive motility were significantly reduced in *Cfap47*-mutated male mice when compared to those of wild-type male mice (Figures 6A and 6B and Table S6). Furthermore, the rates of immotile and non-progressive spermatozoa from *Cfap47*-mutated male mice were significantly higher than

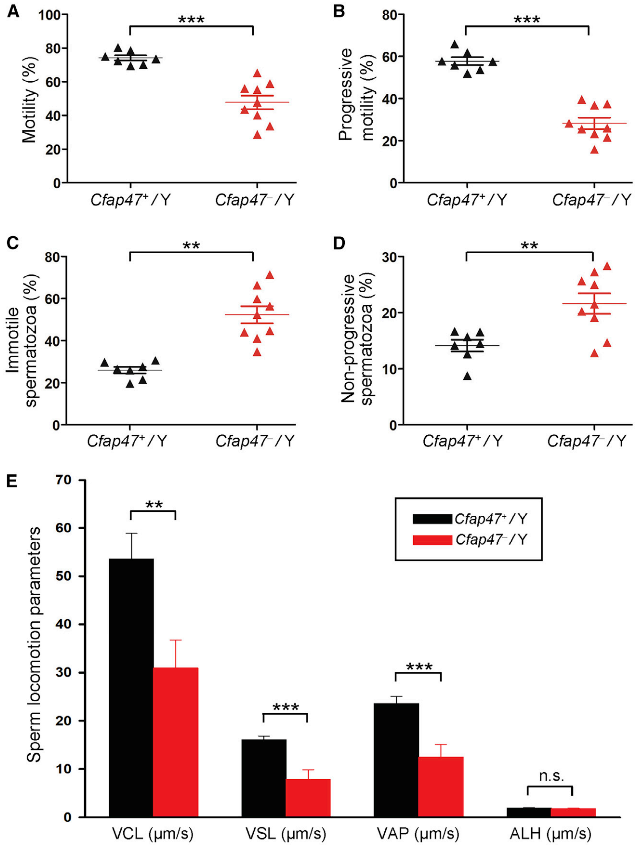


Figure 6. Semen characteristics of *Cfap47*-mutated male mice

(A–D) Semen characteristics by computer-assisted analysis system revealed significantly lower rates of sperm motility (A) and progressive motility (B) but significantly higher rates of immotile spermatozoa (C) and non-progressive spermatozoa (D) in *Cfap47*-mutated (*Cfap47*$^{-}$/Y) male mice when compared with those in wild-type (*Cfap47*$^{+}$/Y) male mice.
(E) Curvilinear velocity (VCL), straight-line velocity (VSL), and average-path velocity (VAP) were also dramatically reduced in *Cfap47*-mutated male mice. Error bars represent the standard error of the mean. $^{**}p < 0.01$; $^{***}p < 0.001$; n.s., not significant (Student's t test). ALH, amplitude of lateral head displacement.

those of wild-type male mice (Figures 6C and 6D). Additionally, three kinematic parameters, including curvilinear velocity (VCL), straight-line velocity (VSL), and average-path velocity (VAP), were significantly lower in *Cfap47*-mutated male mice than in wild-type male mice (Figure 6E). Regarding sperm concentration, there was no obvious difference between wild-type and *Cfap47*-mutated male mice (Table S6). H&E staining and SEM assay revealed that bent flagella were frequently observed in *Cfap47*-mutated male mice (Figure 7). Remarkably, the bent flagella most likely appeared first in the corpus of epididymis, but not in the testis of *Cfap47*-mutated male mice (Videos S3–S8 and Figure S12). Furthermore, although no obvious abnormal ultrastructure was detected in cross sections of the spermatozoa from *Cfap47*-mutated male mice, a higher rate of unevenly distributed fibrous sheaths was observed in the spermatozoa from *Cfap47*-mutated male mice than in those from wild-type male mice (Figures S13 and S14). These experimental observations suggested the possibility of abnormal sperm maturation in the epididymis of *Cfap47*-mutated male mice and/or sperm flagellar frangibility to mechanical stress due to the potential abnormality in flagellar assembly.

Damaged male fertility of *Cfap47*-mutated mice can be rescued by ICSI treatment

To investigate the fertility and reproductive behavior of *Cfap47*-mutated mice, we caged the sexually mature (6 weeks of age or older) wild-type and *Cfap47*-mutated male mice to age-matched wild-type females (one male with two females) for 3 months and counted the numbers of pups per litter. As shown in Figure S15, wild-type male mice routinely produced offspring (3–4 litters produced per female, 8 ± 0.82 pups per litter), whereas *Cfap47*-mutated male mice failed to produce any offspring over 3 months of breeding. We also performed IVF analysis to further investigate the fertility of *Cfap47*-mutated male mice. As expected, the rates of two-cell embryos and blastocysts were significantly lower in the group using the spermatozoa from *Cfap47*-mutated male mice than those of the wild-type male control group (Figure 8A). All these findings indicated that *Cfap47* deficiency causes male infertility in mice.

ICSI treatment has been suggested as an effective way to circumvent the physical limitations experienced by MMAF sperm.[37] To investigate whether mouse *Cfap47*-associated asthenoteratozoospermia could be overcome via ICSI, we performed ICSI by using spermatozoa from wild-type and *Cfap47*-mutated male mice. Notably, the rates of two-cell embryos and blastocysts generated by *Cfap47*-mutated male mice were comparable to those generated by wild-type male mice (Figure 8B). Similarly, three out of the four men harboring hemizygous *CFAP47* variants in this study achieved good clinical outcomes after assisted

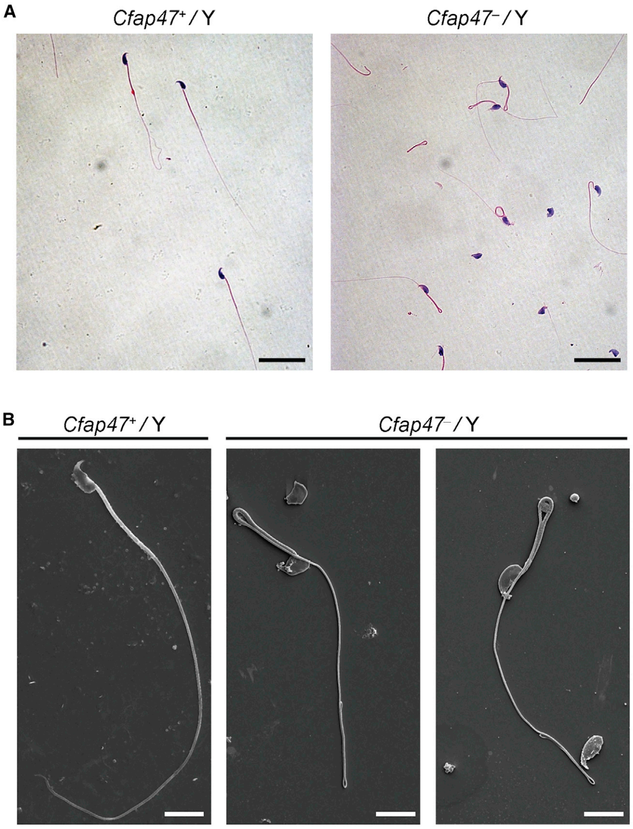


Figure 7. Sperm morphology analysis for *Cfap47*-mutated male mice
H&E staining (A) and SEM analyses (B) revealed higher rates of bent flagella in the spermatozoa of *Cfap47*-mutated (*Cfap47*⁻/Y) male mice when compared with those of wild-type (*Cfap47*⁺/Y) male mice. Scale bars: 20 μm (A) and 5 μm (B).

reproductive therapy by ICSI (Table 3). The fourth couple (subject T115 II-1 and his wife) could not achieve a pregnancy, potentially because of additional female risk factors of infertility (e.g., advanced reproductive age and poor oocyte quality). Therefore, our studies strongly indicated that the damaged male fertility caused by hemizygous *CFAP47* variants can be rescued by ICSI treatment.

Discussion

Since the initial identification of *DNAH1* in 2014 as a pathogenic gene in human asthenoteratozoospermia with MMAF, only an autosomal recessive inheritance has been proposed for MMAF,[4] and an additional 21 such autosomal genes have been confirmed to cause human MMAF.[4–9] Here, we demonstrated that MMAF can be caused by hemizygous mutations in an X-linked gene. We report that hemizygous deleterious variants of X-linked *CFAP47* induce severe asthenoteratozoospermia in four unrelated MMAF-affected individuals from different geographical origins. All these deleterious variants identified in *CFAP47* are absent from human population genome datasets archived in the 1000 Genomes Project, gnomAD, or DGV (Figure 1 and Table 1). Functional experiments further indicated the deficiency of CFAP47 in the spermatozoa from men harboring hemizygous *CFAP47* variants. All these findings indicate that *CFAP47* is an X-linked MMAF-associated gene. Analogously, the X-linked *AKAP4* (MIM: 300185) was previously found to be deleted together with an *AKAP3* deletion on the human chromosome 12 in a man with sperm fibrous sheath dysplasia, but no further molecular evidence of the transcripts or proteins was available to confirm the pathogenicity of *AKAP3* deletion and/or *AKAP4* deletion.[38] Therefore, X-linked *AKAP4* may be also a candidate gene for MMAF, but so far this has not been formally demonstrated.

To date, we do not have cues explaining the molecular mechanism by which deleterious *CFAP47* variants cause abnormal sperm morphology and/or motility. Notably, a potential interaction between CFAP65 and CFAP47 was predicted by the *in silico* tool STRING (Figure S5). Immunostaining assays performed in this study also revealed abnormal localization of CFAP65 in the spermatozoa from *CFAP47*-mutated men and the loss of CFAP47 staining in the sperm cells from *CFAP65*-mutated men. Moreover, co-immunoprecipitation assays performed on the spermatozoa from a fertile control man and the testes of wild-type male mice further support a functional link between these two cilia- and flagella-associated proteins. Previous studies also reported that *CFAP65* deficiency caused severe asthenoteratozoospermia in mice and humans.[28,31–33] Moreover, CFAP65 was found to participate in calcium-mediated retrograde signaling affecting cell differentiation and proliferation.[32,39] Thus, the reduced sperm motility and higher rates of abnormal sperm morphologies in men harboring hemizygous *CFAP47* variants may partially result from the abnormal interaction or regulation between CFAP47 and CFAP65.

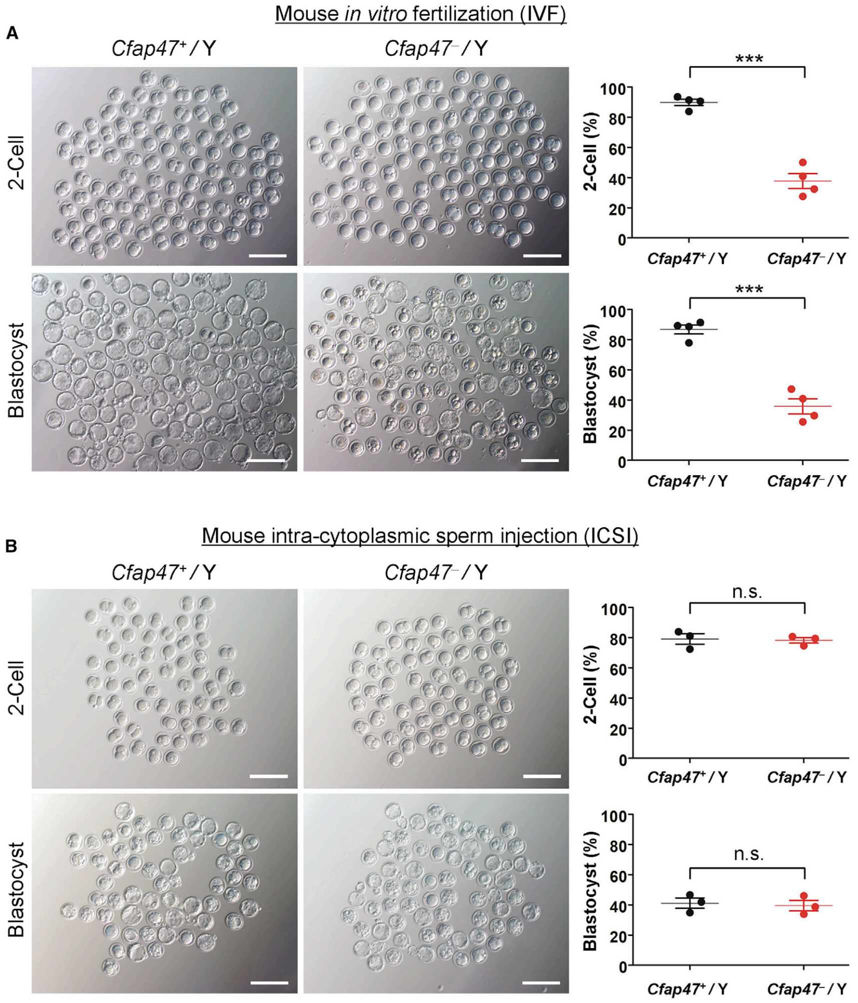


Figure 8. Damaged fertilization capability by deficiency of mouse CFAP47 could be rescued by ICSI

(A) Representative two-cell embryos and blastocysts from IVF in mice. The rates of both two-cell embryos and blastocysts were significantly lower in the group using the spermatozoa from *Cfap47*-mutated (*Cfap47*$^{-}$/Y) male mice than those in the wild-type (*Cfap47*$^{+}$/Y) group. Both groups consisted of four male mice. A total of 716 mouse embryos were counted. Data are represented as means ± standard error of measurement; ***$p < 0.001$. Scale bars: 200 μm.

(B) Representative two-cell embryos and blastocysts from ICSI in mice. The rates of two-cell embryos and blastocysts were counted after sperm heads were injected into oocytes that were collected from superovulated wild-type female mice. Both the *Cfap47*$^{+}$/Y and *Cfap47*$^{-}$/Y groups consisted of three male mice. A total of 367 mouse embryos were counted. Data are represented as means ± standard error of measurement. n.s., not significant. Scale bars: 200 μm.

We also constructed *Cfap47*-mutated mice in this study to further explore the effect of CFAP47 deficiency on sperm flagellar formation. Consistent with clinical presentation of men harboring hemizygous deleterious variants in *CFAP47*, *Cfap47*-mutated male mice were sterile and displayed reduced sperm motility. Light microscopy analysis revealed a higher rate of bent flagella in *Cfap47*-mutated male mice than in wild-type mice, a phenotype first discovered in the corpus epididymides, but not in the testis of the *Cfap47*-mutated male mice. It is well-established

Table 3. Clinical outcomes of ICSI cycles using the spermatozoa from men harboring hemizygous *CFAP47* variants

Subject	T115 II-1	T176 II-1	H025 II-1	MA2603 II-1
Male age (years)	37	26	34	31
Female age (years)	41	26	32	N/A
Number of ICSI cycles	1	1	1	1
Number of oocytes injected	1	9	11	12
Number (and rate) of fertilized oocytes	0 (0%)	8 (89%)	9 (82%)	10 (83%)
Number (and rate) of cleavage embryos	–	8 (100%)	9 (100%)	N/A
Number (and rate) of 8-cells	–	5 (63%)	7 (78%)	N/A
Number (and rate) of blastocysts	–	5 (63%)	7 (78%)	3 (30%)
Number of transfer cycles	–	1	1	3
Number of embryos transferred per cycle	–	2	2	1
Implantation rate	–	100%	50%	66%
Clinical pregnancy rate	–	100%	100%	100%
Miscarriage rate	–	0%	0%	33%

–, not applicable; N/A, not available.

that the epididymis constitutes a critical place for spermatozoa to acquire their fertilizing ability and, in particular, forward motility properties.[40] Thus, a possible impairment of sperm maturation may also occur as a result of the deficiency of mouse CFAP47. In addition, TEM assessment revealed a higher rate of unevenly distributed fibrous sheaths in the spermatozoa from *Cfap47*-mutated male mice than those from wild-type male mice, indicating potential abnormalities in flagellar assembly. It is possible that those fibrous sheath defects could confer the fragility to the sperm flagella, which may further induce bending during epididymal transit and contribute to the compromised sperm motility.[41]

Overall, the sperm phenotype of *Cfap47*-mutated male mice appears milder than the phenotype of men harboring hemizygous *CFAP47* variants. This could be due to evolutionarily divergent protein regulatory networks of CFAP47 or distinct compensatory mechanisms between species or differences in environmental exposure. Nonetheless, our investigations using CASA demonstrate significantly diminished sperm motility and abnormal sperm movements in *Cfap47*-mutated male mice. As such, the cause of sterility in *Cfap47*-mutated male mice is almost certainly due to the inability of sperm to ascend the female reproductive tract and reach the site of fertilization.

ICSI has become an effective method to help infertile couples to achieve a successful pregnancy. For infertile men with MMAF, ICSI treatment constitutes the only choice because of the constitutive flagellar defects.[42] Previous studies and our recent works revealed the good prognosis of ICSI for a series of MMAF-related genes. For example, MMAF-affected individuals with bi-allelic variants in *DNAH1*, *DNAH8*, or *TTC29* have good clinical outcomes following ICSI, while failed pregnancies were reported for *CEP135* (due to abnormal centriole assembly)- or *DNAH17* (due to unknown reason)-associated MMAF.[7,17,37,43,44] In this study, despite the reduction in fertilization observed in mouse IVF using the spermatozoa from *Cfap47*-mutated male mice, the fertilization (two-cell embryo) rate and blastocyst rate when performing ICSI for *Cfap47*-mutated male mice were similar to those observed for wild-type males. Moreover, men harboring hemizygous *CFAP47* variants in this study (except for subject T115 II-1, whose wife had poor oocyte quality) acquired successful pregnancies after ICSI treatment, further supporting that ICSI can be recommended for *CFAP47*-associated asthenoteratozoospermia.

Importantly, while both men harboring deleterious variants in X-linked *CFAP47* and those harboring bi-allelic mutations in autosomal MMAF-associated genes can have offspring via ICSI treatment, different rates of mutation carriers and different risks of male infertility in the offspring are expected between those two modes of inheritance. For the affected men whose MMAF is caused by autosomal recessive variants, both male and female offspring will be fertile heterozygous carriers who may ultimately transmit the autosomal recessive variants to next generations at a 50% probability. Assuming that there are no consanguineous conceptions, in this case of autosomal MMAF loci, the recurrence risk of MMAF will be very low for the offspring. In contrast, for ICSI performed in case of X-linked *CFAP47*-associated MMAF, 100% of female (but no male) offspring will inherit heterozygous deleterious variants of *CFAP47*. These female heterozygous carrier offspring will transmit the X-linked *CFAP47* variants at a 50% probability to the next generations, among which the females will be heterozygous carriers and the males harboring hemizygous *CFAP47* variants will be infertile. Therefore, the findings of this study are meaningful to genetic counseling for X-linked MMAF and asthenoteratozoospermia before ICSI treatment.

In conclusion, our experimental observations on both human subjects and the mouse model indicate that hemizygous *CFAP47* variants can induce MMAF-associated asthenoteratozoospermia. The observed functional associations between CFAP47 and CFAP65 indicated that cilia- and flagella-associated proteins may exist in a mutually regulated manner during spermatogenesis. Furthermore, good pregnancy outcomes could be achieved through ICSI treatment using the spermatozoa from men harboring hemizygous *CFAP47* variants. Overall, our findings provide important information for genetic counselors and clinicians to further understand the genetic causes of male infertility and establish a personalized treatment plan.

Data and code availability

The NCBI reference sequence numbers for human *CFAP47* transcript, CFAP47, and mouse *Cfap47* transcript are GenBank: NM_001304548.2, NP_001291477.1, and NM_001368718.2, respectively.

Supplemental Information

Supplemental Information can be found online at https://doi.org/10.1016/j.ajhg.2021.01.002.

Acknowledgments

We would like to thank the families for participating and supporting this study. We also thank the Center of Cryo-electron Microscopy at Zhejiang University for technical support. This work was supported by National Natural Science Foundation of China (31625015, 31521003, 81971447, 81971441, 81871216, 81901541, and 81601340), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), China Postdoctoral Science Foundation (2020TQ0072 and 2019M662786), State Key Laboratory of Reproductive Medicine (SKLRM-K202002), Science and Technology Major Project of Inner Mongolia Autonomous Region of China (zdzx2018065), the key grant of prevention and treatment of birth defect from Hunan Province (2019SK1012), the National Key Research and Developmental Program of China (2018YFC1004900), Shanghai Municipal Commission for Science and Technology (19QA1407500 and 18ZR1432300), Foundation of the Department of Science and Technology of Anhui Province (2017070802D150), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2019PT310002), the National Key Research and Development Program of China (2016YFC1000204), the Agence Nationale pour la Recherche (FLAGEL-OME ANR-19-CE17-0014), and Innovative Research Team of High-level Local Universities in Shanghai (SSMU-ZLCX20180500). The sequencing and analysis of Australian patients was supported in part by a grant from the National Health and Medical Research Council (APP1120356) and an Investigator Award in Science from The Wellcome Trust (209451). We acknowledge the ongoing support of Professor Rob McLachlan and the Monash IVF Group of infertility clinics in Australia for their role in patient recruitment and treatment.

Declaration of interests

Moira K. O'Bryan is a member of the Monash IVF Research and Education Foundation steering committee. The other authors declare no competing interests.

Received: September 30, 2020
Accepted: December 29, 2020
Published: January 19, 2021

Web resources

1000 Genomes Project, https://www.internationalgenome.org/
AnnotSV, https://lbgi.fr/AnnotSV/
CADD, https://cadd.gs.washington.edu/snv
Database of Genomic Variations, http://dgv.tcag.ca/dgv/app/home
gnomAD, https://gnomad.broadinstitute.org
HOPE, https://www3.cmbi.umcn.nl/hope/
HUGO Gene Nomenclature Committee, https://www.genenames.org/
M-CAP, http://bejerano.stanford.edu/MCAP/
National Center for Biotechnology Information (NCBI), https://www.ncbi.nlm.nih.gov/
Online Mendelian Inheritance in Man (OMIM), https://omim.org/
Picard, https://github.com/broadinstitute/picard
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/
SIFT, https://sift.bii.a-star.edu.sg
STRING, https://string-db.org/
SWISS-MODEL, https://swissmodel.expasy.org/
UCSC Genome Browser, http://genome.ucsc.edu
UCSF Chimera, http://www.cgl.ucsf.edu/chimera
UniProt, https://www.uniprot.org

References

1. Matzuk, M.M., and Lamb, D.J. (2002). Genetic dissection of mammalian fertility pathways. Nat. Cell Biol. *4* (*Suppl*), s41–s49.
2. Tournaye, H., Krausz, C., and Oates, R.D. (2017). Novel concepts in the aetiology of male reproductive impairment. Lancet Diabetes Endocrinol. *5*, 544–553.
3. Shahrokhi, S.Z., Salehi, P., Alyasin, A., Taghiyar, S., and Deemeh, M.R. (2020). Asthenozoospermia: Cellular and molecular contributing factors and treatment strategies. Andrologia *52*, e13463.
4. Ben Khelifa, M., Coutton, C., Zouari, R., Karaouzène, T., Rendu, J., Bidart, M., Yassine, S., Pierre, V., Delaroche, J., Hennebicq, S., et al. (2014). Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. Am. J. Hum. Genet. *94*, 95–104.
5. Touré, A., Martinez, G., Kherraf, Z.E., Cazin, C., Beurois, J., Arnoult, C., Ray, P.F., and Coutton, C. (2020). The genetic architecture of morphological abnormalities of the sperm tail. Hum. Genet. https://doi.org/10.1007/s00439-00020-02113-x.
6. Lv, M., Liu, W., Chi, W., Ni, X., Wang, J., Cheng, H., Li, W.Y., Yang, S., Wu, H., Zhang, J., et al. (2020). Homozygous mutations in *DZIP1* can induce asthenoteratospermia with severe MMAF. J. Med. Genet. *57*, 445–453.
7. Liu, C., Miyata, H., Gao, Y., Sha, Y., Tang, S., Xu, Z., Whitfield, M., Patrat, C., Wu, H., Dulioust, E., et al. (2020). Bi-allelic DNAH8 Variants Lead to Multiple Morphological Abnormalities of the Sperm Flagella and Primary Male Infertility. Am. J. Hum. Genet. *107*, 330–341.
8. Martinez, G., Beurois, J., Dacheux, D., Cazin, C., Bidart, M., Kherraf, Z.E., Robinson, D.R., Satre, V., Le Gac, G., Ka, C., et al. (2020). Biallelic variants in *MAATS1* encoding CFAP91, a calmodulin-associated and spoke-associated complex protein, cause severe astheno-teratozoospermia and male infertility. J. Med. Genet. *57*, 708–716.
9. He, X., Liu, C., Yang, X., Lv, M., Ni, X., Li, Q., Cheng, H., Liu, W., Tian, S., Wu, H., et al. (2020). Bi-allelic Loss-of-function Variants in CFAP58 Cause Flagellar Axoneme and Mitochondrial Sheath Defects and Asthenoteratozoospermia in Humans and Mice. Am. J. Hum. Genet. *107*, 514–526.

10. Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature *423*, 825–837.
11. Mueller, J.L., Mahadevaiah, S.K., Park, P.J., Warburton, P.E., Page, D.C., and Turner, J.M.A. (2008). The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat. Genet. *40*, 794–799.
12. Vockel, M., Riera-Escamilla, A., Tüttelmann, F., and Krausz, C. (2019). The X chromosome and male infertility. Hum. Genet. https://doi.org/10.1007/s00439-00019-02101-w.
13. Vogt, P.H., Edelmann, A., Kirsch, S., Henegariu, O., Hirschmann, P., Kiesewetter, F., Köhn, F.M., Schill, W.B., Farah, S., Ramos, C., et al. (1996). Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum. Mol. Genet. *5*, 933–943.
14. Yatsenko, A.N., Georgiadis, A.P., Röpke, A., Berman, A.J., Jaffe, T., Olszewska, M., Westernströer, B., Sanfilippo, J., Kurpisz, M., Rajkovic, A., et al. (2015). X-linked TEX11 mutations, meiotic arrest, and azoospermia in infertile men. N. Engl. J. Med. *372*, 2097–2107.
15. Patat, O., Pagin, A., Siegfried, A., Mitchell, V., Chassaing, N., Faguer, S., Monteil, L., Gaston, V., Bujan, L., Courtade-Saïdi, M., et al. (2016). Truncating Mutations in the Adhesion G Protein-Coupled Receptor G2 Gene ADGRG2 Cause an X-Linked Congenital Bilateral Absence of Vas Deferens. Am. J. Hum. Genet. *99*, 437–442.
16. Liu, W., He, X., Yang, S., Zouari, R., Wang, J., Wu, H., Kherraf, Z.E., Liu, C., Coutton, C., Zhao, R., et al. (2019). Bi-allelic Mutations in TTC21A Induce Asthenoteratospermia in Humans and Mice. Am. J. Hum. Genet. *104*, 738–748.
17. Liu, C., He, X., Liu, W., Yang, S., Wang, L., Li, W., Wu, H., Tang, S., Ni, X., Wang, J., et al. (2019). Bi-allelic Mutations in TTC29 Cause Male Subfertility with Asthenoteratospermia in Humans and Mice. Am. J. Hum. Genet. *105*, 1168–1181.
18. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.
19. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
20. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
21. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
22. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
23. Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–362.
24. Venselaar, H., Te Beek, T.A., Kuipers, R.K., Hekkelman, M.L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics *11*, 548.
25. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
26. The R Development Core Team (2011). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
27. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics *34*, 3572–3574.
28. Tang, S., Wang, X., Li, W., Yang, X., Li, Z., Liu, W., Li, C., Zhu, Z., Wang, L., Wang, J., et al. (2017). Biallelic Mutations in CFAP43 and CFAP44 Cause Male Infertility with Multiple Morphological Abnormalities of the Sperm Flagella. Am. J. Hum. Genet. *100*, 854–864.
29. Wang, L., Li, M.Y., Qu, C., Miao, W.Y., Yin, Q., Liao, J., Cao, H.T., Huang, M., Wang, K., Zuo, E., et al. (2017). CRISPR-Cas9-mediated genome editing in one blastomere of two-cell embryos reveals a novel Tet3 function in regulating neocortical development. Cell Res. *27*, 815–829.
30. Gu, T.P., Guo, F., Yang, H., Wu, H.P., Xu, G.F., Liu, W., Xie, Z.G., Shi, L., He, X., Jin, S.G., et al. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606–610.
31. Li, W., Wu, H., Li, F., Tian, S., Kherraf, Z.E., Zhang, J., Ni, X., Lv, M., Liu, C., Tan, Q., et al. (2020). Biallelic mutations in *CFAP65* cause male infertility with multiple morphological abnormalities of the sperm flagella in humans and mice. J. Med. Genet. *57*, 89–95.
32. Wang, W., Tu, C., Nie, H., Meng, L., Li, Y., Yuan, S., Zhang, Q., Du, J., Wang, J., Gong, F., et al. (2019). Biallelic mutations in *CFAP65* lead to severe asthenoteratospermia due to acrosome hypoplasia and flagellum malformations. J. Med. Genet. *56*, 750–757.
33. Zhang, X., Shen, Y., Wang, X., Yuan, G., Zhang, C., and Yang, Y. (2019). A novel homozygous CFAP65 mutation in humans causes male infertility with multiple morphological abnormalities of the sperm flagella. Clin. Genet. *96*, 541–548.
34. Wang, Y., Yang, J., Jia, Y., Xiong, C., Meng, T., Guan, H., Xia, W., Ding, M., and Yuchi, M. (2014). Variability in the morphologic assessment of human sperm: use of the strict criteria recommended by the World Health Organization in 2010. Fertil. Steril. *101*, 945–949.
35. Cooper, T.G., Noonan, E., von Eckardstein, S., Auger, J., Baker, H.W., Behre, H.M., Haugen, T.B., Kruger, T., Wang, C., Mbizvo, M.T., and Vogelsong, K.M. (2010). World Health Organization reference values for human semen characteristics. Hum. Reprod. Update *16*, 231–245.
36. Auger, J., Jouannet, P., and Eustache, F. (2016). Another look at human sperm morphology. Hum. Reprod. *31*, 10–23.
37. Wambergue, C., Zouari, R., Fourati Ben Mustapha, S., Martinez, G., Devillard, F., Hennebicq, S., Satre, V., Brouillet, S., Halouani, L., Marrakchi, O., et al. (2016). Patients with multiple morphological abnormalities of the sperm flagella due to DNAH1 mutations have a good prognosis following intracytoplasmic sperm injection. Hum. Reprod. *31*, 1164–1172.

38. Baccetti, B., Collodel, G., Estenoz, M., Manca, D., Moretti, E., and Piomboni, P. (2005). Gene deletions in an infertile man with sperm fibrous sheath dysplasia. Hum. Reprod. *20*, 2790–2794.

39. Lee, W.R., Na, H., Lee, S.W., Lim, W.J., Kim, N., Lee, J.E., and Kang, C. (2017). Transcriptomic analysis of mitochondrial TFAM depletion changing cell morphology and proliferation. Sci. Rep. *7*, 17841.

40. Sullivan, R., and Mieusset, R. (2016). The human epididymis: its function in sperm maturation. Hum. Reprod. Update *22*, 574–587.

41. Eddy, E.M., Toshimori, K., and O'Brien, D.A. (2003). Fibrous sheath of mammalian spermatozoa. Microsc. Res. Tech. *61*, 103–115.

42. Chemes, H.E., and Alvarez Sedo, C. (2012). Tales of the tail and sperm head aches: changing concepts on the prognostic significance of sperm pathologies affecting the head, neck and tail. Asian J. Androl. *14*, 14–23.

43. Whitfield, M., Thomas, L., Bequignon, E., Schmitt, A., Stouvenel, L., Montantin, G., Tissier, S., Duquesnoy, P., Copin, B., Chantot, S., et al. (2019). Mutations in DNAH17, Encoding a Sperm-Specific Axonemal Outer Dynein Arm Heavy Chain, Cause Isolated Male Infertility Due to Asthenozoospermia. Am. J. Hum. Genet. *105*, 198–212.

44. Sha, Y.W., Xu, X., Mei, L.B., Li, P., Su, Z.Y., He, X.Q., and Li, L. (2017). A homozygous CEP135 mutation is associated with multiple morphological abnormalities of the sperm flagella (MMAF). Gene *633*, 48–53.

Human Reproduction, pp. 1–23, 2021
doi:10.1093/humrep/deab099

human reproduction **ORIGINAL ARTICLE** ***Reproductive genetics***

Exome sequencing reveals variants in known and novel candidate genes for severe sperm motility disorders

M.S. Oud[1], B.J. Houston[2,3], L. Volozonoka[4,5], F.K. Mastrorosa[5], G.S. Holt[5], B.K.S. Alobaidi[5], P.F. deVries[1], G. Astuti[1], L. Ramos[6], R.I. Mclachlan[7], M.K. O'Bryan[2,3,†], J.A. Veltman[5,*,†], H.E. Chemes[8,†], and H. Sheth[5,9,†]

[1]Department of Human Genetics, Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Center, Nijmegen, The Netherlands [2]School of Biological Sciences, Monash University, Monash, Australia [3]School of BioSciences, Faculty of Science, The University of Melbourne, Parkville, Australia [4]Scientific Laboratory of Molecular Genetics, Riga Stradins University, Riga, Latvia [5]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK [6]Department of Gynaecology and Obstetrics, Radboud University Medical Center, Nijmegen, The Netherlands [7]Hudson Institute of Medical Research, Monash University, Clayton, Melbourne, Australia [8]Centro de Investigaciones Endocrinológicas "Dr. César Bergadá" CEDIE-CONICET-FEI, Hospital de Niños Ricardo Gutiérrez, Buenos Aires, Argentina [9]Foundation for Research in Genetics and Endocrinology, Institute of Human Genetics, Ahmedabad, India

*Correspondence address. Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE1 4EP, UK. E-mail: joris.veltman@newcastle.ac.uk

Submitted on August 14, 2020; resubmitted on March 30, 2021; editorial decision on April 7, 2021

STUDY QUESTION: What are the causative genetic variants in patients with male infertility due to severe sperm motility disorders?

SUMMARY ANSWER: We identified high confidence disease-causing variants in multiple genes previously associated with severe sperm motility disorders in 10 out of 21 patients (48%) and variants in novel candidate genes in seven additional patients (33%).

WHAT IS KNOWN ALREADY: Severe sperm motility disorders are a form of male infertility characterised by immotile sperm often in combination with a spectrum of structural abnormalities of the sperm flagellum that do not affect viability. Currently, depending on the clinical sub-categorisation, up to 50% of causality in patients with severe sperm motility disorders can be explained by pathogenic variants in at least 22 genes.

STUDY DESIGN, SIZE, DURATION: We performed exome sequencing in 21 patients with severe sperm motility disorders from two different clinics.

PARTICIPANTS/MATERIALS, SETTING, METHOD: Two groups of infertile men, one from Argentina ($n = 9$) and one from Australia ($n = 12$), with clinically defined severe sperm motility disorders (motility $<5\%$) and normal morphology values of 0–4%, were included. All patients in the Argentine cohort were diagnosed with DFS-MMAF, based on light and transmission electron microscopy. Sperm ultrastructural information was not available for the Australian cohort. Exome sequencing was performed in all 21 patients and variants with an allele frequency of $<1\%$ in the gnomAD population were prioritised and interpreted.

MAIN RESULTS AND ROLE OF CHANCE: In 10 of 21 patients (48%), we identified pathogenic variants in known sperm assembly genes: *CFAP43* (3 patients); *CFAP44* (2 patients), *CFAP58* (1 patient), *QRICH2* (2 patients), *DNAH1* (1 patient) and *DNAH6* (1 patient). The diagnostic rate did not differ markedly between the Argentinian and the Australian cohort (55% and 42%, respectively). Furthermore, we identified patients with variants in the novel human candidate sperm motility genes: *DNAH12*, *DRC1*, *MDC1*, *PACRG*, *SSPL2C* and *TPTE2*. One patient presented with variants in four candidate genes and it remains unclear which variants were responsible for the severe sperm motility defect in this patient.

LARGE SCALE DATA: N/A

LIMITATIONS, REASONS FOR CAUTION: In this study, we described patients with either a homozygous or two heterozygous candidate pathogenic variants in genes linked to sperm motility disorders. Due to unavailability of parental DNA, we have not assessed the

[†]These authors contributed equally to this work.

frequency of *de novo* or maternally inherited dominant variants and could not determine the parental origin of the mutations to establish in all cases that the mutations are present on both alleles.

WIDER IMPLICATIONS OF THE FINDINGS: Our results confirm the likely causal role of variants in six known genes for sperm motility and we demonstrate that exome sequencing is an effective method to diagnose patients with severe sperm motility disorders (10/21 diagnosed; 48%). Furthermore, our analysis revealed six novel candidate genes for severe sperm motility disorders. Genome-wide sequencing of additional patient cohorts and re-analysis of exome data of currently unsolved cases may reveal additional variants in these novel candidate genes.

STUDY FUNDING/COMPETING INTEREST(S): This project was supported in part by funding from the Australian National Health and Medical Research Council (APP1120356) to M.K.O.B., J.A.V. and R.I.M.L., The Netherlands Organisation for Scientific Research (918-15-667) to J.A.V., the Royal Society and Wolfson Foundation (WM160091) to J.A.V., as well as an Investigator Award in Science from the Wellcome Trust (209451) to J.A.V. and Grants from the National Research Council of Argentina (PIP 0900 and 4584) and ANPCyT (PICT 9591) to H.E.C. and a UUKi Rutherford Fund Fellowship awarded to B.J.H.

Key words: whole exome sequencing / dysplasia of the fibrous sheath / multiple morphological abnormalities of the sperm flagella / asthenozoospermia / male infertility; sperm motility disorders / candidate novel genes

Introduction

The presence of motile sperm is an absolute requirement for male fertility in all mammals. The sperm flagellum is a modified motile cilium and structural and functional deficiencies of this structure are frequently associated with male infertility (Toure et al., 2020). A normal human sperm tail is composed of a central axoneme consisting of nine peripherally arranged doublet microtubules encircling a central pair (Fig. 1). Each peripheral doublet projects towards the next doublet in a clockwise direction, via the presence of outer and inner dynein arms (ODAs and IDAs), which are the key effector structures underpinning sperm motility. The absence of ODAs and/or IDAs of respiratory cilia and sperm flagella in men leads to sperm immotility and chronic respiratory disease in a syndrome collectively known as primary ciliary dyskinesia (PCD, Afzelius et al., 1975; Rebbe and Pedersen, 1975; Rossman et al., 1981). Key to axoneme function and PCD causality are the dynein complexes within the IDA and ODA, which are ATPases responsible for microtubule sliding within the axoneme of the sperm tail and respiratory cilia. Consistent with the assembly of the sperm tail in a distinct cytoplasmic lobe devoid of protein translation, the loss of function of genes associated with protein transport can lead to sperm motility defects in animal models, spanning all aspects of sperm ultrastructure (Pleuger et al., 2020).

Within the broad spectrum of sperm immotility disorders, a range of pathological sub-types exist. In 1987, Chemes *et al.* introduced the term dysplasia of the fibrous sheath (DFS) to describe a distinct form of human sperm pathology involving axonemal and peri-axonemal structures. This condition can be familial, suggesting a genetic aetiology, and/or can be associated with chronic respiratory disease due to dynein deficiency, suggesting a genetic link as well as a mechanistic overlap with PCD (Afzelius et al., 1975; Baccetti et al., 1981; Chemes et al., 1987; Chemes et al., 1990; Chemes et al., 1998; Rawe et al., 2002). In 2014, Ben Khelifa *et al.* introduced the term multiple morphological abnormalities of the sperm flagellum (MMAF) to describe a similar combination of sperm phenotypes (Ben Khelifa et al., 2014). Whilst the phenotypes identified as DFS and MMAF are overlapping, the main differences reside in the significance put on short and thick tails ('stumpy tails') due to the fibrous sheath disorganisation and associated axonemal anomalies (DFS), and on the relevance given to a lack of the central pair of microtubules or dynein arms (MMAF). In the present paper, we will use the term spectrum of severe sperm motility disorders to include both DFS and MMAF but also the more general description of patients with non-syndromic severe asthenoteratozoospermia.

Unbiased next-generation sequencing methods such as exome sequencing have proven critical in the discovery of the genetic causes underlying severe sperm motility disorders, including variants in Dynein Axonemal Heavy Chain 1 and 6 (*DNAH1* and *DNAH6)* (Ben Khelifa et al., 2014; Tu et al., 2019), Cilia And Flagella Associated Protein 43 and 44 (*CFAP43* and *CFAP44)* (Tang et al., 2017) and Glutamine Rich 2 (*QRICH2*) (Shen et al., 2019). Studies in the past 6 years have associated variants in at least 22 genes with severe sperm motility defects and demonstrated that, indeed, a large portion of these defects are genetic in origin (Supplementary Table SI). Currently 30–60% of all DFS-MMAF cases can be explained genetically (Toure et al., 2020). In the present study, we aimed to determine the diagnostic value of the currently known sperm motility genes in different clinical cohorts of patients with severe sperm motility defects using an unbiased exome sequencing approach. This also allowed us to identify novel candidate genes involved in sperm morphology and motility.

Materials and methods

Patients and sample collection

The current study included two groups of patients, one from Argentina and a second from Australia. All patients were informed of the nature of the study and gave informed consent before collection of blood samples. The collection of samples in Argentina was approved by the Ethics Review Board of Centro de Investigaciones Endocrinológicas, National Research Council, Buenos Aires, Argentina. The collection of samples in Australia was approved by the human ethics panels at three sites: Monash Surgical Private Hospital (Clayton), Monash Medical Centre and Monash University, Australia.

Nine males from Argentina were included, who presented with primary infertility due to severe sperm tail defects and very low motility or immotile sperm (Table I). In addition to the standard semen analysis, their sperm were examined by electron microscopy. All patients were characterised as having a typical DFS-MMAF phenotype

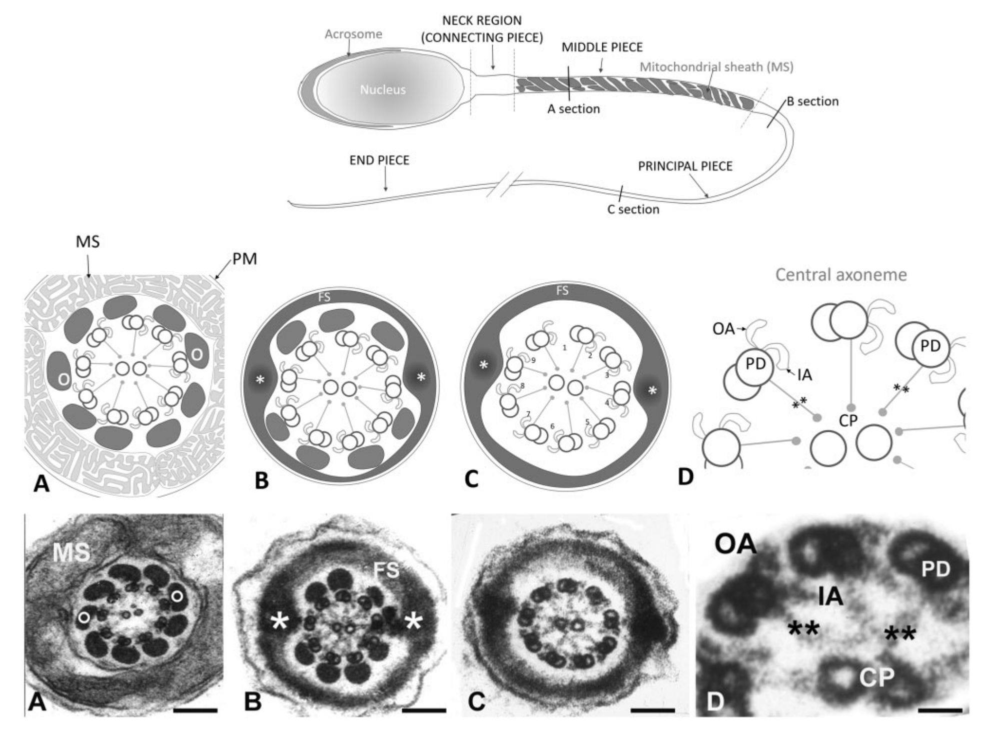


Figure 1. Structure of the tail of a normal human spermatozoon. Upper panel: schematic drawing of a normal human spermatozoon showing three consecutive sections along the length of the tail: middle piece, principal piece and end piece. Transversal lines along its length mark the level of the cross sections displayed in the schematic panel drawings (middle panel) and the electron microscope images (lower panel): (**A**) mid piece, (**B**) proximal principal piece, (**C**) distal principal piece and (**D**) higher magnification detail of the axoneme. Panel A: schematic drawings and sections of the mid piece: circumferential to the axoneme there are nine outer dense fibres (o) each associated to the corresponding peripheral pair. They are surrounded by a helically arranged mitochondrial sheath (MS). At the proximal principal piece (Panel B) mitochondria are replaced by the Fibrous Sheath (FS), which is organised in two longitudinal columns (*) that replace outer dense fibres 3 and 8 and are joined by transverse hemi-circumferential 'ribs' (FS). At the distal principal piece (Panel C) all outer dense fibres disappear and the axoneme is only surrounded by the Fibrous Sheath. Panels D: higher magnification details of three peripheral doublet microtubules (PD) projecting in a clockwise direction toward the next PD OA and IA dynein arm, and radial spokes (**) towards the central pair. Magnification bars: (A) 150 nm, (B) 140 nm, (C) 104 nm, and (D) 26 nm.

(Chemes *et al.*, 1987, 1998). ARG5 has a non-twin brother with DFS, while ARG7 and ARG8 have suffered from chronic respiratory disease and sinusitis since early childhood. ARG6 had a combination of DFS-MMAF with 'acephalic spermatozoa', a phenotype derived from a faulty development of the sperm head-tail attachment (Rawe *et al.*, 2002, Moretti *et al.*, 2011).

In the Australian group, 12 males were recruited following assessment of their semen samples via WHO criteria (World Health Organization, 2010). All presented with infertility due to severe asthenozoospermia and a high percentage of abnormal forms, based on light microscopy as reported by the clinical andrology laboratory (Table II). Specifically, these men had sperm motility values <5% and normal morphology values of 0–4%. Patient AUS3 had a history of chronic sinus congestion with productive cough, suggestive of a ciliary defect. Patient AUS3 also experienced respiratory distress of presumed, but unexplored, environmental origin.

All 21 patients provided a venous blood sample from which DNA was extracted and kept at −80°C until analysis.

Transmission electron microscopy

As indicated above, in addition to the standard semen analyses, an aliquot of fresh semen from each of the Argentinian patients was processed for transmission electron microscopy (TEM) according to the

Table I Light and electron microscopy characteristics of spermatozoa in patients of the Argentinean cohort.

Sample	Shape of tails	Motility: total/ translative	Fibrous sheath thickening	Axoneme**	Dynein arms	Mid piece anomaly	Extension ODFs 3 and 8	Observations
ARG1	Stump*	2/1	Present	8 + 0	Present	Present	Present	Oligozoospermia
ARG2	Stump	0	Present	9 + 0	NE	Present	Present	Oligozoospermia
ARG3	Stump	0	Present	9 + 0 9 + 2	NE	Present	Present	–
ARG4	Stump	NE	Present	NE	NE	NE	NE	Astheno-teratozoospermia.
ARG5	Stump	0	Present	9 + 0	Partial absence	NE	NE	Brother with DFS
ARG6	Stump	0	Present	9 + 0	Present	Present	Present	Combined with acephalic sperm
ARG7	Stump	0	Present	9 + 2	Present	Absent	Absent	Chronic respiratory disease
ARG8	Stump	0	Present	9 + 1 + 1***	Absent	Partial	Partial (?)	Chronic respiratory disease
ARG9	Stump	0	Present	TAD	Absent	NE	NE	Oligozoospermia

*Stump tails: short, thick, of irregular outline.
**Axoneme: 1st digit: number of peripheral doublets, 2nd digit: number of central microtubules.
***9 peripheral doublets + 1 centrally translocated peripheral doublet + 1 central microtubule.
NE, not evaluated because of technical limitations; TAD, Total Axonemal Disruption.
Mid piece anomaly: Short or absent mid pieces due to lack of annulus migration.
Extension ODF 3 and 8: ODF 3 and 8 abnormally extended beyond the mid piece.

Table II Spermiogram and clinical outcome of Australian cohort.

Sample	Concentration (x10^6/ml)*	Motility (% motile)*	Morphology (% abnormal)*	Additional relevant clinical notes	Fertility outcome after ART treatment
AUS1	120	4	95		NA
AUS2	20	1	94		NA
AUS3	5	5	94	Chronic sinus congestion with productive cough	3 ICSI cycles 2 transferred 0 pregnancies
AUS4	29	0	91-95		NA
AUS5	15	2	98		NA
AUS6	12	0	98		NA
AUS7	12	3	96		NA
AUS8	0.5	1	97		2 ICSI cycles 2 pregnancies
AUS9	45	0	100	Presumed smoker's cough. Hematuria	NA
AUS10	0.1	0	98		
AUS11	8.8	5	99		2 ICSI cycles 1 pregnancy
AUS12	1	4	100		10 ICSI cycles 1 foetus loss at 20 weeks 2 pregnancies

*Reference values for normozoospermia according to the World Health Organization: ≥15 × 10^6 sperm per ml; ≥40% motility (progressive motility and non-progressive motility); ≥4% normal forms.

methods previously described (Chemes *et al.*, 1987, 1998). Briefly, within 30–60 min after ejaculation, when liquefaction was complete, samples were diluted in phosphate buffer. After centrifugation pellets were fixed in situ with EM grade glutaraldehyde in phosphate buffer, followed by post-fixation with osmium tetroxide. Sperm pellets were dehydrated followed by infiltration in propylene oxide-epon-araldite mixture, embedded and subsequently polymerised in pure Epon-Araldite (Pelco International, Fresno, CA, USA). Thin sections exhibiting silver to pale golden interference colours were obtained using a Pelco diamond knife in a RMC-7000 ultramicrotome. These sections were mounted on 300 mesh copper grids, double-stained with uranyl acetate and lead citrate, and studied and photographed in a Zeiss 109 electron microscope (Zeiss Oberkochen, Jena, Germany).

Whole exome sequencing

Samples of 100 ng high-quality genomic DNA, measured with Qubit dsDNA HS kit (Invitrogen, Carlsbad, CA, USA), were used for whole exome target capture using Illumina's TruSeq Rapid Exome Capture kit (Illumina, San Diego, CA, USA), according to the manufacturer's protocol. Sample libraries were dual indexed using Illumina's Nextera i7 and i5 primers (Illumina, San Diego, CA, USA). Pooled libraries were sequenced on the NextSeq 500 platform for the Argentinian cohort (Illumina, San Diego, CA, USA) and the NovaSeq 6000 platform for the Australian cohort (Illumina, San Diego, CA, USA). Paired-end sequencing of 150 bp was carried out at an average sequencing depth of 100× per sample. Whole exome sequencing was carried out at the Genomics Core Facility, Biosciences Institute, Faculty of Medical Sciences, Newcastle University, UK.

FASTQ files were aligned against the human reference genome (hg19/GCRh37) using Burrows Wheeler Aligner (BWA MEM 0.7.12) to generate BAM files. Picard toolkit v1.90 was used to mark PCR duplicates and SAMtools v1.6 was used to sort and index BAM files. Genome Analysis Toolkit (GATK) v3.4.46 was used to perform base quality score recalibration and variant calling to generate gVCF file containing SNVs and small indels for each sample. All gVCF files were annotated using Ensembl's Variant Effect Predictor (VEP v92) tool. Homozygosity calling was performed using RareVariantVis (Stokowy *et al*., 2016) and regions of $>$ 1 000 000 bp and a percentage of homozygosity larger than 85 (perc_HMZ $>$85) were classified as stretches of homozygosity.

Variant filtering, prioritisation and validation

For variant filtering and prioritisation, we focused on variants present in exons and canonical splice sites. Variants were excluded from downstream analysis if they did not meet all of the following criteria: (a) variant was more than five reads covering the locus; (b) variant was present in more than 15% of reads covering that locus; and (c) variant had an allele frequency of $<$1% in the gnomAD database (https://gnomad.broadinstitute.org), dbSNP (https://www.ncbi.nlm.nih.gov/SNP/) and our internal database. Variants were classed as homozygous if the variant allele was detected in $>$85% of all reads covering the locus and heterozygous if the variant allele was detected in $>$15% and $<$85% of all the reads covering the locus. Following filtering, variants were prioritised based on the following criteria: (a) variants present in known or candidate severe sperm motility disorder genes (*AK7, AKAP4, ARMC2, CEP135, CFAP43, CFAP44, CFAP58, CFAP65, CFAP69, CFAP70, DNAH1, DNAH2, DNAH6, DNAH17, DZIP1, FSIP2, MAATS1, QRICH2, SPEF2, TTC21A, TTC29* and *WDR66*); (b) genes which were mutated in multiple patients; (c) homozygous variants which were present in homozygosity stretches of $>$1 Mb in length; (d) genes which were reported as having elevated mRNA expression in testis, which is available from the Human Protein Atlas database version 19.1 (https://www.proteinatlas.org/humanproteome/tissue/testis); (e) genes which interact with known sperm motility or cilia related genes in the STRING database version 11.0 (https://string-db.org); and (f) genes which present infertility or astheno-teratozoospermia phenotypes as reported in the Mouse Genomics Institute database (http://www.informatics.jax.org); database last accessed on 8 November 2019 or elsewhere in the literature.

Candidate variants were classified according to the guidelines of the American College of Medical Genetics using five classes: benign (Class 1), likely benign (Class 2), variant of unknown significance (Class 3), likely pathogenic (Class 4) and pathogenic (Class 5) (Richards *et al*., 2015) using the software program Alamut Visual version v.2.13. Missense pathogenicity prediction was performed by Align GVGD, SIFT, MutationTaster and PolyPhen-2 and splicing prediction was performed as described previously (Houston *et al*., 2020). Variants on chromosome X were classified as (likely) benign if the allele frequency in men exceeded 0.05% in any population described in gnomAD. Candidate variants following filtering and prioritisation were visually inspected in the IGV browser (http://software.broadinstitute.org/software/igv/) to evaluate variant quality. Lastly, candidate variants were validated using the conventional Sanger sequencing approach according to the standard protocols.

Control cohort of proven fathers

To assess the frequency of all variants prioritised in our analysis, we used a control cohort of 5784 proven fathers as described previously (Wyrwoll *et al*., 2019). Detailed information regarding child conception was unavailable for these men, but they likely reflect the normal population of fathers in the Netherlands. Currently, approximately 1 in 33 children in the Netherlands is conceived through any form of IVF, ICSI or transfer of previously frozen embryos, and 1 in 98 is conceived through ICSI alone as reported by the Dutch Society for Obstetrics and Gynecology (https://www.degynaecoloog.nl/nuttige-informatie/ivf-resultaten/).

CNV analysis

CNVs were detected from the exome sequencing data using a custom GATK4-based pipeline. This workflow exploits the GATK4 sequence read-depth normalisation (McKenna *et al*., 2010) and a custom R-based segmentation and visualisation. The CNVs identified were annotated with AnnotSV3 (https://lbgi.fr/AnnotSV/). Due to low quality of the CNV data, samples from ARG5, ARG3 and AUS9 were excluded from the analysis. The common CNVs identified in more than 1% of the samples of the Database of Genomic Variations were excluded. For the rare and large CNVs encompassing $\geq$ 20 sequencing probes, the Log2Ratio plots were manually inspected and the genes involved were investigated to find any linked to spermatogenesis and testis function. A panel of known primary ciliary dyskinesia genes comprehensive of 32 (described in Takeuchi *et al*., 2020) as well as the known or candidate severe sperm motility disorder genes reported in Supplementary Table SI were used to screen the genes involved in all of the CNVs detected.

Results

Sperm phenotype under light and electron microscopy

Sperm from all men in the Argentinian cohort exhibited the DFS-MMAF phenotype, as verified at a light and electron microscopic level (Table I and Fig. 2). The main features of this phenotype include severe astheno-teratozoospermia (<5% motility) or total immotility (Table I). Most spermatozoa present with short, thick and irregular tails ('stump tails', Fig. 2A, C and D). There are occasional sperm heads with absent flagella. Ultrastructural examination shows serious architectural disruptions. Thick and short stump tails are packed by disorganised thick filaments corresponding to the ribs of the fibrous sheath and the axonemes depict serious distortions such as partial to complete lack of the central pair (9 + 0 configuration, Fig. 2E and F). Dynein arms (inner or both) are frequently absent from the axoneme peripheral doublets (Fig. 2G, I and J) and, on occasions, the axoneme is completely disrupted (Fig 2I and J). Outer dense fibres 3 and 8 are abnormally extended to the sperm tail principal piece (Table I and Fig. 2E and G). As a consequence of failed caudal migration of the annulus, mitochondria do not assemble properly and the mid piece is missing or substantially reduced to very few mitochondria (Fig. 2A, C and D). Semen samples from the Australian cohort were examined following the WHO 2010 criteria for semen analysis (World Health Organization, 2010) and were characterised as severe astheno-teratozoospermia (Table II).

Exome sequencing in patients with severe sperm motility disorders

Exome sequencing revealed an average of 92 504 variants per patient (Supplementary Table SII). Since severe sperm motility disorders typically follow an autosomal recessive inheritance pattern, we focussed our analysis on compound heterozygous and homozygous variants, supplemented with an analysis of X and Y-linked variants. After exclusion of false-positive variant calls and variants classified as (likely) benign according to the ACMG guidelines, we identified an average of five variants in each patient for further consideration (Supplementary Tables SII and SIII). Parental data were not available to confirm compound heterozygosity of the heterozygous variants. CNV analysis was performed in all patient exome data, but no clinically relevant CNVs were detected.

In 10 out of 21 patients (47.6%), we found homozygous or 2 heterozygous high confidence disease causing variants in genes previously associated with severe sperm motility disorders (Table III and Supplementary Table SIII): *CFAP43* (3 patients: ARG2, AUS8 and AUS9); *CFAP44* (2 patients: ARG6 and ARG9), *CFAP58* (1 patient: ARG5), *QRICH2* (2 patients: AUS5 and AUS12), *DNAH1* (1 patient: AUS2) and *DNAH6* (1 patient: ARG3). The homozygous variants found in ARG5, AUS8, AUS9 and AUS12 were each located in a region of homozygosity indicating consanguinity (Supplementary Table SIV). None of the variants were found to be present as homozygous in a control cohort of 5784 proven fathers (Supplementary Tables SI and SV).

Novel candidate genes for severe sperm motility disorders

Expanding the analysis to consider putative variants in genes not previously associated with human astheno-teratozoospermia, revealed a total of 71 variants in 53 genes in the remaining patients (Supplementary Tables SII and SIII). After assessing the predicted pathogenicity of the variant, gene expression pattern in the testis, protein–protein interactions, relevant animal models and previous publications found in PubMed, we classified an additional 11 genes in seven patients as novel or possible candidate gene for a severe motility disorder. All other variants were classified as unlikely to be disease causing (Supplementary Table SIII).

From the Argentinian cohort, ARG1, a patient with typical DFS-MMAF features and no reported symptoms of PCD, carried two heterozygous variants in Dynein Axonemal Heavy Chain 12 (*DNAH12*) (c.5393T>C; p.(Phe1798Ser) and c.7438C>T; p.(Pro2480Ser)) (Table III). *DNAH12* expression is restricted to the ciliated cells in the brain, fallopian tube, lung and testis (Dumur *et al.*, 1990). The variant allele frequency of these two variants is very similar in control populations, indicating that they are present on the same allele and may thus not be compound heterozygous. It remains unclear whether variants in *DNAH12* cause DFS-MMAF in this patient. ARG4, carried a homozygous nonsense variant (c.369T>A; p.(Tyr123*)) in Parkin Co-regulated (*PACRG*) (Table III), which has not been described before in public databases such as gnomAD. The variant likely results in nonsense-mediated decay of *PACRG* mRNA. Lastly, in patient ARG7, a man with DFS-MMAF in combination with chronic respiratory disease, we identified two heterozygous nonsense variants (c.238C>T; p.(Arg80*)) in exon 2 and (c.352C>T; p.(Gln118*)) in exon 3 in Dynein Regulatory Complex Subunit 1 (*DRC1*) (Table III). DFS-MMAF patient ARG8 carried variants in multiple candidate genes previously associated with ciliated cell development: *DNAH6*, *ATP2B4*, *CEP350* and *CEP290*.

The Australian patient AUS4 carried a homozygous missense variant (c.634C>T; p.(Arg212Trp)) in Signal Peptide Peptidase Like 2C (*SPPL2C*) (Table III). Patient AUS7 carried a homozygous nonsense variant (c.715C>T; p.(Gln239*)) in Transmembrane Phosphoinositide 3-Phosphatase and Tensin Homolog 2 (*TPTE2*), which has a highly testis enriched expression pattern. Finally, we identified two heterozygous nonsense variants (c.472C>T; p.(Gln158*) in Exon 3 and c.2134C>T; p.(Gln712*) in Exon 7) in Mediator of DNA Damage Checkpoint 1 (*MDC1*) in patient AUS11, who suffered from mild oligozoospermia combined with astheno-teratozoospermia (Table III). In humans, *MDC1* is detected in all tissues but is most strongly expressed in the testis (Uhlen *et al.*, 2010; GTEx Consortium, 2015).

Analysis of homozygous loss-of-function variants in proven fathers

In our analysis of 21 patients, we identified four patients with a homozygous loss-of-function (LoF) variant in a gene known to be required for normal sperm tail assembly and function. Given the large number of genes involved in sperm tail assembly, we assessed whether sequencing a control cohort of 5784 proven fathers would result in a similar number of homozygous LoF variants. In all 22 known sperm tail assembly genes as well as the 6 new candidate genes, one homozygous LoF carrier was identified among the 5784 proven fathers

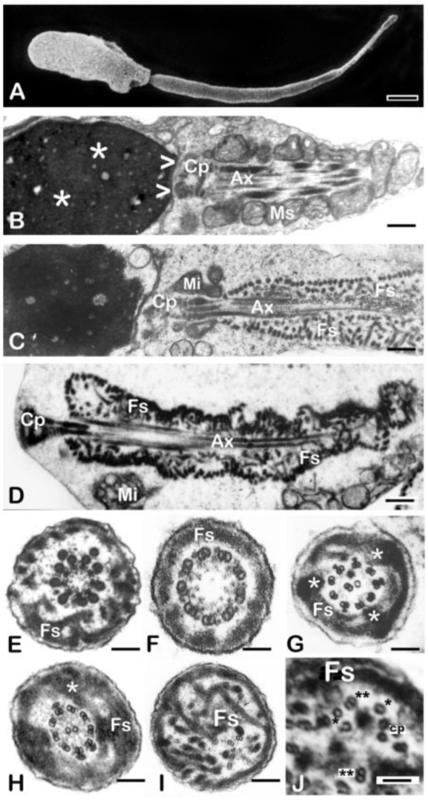


Figure 2. Characterisation of the DFS-MMAF phenotype by scanning and transmission electron microscopy. (**A**) Scanning EM of a typical DFS-MMAF spermatozoon (ARG1). The head is abnormally shaped and the mid piece is absent. The sperm tail is very short and thick (7.7 µm long and 630 nm in diameter (normal values of 50 µm in length and 100–140 nm in diameter). (**B**) Longitudinal section of the head, connecting piece and mid piece of a normal human spermatozoon. The sperm head shows densely condensed chromatin (asterisks). At its caudal pole there is a shallow concavity (the implantation fossa, arrowheads) where the connecting piece (CP) and centrioles of the sperm tail are lodged. A helically arranged mitochondrial sheath (MS) surrounds the first part of the axoneme (Ax) with its central microtubules and peripheral outer dense fibers (see mid piece cross section in Fig. 1). (**C**) Longitudinal section of a typical DFS-MMAF spermatozoon to illustrate the details of the phenotype (patient not included in the genetics part of this study). A largely missing mitochondrial sheath is replaced by few mitochondria (Mi). The axoneme (Ax) and outer dense fibers are surrounded by a thick, multilayered, haphazardly arranged fibrous sheath (FS). (**D**) Longitudinal section of a DFS-MMAF/acephalic spermatozoon (ARG6). The connecting piece (CP) takes up the cranial position, there is no mid piece and a misplaced mitochondrion lays besides the flagellum (Mi). A disorganised, multi-layered fibrous sheath (FS) encloses the centrally located axoneme (Ax) and surrounding outer dense fibres. (**E–J**) Sperm tail transverse sections of DFS-MMAF spermatozoa showing redundant and disorganised fibrous sheaths (FS). In Panel E (ARG6), the central pair is missing, there are two extra outer dense fibres and the FS is thickened and disorganised. Dynein arms are present. Panel F (ARG5) corresponds to a distal section of the tail principal piece. The axoneme is 9 + 0 (lack of central pair). Dynein arms are present. The FS is not redundant (distal part of the flagellum). In Panel G (ARG8), one microtubule is missing from the central pair and there is central translocation of one supernumerary peripheral doublet. Dynein arms are absent, a superfluous and disorganised FS shows three lateral columns (asterisks). In Panel H (ARG7), the axoneme is normal, and a very thick FS presents with one extra lateral column (asterisk). Panel I (ARG9) shows complete axonemal disruption, a hyperplastic FS and two microtubular pairs lacking dynein arms (down and right from the centre). Panel J (ARG9) shows a higher magnification detail of a DFS tail with disrupted axoneme and hyperplastic FS (Fs). Note a dislocated central pair (cp) and various singlet (*) and doublet microtubules (**) with absent dynein arms. Magnification bars: (A) 1.14 µm, (B) 741 nm, (C) 1012 nm, (D) 533 nm, € 272 nm, (F) 92 nm, (G) 270 nm, (H) 274 nm, (I) 190 nm, and (J) 75 nm.

(Supplementary Tables SI and SV). This variant was (NM_001039706.2: c.992del; p.(Gly331Alafs*6)) in Cilia And Flagella Associated Protein 69 (*CFAP69*).

Discussion

With the recent application of exome sequencing to previously unexplained individuals with severe sperm tail assembly disorders, variants in at least 22 genes have now been implicated in the spectrum of motility disorders due to tail abnormalities (Supplementary Table SI). In this study, we set out to find the causative genetic variants in known and novel candidate genes in 21 men suffering from severe asthenoteratozoospermia. In 10 out of 21 patients (47.6%), we identified pathogenic or likely pathogenic mutations in a total of 6 known severe sperm motility disorder genes, *CFAP43* (n = 3), *CFAP44* (n = 2), *CFAP58* (n = 1), *QRICH2* (n = 2), *DNAH1* (n = 1) and *DNAH6* (n = 1). In addition, we identified predicted pathogenic mutations in novel candidate genes in seven other patients (33%).

Exome sequencing is an effective method to identify genetic causes of severe motility disorders

With the use of exome sequencing, we demonstrated that variants in known sperm motility genes likely explained the disorder in 10/21 individuals, reaching a diagnostic yield of approximately 48%. This result is in concordance with previous estimates for the severe sperm motility disorder DFS-MMAF (Coutton *et al.*, 2019; Shen *et al.*, 2019). Interestingly, three patients in our cohort carried variants in the

Table III Selected variants prioritised from the exome sequencing data in severe sperm motility disorders.

Patient	Gene	cDNA*	Protein	Zygosity	GnomAD variant frequency (population with highest frequency)	Variant classification (ACMG)**	Gene expression enriched in testis***	Disease model described	Additional information (see also Supplementary Table SI)	Conclusion

ARG1	*DNAH12*	c.5393T>C
		c.7438C>T
			p.(Phe1798Ser)
			p.(Pro2480Ser)
				Het
				Het
					0.1% (AFR: 0.4%)
					0.0% (AFR: 0.2%)
						VUS
						VUS
							Yes	No	Variants have highly similar allele frequencies suggesting they reside on the same allele	Unclear if disease causing

ARG2	*CFAP43*	c.1442 + 1G>A	p.?	Het	0.0% (NFE: 0.0%)	Likely pathogenic	Yes	Yes, mouse (Tang *et al.*, 2017)	c.1442 + 1G>A is present in trans with c.1040T>C	Probably disease causing
		c.1019T>C	p.(Phe340Ser)	Het (in cis with c.1040T>C)	0.01% (NFE: 0.03%)	Unlikely pathogenic				
		c.1040T>C	p.(Val347Ala)	Het (in cis with c.1019T>C)	0.01% (NFE: 0.02%)	VUS			Known gene c.1040T>C previously reported (Coutton *et al.*, 2018)	
ARG3	*DNAH6*	c.1316 + 1_1316 + 2insC c.7762C>T	N/A p.(Arg2588*)	Het Het	0.0% (NFE: 0.0%) 0.0% (AMR: 0.0%)	Likely pathogenic Pathogenic	Yes	Yes, mouse and zebrafish (Li *et al.*, 2016)	Known gene	Probably disease causing
ARG4	*PACRG*	c.369T>A	p.(Tyr123*)	Hom	0.00%	Likely pathogenic	Yes	Yes, mouse (Lorenzetti *et al.*, 2004)	Associated with the development of sperm flagellum (Lorenzetti *et al.*, 2004; Li *et al.* 2015)	Novel candidate gene
ARG5	*CFAP58*	c.1360C>T	p.(Gln454*)	Hom	0.05% (ASJ: 1.09%)	Pathogenic	Yes	No	In homozygosity region. Variant is relatively common in Ashkenazi Jewish population	Probably disease causing
ARG6	*CFAP44*	c.652del	p.(Arg218Aspfs*37)	Hom	0.03% (ASJ: 0.61%)	Pathogenic	Yes	Yes, mouse (Tang *et al.*, 2017)	Known gene	Disease causing

ARG7	*DRC1*	c.238C>T
		c.352C>T
			p.(Arg80*)
			p.(Gln118*)
				Het
				Het
					0.00% (FIN: 0.03%)
					0.04% (0.07% (NFE)
						Pathogenic
						Pathogenic
							Yes	Yes, Chlamydomonas reinhardtii (Wirschell *et al.*, 2013)	Described in primary ciliary dyskinesia (Wirschell *et al.*, 2013; Morimoto *et al.*, 2019a)	Novel candidate gene

ARG8	*DNAH6*	c.2059C>A	p.(Pro687Thr)	Hom	0.04% (NFE: 0.07%)	VUS	Yes	Yes, mouse and zebrafish (Li *et al.*, 2016)	Known gene	Possible candidate gene
	ATP2B4	c.376G>C	p.(Gly126Arg)	Hom	0.01% (SAS: 0.08%)	VUS	No	Yes, mouse (Schuh *et al.*, 2004)	Mouse displays asthenozoospermia. In homozygosity region	Possible candidate gene
	CEP350	c.229A>G	p.(Arg77Gly)	Hom	0.34% (ASJ: 0.81%)	VUS	No	No	In homozygosity region	Possible candidate gene

	CEP290	c.5998A>G
		c.1092T>G
			p.(Ile2000Val)
			p.(Ile364Met)
				Het
				Het
					0.02% (NFE: 0.04%)
					0.08% (SAS: 0.35%)
						VUS
						VUS
							No	Yes, mouse (Lancaster *et al.*, 2011)	Described in patients with Leber's Congenital Amaurosis and asthenozoospermia (Yzer *et al.*, 2012)	Possible candidate gene

ARG9	*CFAP44*	c.2674A>G
		c.2107A>G
		c.2104A>T
		c.1174T>C
			p.(Met892Val)
			p.(Arg703Gly)
			p.(Ile702Leu)
			p.(Trp392Arg)
				Het
				Het
				Het
				Het
					0.05% (AMR: 0.08%)
					0.00%
					0.00%
					0.00%
						Likely benign
						VUS
						Likely benign
						VUS
							Yes	Yes, mouse (Tang *et al.*, 2017)	Known gene	Probably disease causing

AUS1	-	–	–	–	–	–	–	–	–	

(continued)

Table III Continued

Patient	Gene	cDNA*	Protein	Zygosity	GnomAD variant frequency (population with highest frequency)	Variant classification (ACMG)**	Gene expression enriched in testis***	Disease model described	Additional information (see also Supplementary Table SI)	Conclusion
										No candidate genes found
AUS2	*DNAH1*	c.5105G>A c.10823 + 1G>C	p.(Arg1702Gln) p.?	Het Het	0.00% (NFE: 0.00%) 0.00%	VUS Likely pathogenic	No	Yes, mouse (Neesen *et al.*, 2001)	Known gene	Probably disease causing
AUS3	-	–	–	–	–	–	–	–	–	No candidate genes found
AUS4	*SPPL2C*	c.634C>T	p.(Arg212Trp)	Hom	0.01% (SAS: 0.06)	VUS	Yes	Yes, mouse (Niemeyer *et al.*, 2019)	SPPL2c deficiency leads to a partial loss of elongated spermatids and reduced motility of mature spermatozoa, but preserved fertility in mice (Niemeyer *et al.*, 2019). Possibly involved in acrosome formation (Papadopoulou *et al.*, 2019)	Novel candidate gene
AUS5	*QRICH2*	c.145dup	p.(Thr49Asnfs*31)	Hom	0.00% (NFE: 0.00%)	Pathogenic	Yes	Yes, mouse (Shen *et al.*, 2019)	Known gene	Disease causing
AUS6	-	–	–	–	–	–	–	–	–	No candidate genes found
AUS7	*TPTE2*	c.715C>T	p.(Gln239*)	Hom	0.00% (NFE: 0.19%)	Likely pathogenic	Yes	No	Voltage-sensitive phosphatase (Halaszovich *et al.*, 2012)	Novel candidate gene
AUS8	*CFAP43*	c.335A>T	p.(Asp112Val)	Hom	0.01% (NFE : 0.01%)	VUS	Yes	Yes, mouse (Tang *et al.*, 2017)	Known gene In homozygosity region	Probably disease causing
AUS9	*CFAP43*	c.944del	p.(Gly315Alafs*22)	Hom	0.00%	Pathogenic	Yes	Yes, mouse (Tang *et al.*, 2017)	Known gene. In homozygosity region	Disease causing
AUS10	-	–	–	–	–	–	–	–	–	No candidate genes found

AUS11	*MDC1*	c.472C>T
		c.2134C>T
			p.(Gln158*)
			p.(Gln712*)
				Het
				Het
					0.00%
					0.00%
						Likely pathogenic
						Likely pathogenic
							Yes	Yes, mouse (Lou *et al.*, 2006)	Mouse knock-out possibly has a meiotic defect (Lou *et al.*, 2006)	Novel candidate gene

AUS12	*QRICH2*	c.169G>A	p.(Glu57Lys)	Hom	0.01% (NFE: 0.02)	VUS	Yes	Yes, mouse (Shen *et al.*, 2019)	Known gene. In homozygosity region	Probably disease causing

*gDNA position and transcript information are available in Supplementary Table SIII.
**VUS: Variant of Unknown Significance.
***Based on the Human Protein Atlas version 19.1.
The full table is available in Supplementary Table SIII.

recently discovered genes *CFAP58* (He et al., 2020), *QRICH2* (Shen et al., 2019) and *DNAH6* (Tu et al., 2019), further supporting their role in causing DFS-MMAF. Exome sequencing is therefore a highly efficient method for genetic diagnostics in patients with defective sperm motility disorders. Of note, while the two cohorts included in this study were collected and phenotyped by different clinicians using different levels of resolution (with/without electron microscopy), genetic diagnoses were observed in both cohorts in comparable numbers, with five out of nine Argentinian patients genetically diagnosed versus 5 out of 12 Australian patients. Interestingly, however, exome sequencing revealed variants in either known or candidate genes in all Argentinean patients but not in all Australian patients, leaving four Australian patients without mutations in known or novel candidate genes. The reason for this difference is not currently known but is likely related to the differences in the population background between the two locations.

Patient ARG3 carried two likely pathogenic variants (c.1316+1_1316+2ins (canonical splice site variant) and c.7762C>T; p.(Arg2588*)) in *DNAH6* and no variants in other candidate genes (Table III). *DNAH6* is a dynein gene involved in motile cilia function in numerous tissues, which, when mutated, leads to a primary cilia dyskinesia phenotype in zebra fish and humans (Li et al., 2016). DNAH6 has also been recently implicated in the aetiology of human DFS-MMAF and confirmed in a mouse study (Tu et al., 2019). Herein we confirm that mutations in *DNAH6* are a bona fide cause of human DFS-MMAF and that DFS-MMAF should be considered as part of the spectrum of clinical presentations designated as severe sperm motility syndrome.

In two other cases, we are not convinced that the sperm motility defects can be explained by variants in *DNAH6*. The two *DNAH6* variants in AUS3 (c.9436A>G; p.(Ser3146Gly) and c.12352G>A; p.(Ala4118Thr)) (Supplementary Table SIII) have almost identical minor allele frequencies among the gnomAD populations, indicating they are located on the same allele and are not compound heterozygous. This hypothesis could not be tested due to the unavailability of parental DNA of this patient, but this makes it unlikely that these variants alone cause DFS-MMAF in AUS3. The other patient with a candidate variant in *DNAH6* is ARG8, carrying a homozygous variant of unknown significance (VUS) (c.2059C>A; p.(Pro687Thr)). This patient, however, also carried variants of unclear significance in three other genes: (1) *ATP2B4,* which is associated with asthenozoospermia in mice; (2) *CEP290*, mutations which are a known cause of Leber's Congenital Amaurosis that is associated with asthenozoospermia in males; and (3) *CEP350*, which is known to interact with *CEP290* and the known MMAF gene *CEP135*. It remains uncertain whether variants in any of these genes alone, or combined, are responsible for DFS-MMAF in combination with chronic bronchitis.

The variant (p.(Gln454*)) identified in *CFAP58* in patient ARG5 has a low allele frequency in gnomAD (0.054%), but appears to be more common among the Ashkenazi Jew (ASJ) population reported in the same database (1.086%). This means that an expected 0.012% of this population is homozygous for this variant; a number that is slightly higher than the estimated frequency of DFS-MMAF in the population of Dutch men (0.005–0.01%) (our own observations). Although the variant identified in ARG5 almost certainly disrupts CFAP58 protein function, it remains unclear if this variant is underlying the DFS-MMAF phenotype in the patient due to the allele frequency, which is higher than expected in the ASJ population and additional population studies are required. This variant was located in a homozygosity stretch, indicating consanguinity. The brother of this patient also presented with DFS-MMAF, but DNA was unavailable.

Interestingly, the semen analysis of ARG6, who carried a homozygous frameshift variant (p.(Arg218Aspfs*37)) in *CFAP44,* showed the combination of DFS-MMAF and acephalic sperm previously reported in the literature (Rawe et al., 2002; Moretti et al., 2011). This indicates the possibility of combinations between different sperm phenotypes of genetic origin or involvement of a single gene/protein associated with transport pathways common between the sperm head-tail coupling apparatus and tail proteins (reviewed by Pleuger et al., 2020).

Novel candidate genes for severe sperm motility disorders

Since variants in known genes explain causality in approximately half of our patients, we investigated whether genetic variants were present in genes with a potential role in sperm function. In the current study, we observed missense and null mutations in six novel genes (*DNAH12*, *PACRG*, *DRC1*, *MDC1*, *SSPL2C* and *TPTE2*) that have previously been identified to play a role in axoneme assembly and/or sperm flagellum development and have been shown to interact with genes already implicated in sperm function (Pleuger et al., 2020; Toure et al., 2020).

Patient ARG1 carried two missense variants (p.(Phe1298Ser) and p.(Pro2480Ser)) in *DNAH12,* which is the closest paralog of the *DNAH1* gene. Pathogenic variants in *DNAH1* are known to cause classical PCD or DFS-MMAF without any PCD symptoms (Ben Khelifa et al., 2014; Wambergue et al., 2016; Wang et al., 2017; Amiri-Yekta et al., 2016; Sha et al., 2017; Coutton et al., 2018; Sha et al., 2019b; Coutton et al., 2019; Li et al., 2019b; Hu et al., 2019). The allele frequencies of both *DNAH12* variants are similar in three different populations in gnomAD, which could indicate they reside on the same allele. It is therefore unclear whether these variants are indeed bi-allelic and causal of infertility. Another patient, ARG4, carried a homozygous nonsense variant (p.(Tyr123*)) in *PACRG.* This gene has been implicated in motile cilia function and mutations in mice are known to cause male infertility characterised by defective sperm head and tail formation in combination with hydrocephalus next to fertility problems (Lorenzetti et al., 2004; Wilson et al., 2010; Li et al., 2015). The variants identified in ARG1 and ARG4 likely explain the DFS-MMAF phenotype seen in both patients.

Patient ARG7 carried two nonsense variants (p.(Arg80*) and p.(Gln118*)) in *DRC1*. This gene is known to be important for motile cilia formation, and specifically outer dynein arm formation, as concluded from studies in algae (Wirschell et al., 2013). Loss-of-function point mutations and a recurrent ~28 kb deletion encompassing *DRC1* Exons 1–4 have previously been described in patients with PCD, including a man who had undergone fertility treatment (Wirschell et al., 2013; Morimoto et al., 2019a). Unfortunately, sperm ultrastructure was not examined. The second nonsense variant (p.(Gln118*)) has been described in two Swedish PCD families (Carlen et al., 2003; Wirschell et al., 2013). The importance of *DRC1* in human spermatogenesis is further strengthened by the observed enhanced expression in the testis (along with the brain and fallopian tube) (Uhlen et al., 2010; GTEx Consortium, 2015). Collectively, these data suggest that mutations in *DRC1* cause a spectrum of clinical presentations involving defects in motile cilia function, and that variants in *DRC1* are a novel high confidence cause of male infertility.

In addition to variants in genes with a strong link the clinical aetiology of DFS-MMAF, we also identified variants in genes with less well characterised links to sperm motility. First, patient AUS4 carried a homozygous missense variant of unknown significance (p.(Arg212Trp)) in *SPPL2C*. This gene encodes a testis-specific intermembrane protease residing in the endoplasmic reticulum in somatic cells and in elongating spermatids in the testis (Niemeyer *et al.*, 2019; Papadopoulou *et al.*, 2019). In mice, *Sppl2c* deficiency leads to hypospermatogenesis starting at the level of spermatids, as well as reduced sperm motility and male sub-fertility (Niemeyer *et al.*, 2019). The effect of *Sppl2c* deletion on the sperm ultra-structure was not examined and, as such, a definitive link to DFS-MMAF cannot be made. *SPPL2C* is also one of multiple genes deleted in Koolen–de Vries 17q21.31 microdeletion syndrome (Koolen *et al.*, 2006; Shaw-Smith *et al.*, 2006). A recent case report of Koolen–de Vries syndrome described a patient with intellectual disability and oligoastheno-teratozoospermia. Although it is unlikely that disruption of *SPPL2C* has an effect on intellectual disability, it is possible that its disruption is causative for the infertility described in this patient.

Patient AUS7 carried a homozygous nonsense variant (c.715C>T; p.(Gln239*)) in the voltage-sensitive and membrane-associated phosphatase *TPTE2*. The expression of this gene is highly testis enriched and the protein is localised within the sperm plasma membrane, where it is likely involved in integrating environmental cues into changes in sperm function (Sutton *et al.*, 2012). The homozygous nonsense variant is located in exon 8 and likely results in nonsense-mediated mRNA decay. Based on these data, while we predict mutations in *TPTE2* are likely to result in male infertility, a specific link to the mechanisms underpinning sperm tail assembly is lacking. As such, we classify mutations in *TPTE2* as a possible, but not high confidence cause of severe sperm motility disorders.

Lastly, in patient AUS11, we identified two heterozygous nonsense variants in *MDC1*, a gene essential for the silencing of sex chromosome and genome stability during male meiosis in mice (Ichijima *et al.*, 2011). Unfortunately, parental DNA was not available to prove the biallelic presence of these two heterozygous variants. The knockout mouse model of *Mdc1* revealed a meiotic arrest (Lou *et al.*, 2006), a phenotype that does not directly match the phenotype seen in AUS11. The semen analysis of AUS11 revealed moderate oligozoospermia (8.8 million sperm/ml) and only 5% of sperm showed progressive motility. Based on currently available information on *MDC1*, while we are confident mutations in *MDC1* can lead to human male infertility, they are not a high confidence cause of the severe motility disorder seen in this patient.

Comparison of our results to exome sequencing in a control cohort

The vast majority of known variants causing sperm tail assembly disorders, are homozygous LoF variants (Supplementary Table SI; Oud *et al.*, 2019). Sequencing of a control cohort of 5784 proven fathers did not reveal a similarly high number of homozygous LoF variants in sperm motility genes (Supplementary Table SV). This shows that these disruptive variants occur only rarely in the normal male population, in contrast to males presenting with severe sperm motility disorders. These results further strengthen the evidence for the involvement of these genes in abnormal sperm tail assembly.

Interestingly, we did identify one homozygous LoF variant in a proven father in *CFAP69*, a recently discovered gene which is strongly associated with the DFS-MMAF phenotype (Dong *et al.*, 2018; He *et al.*, 2019). Although it is possible that this man had ICSI to conceive his child, it does indicate that homozygous knock-out of this gene may not always cause complete sterility in human. Due to data usage restrictions, we are unable to search for compound heterozygous variants and could only investigate the zygosity and frequency of all variants in the entire cohort.

Importance for genetic testing in severe sperm motility disorders

Both the European and American guidelines for genetic testing in male infertility, provide a stratified approach to select azoospermic and oligozoospermic patients based on clinical phenotypes to certain genetic tests (Esteves and Chan, 2015; Jungwirth, 2018). The guidelines, however, do not include recommendations for patients with other sperm phenotypes including severe sperm motility disorders. Without a genetic diagnosis, a clinician is very restricted to accurately counsel couples with questions about the causes of their infertility, possible co-morbidities, the potential success of ART treatment and the (reproductive) health of their offspring. Hence, understanding of and testing for genetic causes of severe sperm motility disorders are of enormous value to patients and clinicians. Currently, it remains unclear if genetic abnormalities underlying sperm motility disorders affect the health of potential offspring. As such, the field should consider expanding diagnostic genetic testing for this group of patients, especially since this and other recent studies have reported high (>40%) diagnostic yields in these patient groups (Toure *et al.*, 2020). Furthermore, systematic linking of genetic data with ART success rates as well as patient and offspring health is pivotal for improved counselling in this group of patients.

Conclusion

In summary, our genetic data provided a diagnosis for 10 out of 21 patients with severe sperm motility disorders and we discovered novel candidate genes in seven other patients. Functional data based on literature, propose variants in *DNAH12, DRC1, MDC1, PACRG, SPPL2C* and *TPTE2* as novel genetic causes of severe sperm motility disorders. Our results demonstrate that exome-wide screening for pathogenic variants in these genes is an effective way to diagnose severe forms of motility disorders (Supplementary Table SI). Exome sequencing of additional cases and re-analysis of exome data of currently unsolved cases from other cohorts may reveal additional causative mutations in these novel candidate genes.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Data availability

Raw and processed data are available under controlled access and requires a Data Transfer Agreement from the European Genome-

Phenome Archive (EGA) repository: EGAS00001005018. Data is available using the following link: https://ega-archive.org/studies/EGAS00001005018

Acknowledgements

The authors are grateful for the participation of all patients in this study. The authors would like to thank Lisenka Vissers and Christian Gilissen for technical support in this study. The authors thank Monash IVF for the ongoing access to patients.

Authors' roles

R.M., H.C. and M.O.B. provided the clinical samples and data. Electron microscopy was performed by H.C. M.S.O., B.K.S.A., P.d.V. and L.V. performed the sequencing. M.S.O., L.V., G.A. and F.K.M. performed data analysis and interpretation of results under the supervision of J.A.V. and H.S. B.K.S.A. and P.d.V. performed variant confirmation. L.R., B.H. and M.O.B. provided intellectual support in interpreting and reporting the sequencing results. The figures were prepared by H.C. The manuscript was written by M.S.O., L.V., M.O.B., H.C., J.A.V. and H.S. All authors contributed to this report.

Funding

This project was supported in part by funding from the Australian National Health and Medical Research Council (APP1120356) to M.K.O.B., J.A.V. and R.I.M.L., The Netherlands Organisation for Scientific Research (918-15-667) to J.A.V, the Royal Society and Wolfson Foundation (WM160091) to J.A.V., as well as an Investigator Award in Science from the Wellcome Trust (209451) to J.A.V. and Grants from the National Research Council of Argentina (PIP 0900 and 4584) and Agency for the Promotion of Science and Technology (PICT 9591) to H.E.C. and a UUKi Rutherford Fund Fellowship awarded to B.J.H.

Conflict of interest

The authors have nothing to disclose.

References

Afzelius BA, Eliasson R, Johnsen O, Lindholmer C. Lack of dynein arms in immotile human spermatozoa. *J Cell Biol* 1975;**66**:225–232.

Amiri-Yekta A, Coutton C, Kherraf ZE, Karaouzene T, Le TP, Sanati MH, Sabbaghian M, Almadani N, Sadighi GM, Hosseini SH. *et al.* Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations. *Hum Reprod* 2016;**31**:2872–2880.

Auguste Y, Delague V, Desvignes JP, Longepied G, Gnisci A, Besnier P, Levy N, Beroud C, Megarbane A, Metzler-Guillemain C. *et al.* Loss of calmodulin- and radial-spoke-associated complex protein CFAP251 leads to immotile spermatozoa lacking mitochondria and infertility in men. *Am J Hum Genet* 2018;**103**:413–420.

Baccetti B, Burrini AG, Pallini B, Renieri T. Human dynein and sperm pathology. *J Cell Biol* 1981;**88**:102–107.

Ben Khelifa M, Coutton C, Zouari R, Karaouzene T, Rendu J, Bidart M, Yassine S, Pierre V, Delaroche J, Hennebicq S. *et al.* Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *Am J Hum Genet* 2014;**94**:95–104.

Beurois J, Martinez G, Cazin C, Kherraf ZE, Amiri-Yekta A, Thierry-Mieg N, Bidart M, Petre G, Satre V, Brouillet S. *et al.* CFAP70 mutations lead to male infertility due to severe astheno-teratozoospermia. A case report. *Hum Reprod* 2019;**34**:2071–2079.

Brown PR, Miki K, Harper DB, Eddy EM. A-kinase anchoring protein 4 binding proteins in the fibrous sheath of the sperm flagellum. *Biol Reprod* 2003;**68**:2241–2248.

Carlen B, Lindberg S, Stenram U. Absence of nexin links as a possible cause of primary ciliary dyskinesia. *Ultrastruct Pathol* 2003;**27**:123–126.

Chemes HE, Brugo S, Zanchetti F, Carrere C, Lavieri JC. Dysplasia of the fibrous sheath: an ultrastructural defect of human spermatozoa associated with sperm immotility and primary sterility. *Fertil Steril* 1987;**48**:664–669.

Chemes HE, Morero JL, Lavieri JC. Extreme asthenozoospermia and chronic respiratory disease: a new variant of the immotile cilia syndrome. *Int J Androl* 1990;**13**:216–222.

Chemes HE, Olmedo SB, Carrere C, Oses R, Carizza C, Leisner M, Blaquier J. Ultrastructural pathology of the sperm flagellum: association between flagellar pathology and fertility prognosis in severely asthenozoospermic men. *Hum Reprod* 1998;**13**:2521–2526.

Coutton C, Martinez G, Kherraf ZE, Amiri-Yekta A, Boguenet M, Saut A, He X, Zhang F, Cristou-Kent M, Escoffier J. *et al.* Bi-allelic mutations in ARMC2 lead to severe astheno-teratozoospermia due to sperm flagellum malformations in humans and mice. *Am J Hum Genet* 2019;**104**:331–340.

Coutton C, Vargas AS, Amiri-Yekta A, Kherraf ZE, Ben MS, Le TP, Wambergue-Legrand C, Karaouzene T, Martinez G, Crouzy S. *et al.* Mutations in CFAP43 and CFAP44 cause male infertility and flagellum defects in Trypanosoma and human. *Nat Commun* 2018;**9**:686.

Dong FN, Amiri-Yekta A, Martinez G, Saut A, Tek J, Stouvenel L, Lores P, Karaouzene T, Thierry-Mieg N, Satre V. *et al.* Absence of CFAP69 causes male infertility due to multiple morphological abnormalities of the flagella in human and mouse. *Am J Hum Genet* 2018;**102**:636–648.

Dumur V, Gervais R, Rigot JM, Lafitte JJ, Manouvrier S, Biserte J, Mazeman E, Roussel P. Abnormal distribution of CF delta F508 allele in azoospermic men with congenital aplasia of epididymis and vas deferens. *Lancet* 1990;**336**:512.

Escudier E, Duquesnoy P, Papon JF, Amselem S. Ciliary defects and genetics of primary ciliary dyskinesia. *Paediatr Respir Rev* 2009;**10**:51–54.

Esteves SC, Chan P. A systematic review of recent clinical practice guidelines and best practice statements for the evaluation of the infertile male. *Int Urol Nephrol* 2015;**47**:1441–1456.

Fernandez-Gonzalez A, Kourembanas S, Wyatt TA, Mitsialis SA. Mutation of murine adenylate kinase 7 underlies a primary ciliary dyskinesia phenotype. *Am J Respir Cell Mol Biol* 2009;**40**:305–313.

Gershoni M, Hauser R, Yogev L, Lehavi O, Azem F, Yavetz H, Pietrokovski S, Kleiman SE. A familial study of azoospermic men identifies three novel causative mutations in three new human azoospermia genes. *Genet Med* 2017;**19**:998–1006.

GTEx C. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**;648–660.

Halaszovich CR, Leitner MG, Mavrantoni A, Le A, Frezza L, Feuer A, Schreiber DN, Villalba-Galea CA, Oliver D. A human phospholipid phosphatase activated by a transmembrane control module. *J Lipid Res* 2012;**53**:2266–2274.

He X, Li W, Wu H, Lv M, Liu W, Liu C, Zhu F, Li C, Fang Y, Yang C. *et al.* Novel homozygous CFAP69 mutations in humans and mice cause severe asthenoteratospermia with multiple morphological abnormalities of the sperm flagella. *J Med Genet* 2019;**56**: 96–103.

He X, Liu C, Yang X, Lv M, Ni X, Li Q, Cheng H, Liu W, Tian S, Wu H. *et al.* Bi-allelic loss-of-function variants in CFAP58 cause flagellar axoneme and mitochondrial sheath defects and asthenoteratozoospermia in humans and mice. *Am J Hum Genet* 2020;**107**: 514–526.

Houston BJ, Oud MS, Aguirre DM, Merriner DJ, O'Connor AE, Okutman O, Viville S, Burke R, Veltman JA, O'Bryan MK. Programmed cell death 2-like (Pdcd2l) is required for mouse embryonic development. *G3 (Bethesda)* 2020;**7**:247–255.

Hu J, Lessard C, Longstaff C, O'Brien M, Palmer K, Reinholdt L, Eppig J, Schimenti J, Handel MA. ENU-induced mutant allele of Dnah1, ferf1, causes abnormal sperm behavior and fertilization failure in mice. *Mol Reprod Dev* 2019;**86**:416–425.

Ichijima Y, Ichijima M, Lou Z, Nussenzweig A, Camerini-Otero RD, Chen J, Andreassen PR, Namekawa SH. MDC1 directs chromosome-wide silencing of the sex chromosomes in male germ cells. *Genes Dev* 2011;**25**:959–971.

Jungwirth A,T, Kopa Z, Krausz C, Minhas S, Tournaye H. European Association of Urology guidelines on Male Infertility edition presented at the EAU Annual Congress Copenhagen 2018; **2018**.

Keicho N, Hijikata M, Morimoto K, Homma S, Taguchi Y, Azuma A, Kudoh S. Primary ciliary dyskinesia caused by a large homozygous deletion including exons 1-4 of DRC1 in Japanese patients with recurrent sinopulmonary infection. *Mol Genet Genomic Med* 2020;**8**: e1033.

Kherraf ZE, Amiri-Yekta A, Dacheux D, Karaouzene T, Coutton C, Christou-Kent M, Martinez G, Landrein N, Le TP, Fourati BMS. *et al.* A homozygous ancestral SVA-insertion-mediated deletion in WDR66 induces multiple morphological abnormalities of the sperm flagellum and male infertility. *Am J Hum Genet* 2018;**103**: 400–412.

Kherraf ZE, Cazin C, Coutton C, Amiri-Yekta A, Martinez G, Boguenet M, Fourati BMS, Kharouf M, Gourabi H, Hosseini SH. *et al.* Whole exome sequencing of men with multiple morphological abnormalities of the sperm flagella reveals novel homozygous QRICH2 mutations. *Clin Genet* 2019;**96**:394–401.

Kraatz S, Guichard P, Obbineni JM, Olieric N, Hatzopoulos GN, Hilbert M, Sen I, Missimer J, Gonczy P, Steinmetz MO. The human centriolar protein CEP135 contains a two-stranded coiled-coil domain critical for microtubule binding. *Structure* 2016;**24**: 1358–1371.

Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 2006;**38**:999–1001.

Lancaster MA, Gopal DJ, Kim J, Saleem SN, Silhavy JL, Louie CM, Thacker BE, Williams Y, Zaki MS, Gleeson JG. Defective Wnt-dependent cerebellar midline fusion in a mouse model of Joubert syndrome. *Nat Med* 2011;**17**:726–731.

Li W, Tang W, Teves ME, Zhang Z, Zhang L, Li H, Archer KJ, Peterson DL, Williams DC Jr, Strauss JF, 3rd. *et al.* A MEIG1/PACRG complex in the manchette is essential for building the sperm flagella. *Development* 2015;**142**:921–930.

Li W, Wu H, Li F, Tian S, Kherraf ZE, Zhang J, Ni X, Lv M, Liu C, Tan Q *et al.*. Biallelic mutations in CFAP65 cause male infertility with multiple morphological abnormalities of the sperm flagella in humans and mice. J Med Genet 2029a;**57**:89–95.

Li Y, Sha Y, Wang X, Ding L, Liu W, Ji Z, Mei L, Huang X, Lin S, Kong S. *et al.* DNAH2 is a novel candidate gene associated with multiple morphological abnormalities of the sperm flagella. *Clin Genet* 2019b;**95**:590–600.

Li Y, Yagi H, Onuoha EO, Damerla RR, Francis R, Furutani Y, Tariq M, King SM, Hendricks G, Cui C. *et al.* DNAH6 and its interactions with PCD genes in heterotaxy and primary ciliary dyskinesia. *PLoS Genet* 2016;**12**:e1005821.

Liu C, He X, Liu W, Yang S, Wang L, Li W, Wu H, Tang S, Ni X, Wang J. *et al.* Bi-allelic mutations in TTC29 cause male subfertility with asthenoteratospermia in humans and mice. *Am J Hum Genet* 2019a;**105**:1168–1181.

Liu C, Lv M, He X, Zhu Y, Amiri-Yekta A, Li W, Wu H, Kherraf ZE, Liu W, Zhang J *et al.*. Homozygous mutations in SPEF2 induce multiple morphological abnormalities of the sperm flagella and male infertility. J Med Genet 2029b;**57**:31–37.

Liu W, He X, Yang S, Zouari R, Wang J, Wu H, Kherraf ZE, Liu C, Coutton C, Zhao R. *et al.* Bi-allelic mutations in TTC21A induce asthenoteratospermia in humans and mice. *Am J Hum Genet* 2019c;**104**:738–748.

Liu W, Sha Y, Li Y, Mei L, Lin S, Huang X, Lu J, Ding L, Kong S, Lu Z. Loss-of-function mutations in SPEF2 cause multiple morphological abnormalities of the sperm flagella (MMAF). *J Med Genet* 2019d;**56**:678–684.

Liu W, Wu H, Wang L, Yang X, Liu C, He X, Li W, Wang J, Chen Y, Wang H. *et al.* Homozygous loss-of-function mutations in FSIP2 cause male infertility with asthenoteratospermia. *J Genet Genomics* 2019e;**46**:53–56.

Lorenzetti D, Bishop CE, Justice MJ. Deletion of the Parkin coregulated gene causes male sterility in the quaking(viable) mouse mutant. *Proc Natl Acad Sci U S A* 2004;**101**:8402–8407.

Lores P, Coutton C, El Khouri E, Stouvenel L, Givelet M, Thomas L, Rode B, Schmitt A, Louis B, Sakheli Z. *et al.* Homozygous missense mutation L673P in adenylate kinase 7 (AK7) leads to primary male infertility and multiple morphological anomalies of the flagella but not to primary ciliary dyskinesia. *Hum Mol Genet* 2018;**27**: 1196–1211.

Lores P, Dacheux D, Kherraf ZE, Nsota MJ, Coutton C, Stouvenel L, Ialy-Radio C, Amiri-Yekta A, Whitfield M, Schmitt A. *et al.* Mutations in TTC29, encoding an evolutionarily conserved axonemal protein, result in asthenozoospermia and male infertility. *Am J Hum Genet* 2019;**105**:1148–1167.

Lou Z, Minter-Dykhouse K, Franco S, Gostissa M, Rivera MA, Celeste A, Manis JP, van Deursen J, Nussenzweig A, Paull TT. *et al.* MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. *Mol Cell* 2006; **21**:187–200.

Lv M, Liu W, Chi W, Ni X, Wang J, Cheng H, Li WY, Yang S, Wu H, Zhang J. *et al.* Homozygous mutations in DZIP1 can induce asthenoteratospermia with severe MMAF. *J Med Genet* 2020;**57**: 445–453.

Martinez G, Beurois J, Dacheux D, Cazin C, Bidart M, Kherraf ZE, Robinson Dr, Satre V, Le GG, Ka C. *et al.* Biallelic variants in MAATS1 encoding CFAP91, a calmodulin-associated and spoke-associated complex protein, cause severe astheno-teratozoospermia and male infertility. *J Med Genet* 2020;**0**:1–9.

Martinez G, Kherraf ZE, Zouari R, Fourati BMS, Saut A, Pernet-Gallay K, Bertrand A, Bidart M, Hograindleur JP, Amiri-Yekta A. *et al.* Whole-exome sequencing identifies mutations in FSIP2 as a recurrent cause of multiple morphological abnormalities of the sperm flagella. *Hum Reprod* 2018;**33**:1973–1984.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**:1297–1303.

Miki K, Willis WD, Brown PR, Goulding EH, Fulcher KD, Eddy EM. Targeted disruption of the Akap4 gene causes defects in sperm flagellum and motility. *Dev Biol* 2002;**248**:331–342.

Moretti E, Geminiani M, Terzuoli G, Renieri T, Pascarelli N, Collodel G. Two cases of sperm immotility: a mosaic of flagellar alterations related to dysplasia of the fibrous sheath and abnormalities of head-neck attachment. *Fertil Steril* 2011;**95**:1787.e1719–1723.

Morimoto K, Hijikata M, Zariwala MA, Nykamp K, Inaba A, Guo TC, Yamada H, Truty R, Sasaki Y, Ohta K. *et al.* Recurring large deletion in DRC1 (CCDC164) identified as causing primary ciliary dyskinesia in two Asian patients. *Mol Genet Genomic Med* 2019a;**7**: e838.

Morimoto Y, Yoshida S, Kinoshita A, Satoh C, Mishima H, Yamaguchi N, Matsuda K, Sakaguchi M, Tanaka T, Komohara Y. *et al.* Nonsense mutation in CFAP43 causes normal-pressure hydrocephalus with ciliary abnormalities. *Neurology* 2019b;**92**: e2364–e2374.

Neesen J, Kirschner R, Ochs M, Schmiedl A, Habermann B, Mueller C, Holstein AF, Nuesslein T, Adham I, Engel W. Disruption of an inner arm dynein heavy chain gene results in asthenozoospermia and reduced ciliary beat frequency. *Hum Mol Genet* 2001;**10**: 1117–1128.

Niemeyer J, Mentrup T, Heidasch R, Muller SA, Biswas U, Meyer R, Papadopoulou AA, Dederer V, Haug-Kroper M, Adamski V. *et al.* The intramembrane protease SPPL2c promotes male germ cell development by cleaving phospholamban. *EMBO Rep* 2019;**20**: e46449.

Oud MS, Volozonoka L, Smits RM, Vissers LELM, Ramos L, Veltman JA. A systematic review and standardized clinical validity assessment of male infertility genes. *Hum Reprod* 2019;**34**:932–941.

Panayiotou C, Solaroli N, Xu Y, Johansson M, Karlsson A. The characterization of human adenylate kinases 7 and 8 demonstrates differences in kinetic parameters and structural organization among the family of adenylate kinase isoenzymes. *Biochem J* 2011;**433**: 527–534.

Papadopoulou AA, Muller SA, Mentrup T, Shmueli MD, Niemeyer J, Haug-Kroper M, von BJ, Mayerhofer A, Feederle R, Schroder B. *et al.* Signal Peptide Peptidase-Like 2c (SPPL2c) impairs vesicular transport and cleavage of SNARE proteins. *EMBO Rep* 2019;**20**: e46451.

Pleuger C, Lehti MS, Dunleavy JEM, Fietz D, O'Bryan MK. Haploid male germ cells – the Grand Central Station of protein transport. *Hum Reprod Update* 2020;**26**:474–500.

Rawe VY, Terada Y, Nakamura S, Chillik CF, Olmedo SB, Chemes HE. A pathology of the sperm centriole responsible for defective sperm aster formation, syngamy and cleavage. *Hum Reprod* 2002; **17**:2344–2349.

Rebbe H, Pedersen H. Absence of arms in the axoneme of immobile human spermatozoa1. *Biol Reprod* 1975;**12**:541–544.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E. *et al.* 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. *Genet Med* 2015;**17**:405–424.

Rossman CM, Forrest JB, Lee RM, Newhouse AF, Newhouse MT. The dyskinetic cilia syndrome; abnormal ciliary motility in association with abnormal ciliary ultrastructure. *Chest* 1981;**80**:860–865.

Schuh K, Cartwright EJ, Jankevics E, Bundschu K, Liebermann J, Williams JC, Armesilla AL, Emerson M, Oceandy D, Knobeloch KP. *et al.* Plasma membrane Ca2+ ATPase 4 is required for sperm motility and male fertility. *J Biol Chem* 2004;**279**:28220–28226.

Sha Y, Wei X, Ding L, Mei L, Huang X, Lin S, Su Z, Kong L, Zhang Y, Ji Z. DNAH17 is associated with asthenozoospermia and multiple morphological abnormalities of sperm flagella. *Ann Hum Genet* 2019a;**84**:271–279.

Sha Y, Yang X, Mei L, Ji Z, Wang X, Ding L, Li P, and, Yang S. DNAH1 gene mutations and their potential association with dysplasia of the sperm fibrous sheath and infertility in the Han Chinese population. *Fertil Steril* 2017;**107**:1312,1318–e1312.

Sha YW, Wang X, Su ZY, Mei LB, Ji ZY, Bao H, Li P. Patients with multiple morphological abnormalities of the sperm flagella harbouring CFAP44 or CFAP43 mutations have a good pregnancy outcome following intracytoplasmic sperm injection. *Andrologia* 2019b; **51**:e13151.

Sha YW, Xu X, Mei LB, Li P, Su ZY, He Xq, Li L. A homozygous CEP135 mutation is associated with multiple morphological abnormalities of the sperm flagella (MMAF). *Gene* 2017;**633**:48–53.

Shamoto N, Narita K, Kubo T, Oda T, Takeda S. CFAP70 is a novel axoneme-binding protein that localizes at the base of the outer dynein arm and regulates ciliary motility. *Cells* 2018;**7**:124.

Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D. *et al.* Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 2006;**38**:1032–1037.

Shen Y, Zhang F, Li F, Jiang X, Yang Y, Li X, Li W, Wang X, Cheng J, Liu M. *et al.* Loss-of-function mutations in QRICH2 cause male infertility with multiple morphological abnormalities of the sperm flagella. *Nat Commun* 2019;**10**:433.

Stokowy T, Garbulowski M, Fiskerstrand T, Holdhus R, Labun K, Sztromwasser P, Gilissen C, Hoischen A, Houge G, Petersen K. *et al.* RareVariantVis: new tool for visualization of causative variants in rare monogenic disorders using whole genome sequencing data. *Bioinformatics* 2016;**32**:3018–3020.

Sutton KA, Jungnickel MK, Jovine L, Florman HM. Evolution of the voltage sensor domain of the voltage-sensitive phosphoinositide phosphatase VSP/TPTE suggests a role as a proton channel in eutherian mammals. *Mol Biol Evol* 2012;**29**:2147–2155.

Takeuchi K, Xu Y, Kitano M, Chiyonobu K, Abo M, Ikegami K, Ogawa S, Ikejiri M, Kondo M, Gotoh S. *et al.* Copy number variation in DRC1 is the major cause of primary ciliary dyskinesia in the Japanese population. *Mol Genet Genomic Med* 2020;**8**:e1137.

Tang S, Wang X, Li W, Yang X, Li Z, Liu W, Li C, Zhu Z, Wang L, Wang J. *et al.* Biallelic mutations in CFAP43 and CFAP44 cause male infertility with multiple morphological abnormalities of the sperm flagella. *Am J Hum Genet* 2017;**100**:854–864.

Toure A, Martinez G, Kherraf ZE, Cazin C, Beurois J, Arnoult C, Ray PF, Coutton C. The genetic architecture of morphological abnormalities of the sperm tail. Hum Genet 2021;**140**:21–42.

Tu C, Nie H, Meng L, Wang W, Li H, Yuan S, Cheng D, He W, Liu G, Du J. *et al.* Novel mutations in SPEF2 causing different defects between flagella and cilia bridge: the phenotypic link between MMAF and PCD. *Hum Genet* 2020;**139**:257–271.

Tu C, Nie H, Meng L, Yuan S, He W, Luo A, Li H, Li W, Du J, Lu G. *et al.* Identification of DNAH6 mutations in infertile men with multiple morphological abnormalities of the sperm flagella. *Sci Rep* 2019;**9**;15864.

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 2010;**28**: 1248–1250.

Visser L, Westerveld GH, Xie F, van Daalen SK, van der Veen F, Lombardi MP, Repping S. A comprehensive gene mutation screen in men with asthenozoospermia. *Fertil Steril* 2011;**95**:1020–1024. e1021–1029.

Wambergue C, Zouari R, Fourati BMS, Martinez G, Devillard F, Hennebicq S, Satre V, Brouillet S, Halouani L, Marrakchi O. *et al.* Patients with multiple morphological abnormalities of the sperm flagella due to DNAH1 mutations have a good prognosis following intracytoplasmic sperm injection. *Hum Reprod* 2016;**31**:1164–1172.

Wang W, Tu C, Nie H, Meng L, Li Y, Yuan S, Zhang Q, Du J, Wang J, Gong F. *et al.* Biallelic mutations in CFAP65 lead to severe asthenoteratospermia due to acrosome hypoplasia and flagellum malformations. *J Med Genet* 2019;**56**:750–757.

Wang X, Jin H, Han F, Cui Y, Chen J, Yang C, Zhu P, Wang W, Jiao G, Wang W. *et al.* Homozygous DNAH1 frameshift mutation causes multiple morphological anomalies of the sperm flagella in Chinese. *Clin Genet* 2017;**91**:313–321.

Whitfield M, Thomas L, Bequignon E, Schmitt A, Stouvenel L, Montantin G, Tissier S, Duquesnoy P, Copin B, Chantot S. *et al.* Mutations in DNAH17, encoding a sperm-specific axonemal outer dynein arm heavy chain, cause isolated male infertility due to asthenozoospermia. *Am J Hum Genet* 2019;**105**:198–212.

Wilson GR, Wang HX, Egan GF, Robinson PJ, Delatycki MB, O'Bryan MK, Lockhart PJ. Deletion of the Parkin co-regulated gene causes defects in ependymal ciliary motility and hydrocephalus in the quakingviable mutant mouse. *Hum Mol Genet* 2010;**19**: 1593–1602.

Wirschell M, Olbrich H, Werner C, Tritschler D, Bower R, Sale WS, Loges NT, Pennekamp P, Lindberg S, Stenram U. *et al.* The nexin-dynein regulatory complex subunit DRC1 is essential for motile cilia function in algae and humans. *Nat Genet* 2013;**45**:262–268.

Wu H, Li W, He X, Liu C, Fang Y, Zhu F, Jiang H, Liu W, Song B, Wang X. *et al.* Novel CFAP43 andCFAP44 mutations cause male infertility with multiple morphological abnormalities of the sperm flagella (MMAF). *Reprod Biomed Online* 2019;**38**:769–778.

Wyrwoll MJ, Temel ŞG, Nagirnaja L, Oud MS, Lopes AM, van der Heijden GW, Rotte N, Wistuba J, Wöste M, Ledig S. *et al.* Biallelic mutations in M1AP are a frequent cause of meiotic arrest leading to male infertility. *bioRxiv* 2019;**803346**.

Yzer S, Hollander AI, Lopez I, Pott JW, de Faber JT, Cremers FP, Koenekoop RK, van den Born LI. Ocular and extra-ocular features of patients with Leber congenital amaurosis and mutations in CEP290. *Mol Vis* 2012;**18**:412–425.

Zhang B, Ma H, Khan T, Ma A, Li T, Zhang H, Gao J, Zhou J, Li Y, Yu C. *et al.* A DNAH17 missense variant causes flagella destabilization and asthenozoospermia. *J Exp Med* 2019a;**217**:e20182365.

Zhang X, Shen Y, Wang X, Yuan G, Zhang C, Yang Y. A novel homozygous CFAP65 mutation in humans causes male infertility with multiple morphological abnormalities of the sperm flagella. *Clin Genet* 2019b; **96**:541–548.

9.2 Appendix B: List of Genes Carrying *De Novo* Mutations of Unknown Significance or Classified as Possibly Causative of the Disease in the Trio Cohort.

The *de novo* mutations were identified and classified by Dr Manon Oud and Hannah Smith. The classification criteria are described in Oud *et al.* 2020.

ABCF3, ABLIM1, AMPD2, APC2, ARHGAP33, ASIC5, ATP1A1, ATP8B4, C12orf49, C6orf25, C9orf50, CD81, CDC5L, CDK5RAP2, CELSR2, CLUH, CRHR1, CTNND2, CWC27, DHX36, DNAJC2, DNMT1, EMILIN1, ERG, EVC, EXOSC10, FBXO5, FIZ1, FLNC, FNDC8, FOXF2, FUS, GHRHR, GPR75-ASB3, GREB1L, HELZ2, HNRNPL, HOXA1, HR, HTT, ILVBL, INO80, IQSEC1, ITSN2, KBTBD12, KLC1, LEO1, LRRN2, MAP3K3, MCM6, MPRIP, MRPS9, MSH5, MYH3, MYOF, NEO1, ODF1, OR5P3, OSBPL3, PCDHB1, PLCL1, POPDC3, PPP1R7, PRDM16, PRPF4B, RASAL2, RASEF, RBM5, REN, RPA1, SDF4, SIKE1, SIPA1L3, SMC2, SOGA1, SPECC1L, SSH2, STARD10, STXBP2, TACC2, TAF9, TCFL5, TENM2, TLN2, TMEM62, TMPPE, TMPRSS11B, TOPAZ1, TP53TG5, TRAF7, U2AF2, WDR17, ZFHX4, ZNF292, ZNF469, ZNF629

Chapter 10. References

Acuna-Hidalgo, R., Veltman, J. A. and Hoischen, A. (2016) 'New insights into the generation and role of de novo mutations in health and disease.', *Genome biology*, 17(1), p. 241. doi: 10.1186/s13059-016-1110-1.

Adamson, G. D. *et al.* (2018) 'International Committee for Monitoring Assisted Reproductive Technology: world report on assisted reproductive technology, 2011', *Fertility and Sterility*, 110(6), pp. 1067–1080. doi: 10.1016/j.fertnstert.2018.06.039.

Aitken, R. J. (2018) 'Not every sperm is sacred; a perspective on male infertility', *Molecular Human Reproduction*, 24(6), pp. 287–298. doi: 10.1093/molehr/gay010.

Aitken, R. J. and Koppers, A. J. (2011) 'Apoptosis and DNA damage in human spermatozoa', *Asian Journal of Andrology*, 13(1), pp. 36–42. doi: 10.1038/aja.2010.68.

Alksere, B. *et al.* (2019) 'Case of Inherited Partial AZFa Deletion without Impact on Male Fertility', *Case Reports in Genetics*, 2019, pp. 1–5. doi: 10.1155/2019/3802613.

Arafat, M. *et al.* (2017) 'Mutation in TDRD9 causes non-obstructive azoospermia in infertile men', *Journal of Medical Genetics*, 54(9), pp. 633–639. doi: 10.1136/jmedgenet-2017-104514.

Aston, K. I. *et al.* (2010) 'Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of European descent', *Human Reproduction*, 25(6), pp. 1383–1397. doi: 10.1093/humrep/deq081.

Aston, K. I. (2014) 'Genetic susceptibility to male infertility: news from genome-wide association studies', *Andrology*, 2(3), pp. 315–321. doi: 10.1111/j.2047-2927.2014.00188.x.

Aston, K. I. and Carrell, D. T. (2009) 'Genome-wide study of single-nucleotide polymorphisms associated with azoospermia and severe oligozoospermia', *Journal of Andrology*, 30(6), pp. 711–725. doi: 10.2164/jandrol.109.007971.

Van der Auwera, G. A. *et al.* (2013) 'From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline', in *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.

Avidan, N. *et al.* (2003) 'CATSPER2, a human autosomal nonsyndromic male infertility gene', *European Journal of Human Genetics*, 11(7), pp. 497–502. doi:

10.1038/sj.ejhg.5200991.

Babakhanzadeh, E. *et al.* (2020) 'Testicular expression of TDRD1, TDRD5, TDRD9 and TDRD12 in azoospermia', *BMC Medical Genetics*, 21(1), p. 33. doi: 10.1186/s12881-020-0970-0.

Bamshad, M. J. *et al.* (2011) 'Exome sequencing as a tool for Mendelian disease gene discovery', *Nature Reviews Genetics*, 12(11), pp. 745–755. doi: 10.1038/nrg3031.

Bamshad, M. J., Nickerson, D. A. and 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. doi: 10.1016/j.ajhg.2019.07.011.

Belva, F. *et al.* (2016) 'Semen quality of young adult ICSI offspring: The first results', *Human Reproduction*, 31(12), pp. 2811–2820. doi: 10.1093/humrep/dew245.

Belva, F., Bonduelle, M. and Tournaye, H. (2019) 'Endocrine and reproductive profile of boys and young adults conceived after ICSI', *Current Opinion in Obstetrics and Gynecology*. Lippincott Williams and Wilkins, pp. 163–169. doi: 10.1097/GCO.0000000000000538.

Belyeu, J. R. *et al.* (2021) 'De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families', *The American Journal of Human Genetics*, 0(0). doi: 10.1016/j.ajhg.2021.02.012.

Ben Khelifa, M. *et al.* (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. doi: 10.1016/J.AJHG.2013.11.017.

Boraldi, F. *et al.* (2019) 'Exome sequencing and bioinformatic approaches reveals rare sequence variants involved in cell signalling and elastic fibre homeostasis: new evidence in the development of ectopic calcification', *Cellular Signalling*. Elsevier Inc., pp. 131–140. doi: 10.1016/j.cellsig.2019.03.020.

Braun, R. E. *et al.* (1989) 'Genetically haploid spermatids are phenotypically diploid', *Nature*, 337(6205), pp. 373–376. doi: 10.1038/337373a0.

Brugh, V. M. and Lipshultz, L. I. (2004) 'Male factor infertility: evaluation and management.', *Medical Clinics of North America*, 88(2), pp. 367–385. doi: 10.1016/S0025-7125(03)00150-0.

Bui, C. *et al.* (2014) 'XYLT1 mutations in desbuquois dysplasia type 2', *American*

Journal of Human Genetics, 94(3), pp. 405–414. doi: 10.1016/j.ajhg.2014.01.020.

Burgoyne, P. S. *et al.* (1992) 'Fertility in mice requires X-Y pairing and a Y-chromosomal "Spermiogenesis" gene mapping to the long arm', *Cell*, 71(3), pp. 391–398. doi: 10.1016/0092-8674(92)90509-B.

Capel, B. (1998) 'SEX IN THE 90s: *SRY* and the Switch to the Male Pathway', *Annual Review of Physiology*, 60(1), pp. 497–523. doi: 10.1146/annurev.physiol.60.1.497.

Catford, S. R. *et al.* (2017) 'Long-term follow-up of intra-cytoplasmic sperm injection-conceived offspring compared with in vitro fertilization-conceived offspring: a systematic review of health outcomes beyond the neonatal period', *Andrology*. Blackwell Publishing Ltd, pp. 610–621. doi: 10.1111/andr.12369.

Chandley, A. C. *et al.* (1972) 'Translocation heterozygosity and associated subfertility in man.', *Cytogenetics*, 11(6), pp. 516–533. doi: 10.1159/000130218.

CHANDLEY, A. C. *et al.* (1975) 'Cytogenetics and infertility in man: I. Karyotype and seminal analysis: Results of a five-year survey of men attending a subfertility clinic', *Annals of Human Genetics*, 39(2), pp. 231–254. doi: 10.1111/j.1469-1809.1975.tb00126.x.

Chen, H. *et al.* (2017) 'Human Spermatogenesis and Its Regulation', in *Male Hypogonadism*. Cham: Springer International Publishing, pp. 49–72. doi: 10.1007/978-3-319-53298-1_3.

Chen, W. S. *et al.* (2001) 'Growth retardation and increased apoptosis in mice with homozygous disruption of the akt1 gene', *Genes and Development*, 15(17), pp. 2203–2208. doi: 10.1101/gad.913901.

Cherry, N. *et al.* (2001) 'Occupational exposure to solvents and male infertility', *Occupational and Environmental Medicine*, 58(10), pp. 635–640. doi: 10.1136/oem.58.10.635.

Chianese, C. *et al.* (2014) 'X Chromosome-Linked CNVs in Male Infertility: Discovery of Overall Duplication Load and Recurrent, Patient-Specific Gains with Potential Clinical Relevance', *PLoS ONE*. Edited by J. R. Drevet, 9(6), p. e97746. doi: 10.1371/journal.pone.0097746.

Chilamakuri, C. S. R. *et al.* (2014) 'Performance comparison of four exome capture systems for deep sequencing', *BMC Genomics*, 15(1), p. 449. doi: 10.1186/1471-2164-15-449.

Chuma, S. *et al.* (2006) 'Tdrd1/Mtr-1, a tudor-related gene, is essential for male germ-cell differentiation and nuage/germinal granule formation in mice', *Proceedings of the National Academy of Sciences of the United States of America*, 103(43), pp. 15894–15899. doi: 10.1073/pnas.0601878103.

Collins, R. L. *et al.* (2020) 'A structural variation reference for medical and population genetics', *Nature*, 581(7809), pp. 444–451. doi: 10.1038/s41586-020-2287-8.

Corbett, M. A. *et al.* (2018) 'Pathogenic copy number variants that affect gene expression contribute to genomic burden in cerebral palsy', *npj Genomic Medicine*, 3(1), p. 33. doi: 10.1038/s41525-018-0073-4.

Coutton, C. *et al.* (2019) 'Bi-allelic Mutations in ARMC2 Lead to Severe Astheno-Teratozoospermia Due to Sperm Flagellum Malformations in Humans and Mice', *American Journal of Human Genetics*, 104(2), pp. 331–340. doi: 10.1016/j.ajhg.2018.12.013.

Cozzolino, D. J. and Lipshultz, L. I. (2001) 'Varicocele as a progressive lesion: positive effect of varicocele repair', *Human Reproduction Update*, 7(1), pp. 55–58. doi: 10.1093/humupd/7.1.55.

Dalgaard, M. D. *et al.* (2012) 'A genome-wide association study of men with symptoms of testicular dysgenesis syndrome and its network biology interpretation', *Journal of Medical Genetics*, 49(1), pp. 58–65. doi: 10.1136/jmedgenet-2011-100174.

Debiec-Rychter, M. *et al.* (1992) 'Two familial 9;17 translocations with variable effect on male carriers fertility', *Fertility and Sterility*, 57(4), pp. 933–935. doi: 10.1016/S0015-0282(16)54985-1.

Dixon, G. *et al.* (2021) 'QSER1 protects DNA methylation valleys from de novo methylation', *Science*, 372(6538), p. eabd0875. doi: 10.1126/science.abd0875.

Dong, Y. *et al.* (2015) 'Copy number variations in spermatogenic failure patients with chromosomal abnormalities and unexplained azoospermia', *Genetics and molecular research: GMR*, 14(4), pp. 16041–16049. doi: 10.4238/2015.December.7.17.

Dong, Y. and Simske, J. S. (2016) 'Vertebrate Claudin/PMP22/EMP22/MP20 family protein TMEM47 regulates epithelial cell junction maturation and morphogenesis', *Developmental Dynamics*, 245(6), pp. 653–666. doi: 10.1002/dvdy.24404.

Dyer, S. *et al.* (2016) 'International committee for monitoring assisted reproductive technologies world report: Assisted reproductive technology 2008, 2009 and 2010†', *Human Reproduction*, 31(7), pp. 1588–1609. doi: 10.1093/humrep/dew082.

Dym, M. and Fawcett, D. W. (1971) 'Further observations on the numbers of spermatogonia, spermatocytes, and spermatids connected by intercellular bridges in the mammalian testis.', *Biology of reproduction*, 4(2), pp. 195–215. doi: 10.1093/biolreprod/4.2.195.

Eggers, S. *et al.* (2015) 'Copy number variation associated with meiotic arrest in idiopathic male infertility', *Fertility and Sterility*, 103(1), pp. 214–219. doi: 10.1016/j.fertnstert.2014.09.030.

Elinati, E. *et al.* (2012) 'Globozoospermia is mainly due to dpy19l2 deletion via non-allelic homologous recombination involving two recombination hotspots', *Human Molecular Genetics*, 21(16), pp. 3695–3702. doi: 10.1093/hmg/dds200.

Emanuele, M. A. and Emanuele, N. V (1998) 'Alcohol's effects on male reproduction.', *Alcohol health and research world*, 22(3), pp. 195–201.

Epilepsy Phenome/Genome Project Epi4K Consortium (2015) 'Copy number variant analysis from exome data in 349 patients with epileptic encephalopathy', *Annals of Neurology*, 78(2), pp. 323–328. doi: 10.1002/ana.24457.

Ergin, R. N. *et al.* (2018) 'Social stigma and familial attitudes related to infertility', *Turk Jinekoloji ve Obstetrik Dernegi Dergisi*, 15(1), pp. 46–49. doi: 10.4274/tjod.04307.

Van Esch, H. *et al.* (2005) 'Deletion of VCX-A due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis', *Human Molecular Genetics*, 14(13), pp. 1795–1803. doi: 10.1093/hmg/ddi186.

Esplin, E. D. *et al.* (2014) 'Nine patients with Xp22.31 microduplication, cognitive deficits, seizures, and talipes anomalies', *American Journal of Medical Genetics, Part A*, 164(8), pp. 2097–2103. doi: 10.1002/ajmg.a.36598.

Esteves, S. C. (2015) 'Clinical management of infertile men with nonobstructive azoospermia', *Asian Journal of Andrology*. Medknow Publications, pp. 459–470. doi: 10.4103/1008-682X.148719.

Esteves, S. C. *et al.* (2018) 'Intracytoplasmic sperm injection for male infertility and consequences for offspring', *Nature Reviews Urology*. Nature Publishing Group, pp. 535–562. doi: 10.1038/s41585-018-0051-8.

Ferguson-Smith, M. A. *et al.* (1957) 'Klinefelter's syndrome; frequency and testicular morphology in relation to nuclear sex.', *The Lancet*, 270(6987), pp. 167–169. doi: 10.1016/S0140-6736(57)90617-7.

Ferlin, A. *et al.* (2006) 'Male infertility and androgen receptor gene mutations: Clinical features and identification of seven novel mutations', *Clinical Endocrinology*, 65(5), pp. 606–610. doi: 10.1111/j.1365-2265.2006.02635.x.

Fernandes, D. *et al.* (2007) 'RNAi - induced silencing of the plasma membrane Ca2+ - ATPase 2 in neuronal cells: Effects on Ca2+ homeostasis and cell viability', *Journal of Neurochemistry*, 102(2), pp. 454–465. doi: 10.1111/j.1471-4159.2007.04592.x.

Firth, H.V. et al. (2009) 'DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.', American journal of human genetics, 84(4), pp. 524–33. doi:10.1016/j.ajhg.2009.03.010.

Foresta, C. *et al.* (2000) 'Role of the AZFa candidate genes in male infertility', *Journal of Endocrinological Investigation*. Editrice Kurtis s.r.l., pp. 646–651. doi: 10.1007/BF03343788.

Fromer, M., Moran, Jennifer L, *et al.* (2012) 'Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth.', *American journal of human genetics*, 91(4), pp. 597–607. doi: 10.1016/j.ajhg.2012.08.005.

Fromer, M., Moran, Jennifer L., *et al.* (2012) 'Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth', *The American Journal of Human Genetics*, 91(4), pp. 597–607. doi: 10.1016/J.AJHG.2012.08.005.

Fromer, M. and Purcell, S. M. (2014) 'Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data', in *Current Protocols in Human Genetics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 7.23.1-7.23.21. doi: 10.1002/0471142905.hg0723s81.

Gabriel-Robez, O. *et al.* (1990) 'Deletion of the pseudoautosomal region and lack of sex-chromosome pairing at pachytene in two infertile man carrying an X;Y translocation', *Cytogenetics and Cell Genetics*, 54(1–2), pp. 38–42. doi: 10.1159/000132951.

Gambin, T. *et al.* (2017) 'Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort', *Nucleic acids research*, 45(4), pp. 1633–1648. doi: 10.1093/nar/gkw1237.

Gannon, K., Glover, L. and Abel, P. (2004) 'Masculinity, infertility, stigma and media reports', *Social Science and Medicine*, 59(6), pp. 1169–1175. doi: 10.1016/j.socscimed.2004.01.015.

Gel, B. and Serra, E. (2017) 'karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data', *Bioinformatics*, 33(19), pp. 3088–3090. doi: 10.1093/bioinformatics/btx346.

Geoffroy, V. *et al.* (2018) 'AnnotSV: an integrated tool for structural variations annotation', *Bioinformatics*. Edited by B. Berger, 34(20), pp. 3572–3574. doi: 10.1093/bioinformatics/bty304.

Gershoni, M. *et al.* (2017) 'A familial study of azoospermic men identifies three novel causative mutations in three new human azoospermia genes', *Genetics in Medicine*, 19(9), pp. 998–1006. doi: 10.1038/gim.2016.225.

Lo Giacco, D. *et al.* (2014) 'Recurrent X chromosome-linked deletions: Discovery of new genetic factors in male infertility', *Journal of Medical Genetics*, 51(5), pp. 340–344. doi: 10.1136/jmedgenet-2013-101988.

Gilissen, C. *et al.* (2014) 'Genome sequencing identifies major causes of severe intellectual disability', *Nature*, 511(7509), pp. 344–347. doi: 10.1038/nature13394.

Girirajan, S. *et al.* (2011) 'Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes', *PLOS Genetics*, 7(11), p. e1002334. doi: 10.1371/journal.pgen.1002334.

Goldmann, J. M. *et al.* (2016) 'Parent-of-origin-specific signatures of de novo mutations', *Nature Genetics*, 48(8), pp. 935–939. doi: 10.1038/ng.3597.

Goriely, A. et al. (2003) 'Evidence for Selective Advantage of Pathogenic FGFR2 Mutations in the Male Germ Line', Science, 301(5633), pp. 643–646. doi:10.1126/science.1085710.

Gou, L. T. *et al.* (2017) 'Ubiquitination-Deficient Mutations in Human Piwi Cause Male Infertility by Impairing Histone-to-Protamine Exchange during Spermiogenesis', *Cell*, 169(6), pp. 1090-1104.e13. doi: 10.1016/j.cell.2017.04.034.

Griswold, M. D. (1998) 'The central role of Sertoli cells in spermatogenesis', *Seminars in Cell & Developmental Biology*, 9(4), pp. 411–416. doi: 10.1006/scdb.1998.0203.

Griswold, M. D. (2016) 'Spermatogenesis: The Commitment to Meiosis', *Physiological Reviews*, 96(1), pp. 1–17. doi: 10.1152/physrev.00013.2015.

Grozdanov, P. N. *et al.* (2018) 'Cstf2t Regulates expression of histones and histone-like proteins in male germ cells', *Andrology*, 6(4), pp. 605–615. doi: 10.1111/andr.12488.

Guo, Y. *et al.* (2013) 'Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control', *BioMed Research International*, 2013. doi: 10.1155/2013/915636.

Harbuz, R. *et al.* (2011) 'A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation', *American Journal of Human Genetics*, 88(3), pp. 351–361. doi: 10.1016/j.ajhg.2011.02.007.

He, X. *et al.* (2020) 'Bi-allelic Loss-of-function Variants in CFAP58 Cause Flagellar Axoneme and Mitochondrial Sheath Defects and Asthenoteratozoospermia in Humans and Mice', *American Journal of Human Genetics*, 107(3), pp. 514–526. doi: 10.1016/j.ajhg.2020.07.010.

Hehir-Kwa, J. Y. *et al.* (2011) 'De novo copy number variants associated with intellectual disability have a paternal origin and age bias', *Journal of Medical Genetics*, 48(11), pp. 776–778. doi: 10.1136/jmedgenet-2011-100147.

Hehir-Kwa, J. Y., Pfundt, R. and Veltman, J. A. (2015) 'Exome sequencing and whole genome sequencing for the detection of copy number variation', *Expert Review of Molecular Diagnostics*. Taylor and Francis Ltd, pp. 1023–1032. doi: 10.1586/14737159.2015.1053467.

Helena Mangs, A. and Morris, B. (2007) 'The Human Pseudoautosomal Region (PAR): Origin, Function and Future', *Current Genomics*, 8(2), pp. 129–136. doi: 10.2174/138920207780368141.

Herold, A. *et al.* (2000) 'TAP (NXF1) Belongs to a Multigene Family of Putative RNA Export Factors with a Conserved Modular Architecture', *Molecular and Cellular Biology*, 20(23), pp. 8996–9008. doi: 10.1128/mcb.20.23.8996-9008.2000.

Hodžić, A. *et al.* (2020) 'De novo mutations in idiopathic male infertility - A pilot study', *Andrology*, p. andr.12897. doi: 10.1111/andr.12897.

Hoorsan, H. *et al.* (2017) 'Congenital malformations in infants of mothers undergoing assisted reproductive technologies: A systematic review and meta-analysis study', *Journal of Preventive Medicine and Public Health*. Korean Society for Preventive Medicine, pp. 347–360. doi: 10.3961/jpmph.16.122.

Hoppman, N. *et al.* (2013) 'Genetic testing for hearing loss in the United States should include deletion/duplication analysis for the deafness/infertility locus at 15q15.3', *Molecular Cytogenetics*, 6(1), p. 19. doi: 10.1186/1755-8166-6-19.

Hu, Z. *et al.* (2012) 'A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia', *Nature Genetics*, 44(2), pp. 183–186. doi: 10.1038/ng.1040.

Hu, Z. *et al.* (2014) 'Association analysis identifies new risk loci for non-obstructive azoospermia in Chinese men', *Nature Communications*, 5(1), pp. 1–7. doi: 10.1038/ncomms4857.

Huang, L. *et al.* (2016) 'Proteasome activators, PA28γ and PA200, play indispensable roles in male fertility', *Scientific Reports*, 6(1), pp. 1–9. doi: 10.1038/srep23171.

Ibrahim, O. *et al.* (2020) 'Saliva as a comparable-quality source of DNA for Whole Exome Sequencing on Ion platforms', *Genomics*, 112(2), pp. 1437–1443. doi: 10.1016/j.ygeno.2019.08.014.

Jacobs, P. A. and Strong, J. A. (1959) 'A case of human intersexuality having a possible XXY sex-determining mechanism', *Nature*, 183(4657), pp. 302–303. doi: 10.1038/183302a0.

Jamsai, D. and O'Bryan, M. K. (2011) 'Mouse models in male fertility research', *Asian Journal of Andrology*, 13(1), pp. 139–151. doi: 10.1038/aja.2010.101.

Jarow, J. P. (2001) 'Effects of varicocele on male fertility', *Human Reproduction Update*, 7(1), pp. 59–64. doi: 10.1093/humupd/7.1.59.

Ji, J. *et al.* (2016a) 'Copy number gain of VCX, X-linked multi-copy gene, leads to cell proliferation and apoptosis during spermatogenesis', *Oncotarget*, 7(48), pp. 78532–78540. doi: 10.18632/oncotarget.12397.

Jin, S. K. and Yang, W. X. (2017) 'Factors and pathways involved in capacitation: How are they regulated?', *Oncotarget*. Impact Journals LLC, pp. 3600–3627. doi: 10.18632/oncotarget.12274.

Jónsson, H. *et al.* (2017) 'Parental influence on human germline de novo mutations in 1,548 trios from Iceland', *Nature*, 549(7673), pp. 519–522. doi: 10.1038/nature24018.

Jungwirth A.D.T., Kopa Z., Krausz C., Minhas S. and Tournaye H. (2018) *European Association of Urology guidelines on male infertility presented at the EAU Annual Congress Copenhagen 2018*.

Kamata, A. *et al.* (2007) 'Spatiotemporal expression of four isoforms of Ca2+/calmodulin-dependent protein kinase I in brain and its possible roles in hippocampal dendritic growth', *Neuroscience Research*, 57(1), pp. 86–97. doi:

10.1016/j.neures.2006.09.013.

Karczewski, K. J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581(7809), pp. 434–443. doi: 10.1038/s41586-020-2308-7.

Khan, M. *et al.* (2018) 'The evolutionarily conserved genes: Tex37, Ccdc73, Prss55 and Nxt2 are dispensable for fertility in mice', *Scientific Reports*, 8(1), p. 4975. doi: 10.1038/s41598-018-23176-x.

Kherraf, Z. *et al.* (2019) 'Whole exome sequencing of men with multiple morphological abnormalities of the sperm flagella reveals novel homozygous *QRICH2* mutations', *Clinical Genetics*, 96(5), pp. 394–401. doi: 10.1111/cge.13604.

Kherraf, Z. E. *et al.* (2018) 'A Homozygous Ancestral SVA-Insertion-Mediated Deletion in WDR66 Induces Multiple Morphological Abnormalities of the Sperm Flagellum and Male Infertility', *American Journal of Human Genetics*, 103(3), pp. 400–412. doi: 10.1016/j.ajhg.2018.07.014.

Khor, B. *et al.* (2006) 'Proteasome Activator PA200 Is Required for Normal Spermatogenesis', *Molecular and Cellular Biology*, 26(8), pp. 2999–3007. doi: 10.1128/mcb.26.8.2999-3007.2006.

Kidd, J. M. *et al.* (2014) 'Exome capture from saliva produces high quality genomic and metagenomic data', *BMC Genomics*, 15(1), p. 262. doi: 10.1186/1471-2164-15-262.

Kim, S. T., Omurtag, K. and Moley, K. H. (2012) 'Decreased spermatogenesis, fertility, and altered Slc2A expression in Akt1-/-and Akt2-/-testes and sperm', *Reproductive Sciences*, 19(1), pp. 31–42. doi: 10.1177/1933719111424449.

Kimura, K., Cuvier, O. and Hirano, T. (2001) 'Chromosome Condensation by a Human Condensin Complex in Xenopus Egg Extracts', *Journal of Biological Chemistry*, 276(8), pp. 5417–5420. doi: 10.1074/jbc.C000873200.

Klinefelter, H. F., Reifenstein, E. C. and 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. doi: 10.1210/jcem-2-11-615.

Kloosterman, W. P. *et al.* (2015) 'Characteristics of de novo structural changes in the human genome', *Genome Research*, 25(6), pp. 792–801. doi: 10.1101/gr.185041.114.

Kong, S. W. *et al.* (2018) 'Measuring coverage and accuracy of whole-exome sequencing in clinical context', *Genetics in Medicine*, 20(12), pp. 1617–1626. doi: 10.1038/gim.2018.51.

Koppers, A. J. *et al.* (2011) 'Phosphoinositide 3-kinase signalling pathway involvement in a truncated apoptotic cascade associated with motility loss and oxidative DNA damage in human spermatozoa', *Biochemical Journal*, 436(3), pp. 687–698. doi: 10.1042/BJ20110114.

Kort, H. I. *et al.* (2006) 'Impact of body mass index values on sperm quantity and quality', *Journal of Andrology*, 27(3), pp. 450–452. doi: 10.2164/jandrol.05124.

Koscinski, I. *et al.* (2011) 'DPY19L2 deletion as a major cause of globozoospermia', *American Journal of Human Genetics*, 88(3), pp. 344–350. doi: 10.1016/j.ajhg.2011.01.018.

Koshikawa, N. *et al.* (1994) 'Identification of one- and two-chain forms of trypsinogen 1 produced by a human gastric adenocarcinoma cell line', *Biochemical Journal*, 303(1), pp. 187–190. doi: 10.1042/bj3030187.

Kosova, G. *et al.* (2012) 'Genome-wide association study identifies candidate genes for male fertility traits in humans', *American Journal of Human Genetics*, 90(6), pp. 950–961. doi: 10.1016/j.ajhg.2012.04.016.

Krausz, C. *et al.* (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. doi: 10.1093/hmg/ddl198.

Krausz, C. (2011) 'Male infertility: Pathogenesis and clinical diagnosis', *Best Practice & Research Clinical Endocrinology & Metabolism*, 25(2), pp. 271–285. doi: 10.1016/j.beem.2010.08.006.

Krausz, C. *et al.* (2012) 'High Resolution X Chromosome-Specific Array-CGH Detects New CNVs in Infertile Males', *PLoS ONE*, 7(10). doi: 10.1371/journal.pone.0044887.

Krausz, C., Escamilla, A. R. and Chianese, C. (2015) 'Genetics of male infertility: from research to clinic', *Reproduction*, 150(5), pp. R159–R174. doi: 10.1530/REP-15-0261.

Krausz, C. and Riera-Escamilla, A. (2018) 'Genetics of male infertility', *Nature Reviews Urology*, 15(6), pp. 369–384. doi: 10.1038/s41585-018-0003-3.

Krumm, N. *et al.* (2012) 'Copy number variation detection and genotyping from exome sequence data.', *Genome research*, 22(8), pp. 1525–32. doi: 10.1101/gr.138115.112.

Kuramochi-Miyagawa, S. *et al.* (2004) 'Mili, a mammalian member of piwi family gene, is essential for spermatogenesis', *Development*, 131(4), pp. 839–849. doi: 10.1242/dev.00973.

de la Chapelle, A. *et al.* (1964) 'XX Sex Chromosomes in a Human Male: First Case', *Acta Medica Scandinavica*, 175, pp. 25–38. doi: 10.1111/j.0954-6820.1964.tb04630.x.

Lahn, B. T. and Page, D. C. (2000) 'A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins', *Human Molecular Genetics*, 9(2), pp. 311–319. doi: 10.1093/hmg/9.2.311.

Lalitha, C. *et al.* (2013) 'Hormonal factors associated with hypogonadismand infertility in males -chromosomal abnormality', *IOSR Journal of Dental and Medical Sciences*, 10(1), pp. 2279–861.

Lee, K. H. *et al.* (2003) 'Ubiquitin-specific protease activity of USP9Y, a male infertility gene on the Y chromosome', *Reproduction, Fertility and Development*, 15(1–2), pp. 129–133. doi: 10.1071/rd03002.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536(7616), pp. 285–291. doi: 10.1038/nature19057.

Lelieveld, S. H. *et al.* (2015) 'Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions', *Human Mutation*, 36(8), pp. 815–822. doi: 10.1002/humu.22813.

Lelieveld, S. H. *et al.* (2016) 'Novel bioinformatic developments for exome sequencing', *Human Genetics*, 135, pp. 603–614. doi: 10.1007/s00439-016-1658-6.

Leppa, V. M. *et al.* (2016) 'Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families', *The American Journal of Human Genetics*, 99(3), pp. 540–554. doi: 10.1016/j.ajhg.2016.06.036.

Li, F. *et al.* (2010) 'Interstitial microduplication of Xp22.31: Causative of intellectual disability or benign copy number variant?', *European Journal of Medical Genetics*, 53(2), pp. 93–99. doi: 10.1016/j.ejmg.2010.01.004.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, W. *et al.* (2020) 'Biallelic mutations in *CFAP65* cause male infertility with multiple morphological abnormalities of the sperm flagella in humans and mice', *Journal of Medical Genetics*, 57(2), pp. 89–95. doi: 10.1136/jmedgenet-2019-106344.

Li, Y. R. *et al.* (2020) 'Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations', *Nature Communications*, 11(1), pp. 1–9. doi: 10.1038/s41467-019-13624-1.

de Ligt, J. *et al.* (2012) 'Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability', *New England Journal of Medicine*, 367(20), pp. 1921–1929. doi: 10.1056/NEJMoa1206524.

de Ligt, J. *et al.* (2013) 'Detection of Clinically Relevant Copy Number Variants with Whole-Exome Sequencing', *Human Mutation*, 34(10), pp. 1439–1448. doi: 10.1002/humu.22387.

Lim, Y. *et al.* (2017) 'The saliva microbiome profiles are minimally affected by collection method or DNA extraction protocols', *Scientific Reports*, 7, p. 8523. doi: 10.1038/s41598-017-07885-3.

Lima, A. C. *et al.* (2015) 'Rare double sex and mab-3-related transcription factor 1 regulatory variants in severe spermatogenic failure', *Andrology*, 3(5), pp. 825–833. doi: 10.1111/andr.12063.

Lishko, P. V., Botchkina, I. L. and Kirichok, Y. (2011) 'Progesterone activates the principal Ca2+ channel of human sperm', *Nature*, 471(7338), pp. 387–392. doi: 10.1038/nature09767.

Liu, C. *et al.* (2019) 'Bi-allelic Mutations in TTC29 Cause Male Subfertility with Asthenoteratospermia in Humans and Mice', *American Journal of Human Genetics*, 105(6), pp. 1168–1181. doi: 10.1016/j.ajhg.2019.10.010.

Liu, C. *et al.* (2020) 'Bi-allelic DNAH8 Variants Lead to Multiple Morphological Abnormalities of the Sperm Flagella and Primary Male Infertility', *American Journal of Human Genetics*, 107(2), pp. 330–341. doi: 10.1016/j.ajhg.2020.06.004.

Liu, C. *et al.* (2021) 'Deleterious variants in X-linked CFAP47 induce asthenoteratozoospermia and primary male infertility', *American Journal of Human Genetics*, 108(2), pp. 309–323. doi: 10.1016/j.ajhg.2021.01.002.

Lopes, A. M. *et al.* (2013) 'Human Spermatogenic Failure Purges Deleterious Mutation Load from the Autosomes and Both Sex Chromosomes, including the Gene

DMRT1', *PLoS Genetics*. Edited by E. Hollox, 9(3), p. e1003349. doi: 10.1371/journal.pgen.1003349.

Luddi, A. *et al.* (2009) ' Spermatogenesis in a Man with Complete Deletion of USP9Y ', *New England Journal of Medicine*, 360(9), pp. 881–885. doi: 10.1056/nejmoa0806218.

Luo, T. *et al.* (2019) 'A novel copy number variation in CATSPER2 causes idiopathic male infertility with normal semen parameters', *Human Reproduction*, 34(3). doi: 10.1093/humrep/dey377.

Lv, M. *et al.* (2020) 'Homozygous mutations in DZIP1 can induce asthenoteratospermia with severe MMAF', *Journal of Medical Genetics*, 57(7), pp. 445–453. doi: 10.1136/jmedgenet-2019-106479.

Ma, R. *et al.* (2017) 'A clear bias in parental origin of de novo pathogenic CNVs related to intellectual disability, developmental delay and multiple congenital anomalies', *Scientific Reports*, 7. doi: 10.1038/srep44446.

MacDonald, J. R. *et al.* (2014) 'The Database of Genomic Variants: a curated collection of structural variation in the human genome.', *Nucleic acids research*, 42(Database issue), pp. D986-92. doi: 10.1093/nar/gkt958.

Mahat, K. R. *et al.* (2016) 'Risk Factors and Causes of Male Infertility-A Review', *Biochemistry & Analytical Biochemistry*, 5(2). doi: 10.4172/2161-1009.1000271.

Marchuk, D. S. *et al.* (2018) 'Increasing the diagnostic yield of exome sequencing by copy number variant analysis', *PLOS ONE*. Edited by O. R. Bandapalli, 13(12), p. e0209185. doi: 10.1371/journal.pone.0209185.

Marconi, M. *et al.* (2009) 'Impact of infection on the secretory capacity of the male accessory glands', *International braz j urol*, 35(3), pp. 299–309. doi: 10.1590/S1677-55382009000300006.

Martinez, G. *et al.* (2020) 'Biallelic variants in MAATS1 encoding CFAP91, a calmodulin-associated and spoke-associated complex protein, cause severe astheno-teratozoospermia and male infertility', *Journal of Medical Genetics*, 57(10), pp. 708–716. doi: 10.1136/jmedgenet-2019-106775.

Masarani, M., Wazait, H. and Dinneen, M. (2006) 'Mumps orchitis.', *Journal of the Royal Society of Medicine*, 99(11), pp. 573–5. doi: 10.1258/jrsm.99.11.573.

Massaro, P. A. *et al.* (2015) 'Does intracytoplasmic sperm injection pose an increased risk of genitourinary congenital malformations in offspring compared to in vitro fertilization? A systematic review and meta-analysis', *Journal of Urology*, 193(5), pp. 1837–1842. doi: 10.1016/j.juro.2014.10.113.

Matzuk, M.M. and Lamb, D.J. (2008) 'The biology of infertility: research advances and clinical challenges', Nature Medicine, 14(11), pp. 1197–1213. doi:10.1038/nm.f.1895.

McLachlan, R. I. and O'Bryan, M. K. (2010) 'State of the art for genetic testing of infertile men', *Journal of Clinical Endocrinology and Metabolism*. Endocrine Society, pp. 1013–1024. doi: 10.1210/jc.2009-1925.

Meyerson, M., Gabriel, S. and Getz, G. (2010) 'Advances in understanding cancer genomes through second-generation sequencing', *Nature Reviews. Genetics*, 11(10), pp. 685–696. doi: 10.1038/nrg2841.

Meynert, A. M. *et al.* (2014) 'Variant detection sensitivity and biases in whole genome and exome sequencing', *BMC Bioinformatics*, 15(1), p. 247. doi: 10.1186/1471-2105-15-247.

Mieusset, R. *et al.* (1987) 'Hyperthermia and human spermatogenesis: enhancement of the inhibitory effect obtained by "artificial cryptorchidism"', *International Journal of Andrology*, 10(4), pp. 571–580. doi: 10.1111/j.1365-2605.1987.tb00356.x.

Mieusset, R. and B'ujan, L. (1994) 'The potential of mild testicular heating as a safe, effective and reversible contraceptive method for men', *International Journal of Andrology*, 17(4), pp. 186–191. doi: 10.1111/j.1365-2605.1994.tb01241.x.

Miller, D. T. *et al.* (2010) 'Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies', *American Journal of Human Genetics*, 86(5), pp. 749–764. doi: 10.1016/j.ajhg.2010.04.006.

Misell, L. M. *et al.* (2006) 'A stable isotope-mass spectrometric method for measuring human spermatogenesis kinetics in vivo', *Journal of Urology*, 175(1), pp. 242–246. doi: 10.1016/S0022-5347(05)00053-4.

Miyamoto, T. *et al.* (2003) 'Azoospermia in patients heterozygous for a mutation in SYCP3', *The Lancet*, 362(9397), pp. 1714–1719. doi: 10.1016/S0140-6736(03)14845-3.

Mobasseri, N. *et al.* (2018) 'Androgen receptor (AR)-CAG trinucleotide repeat length and idiopathic male infertility: A case-control trial and a meta-analysis', *EXCLI Journal*, 17, pp. 1167–1179. doi: 10.17179/excli2018-1744.

Mohandas, T. K. *et al.* (1992) 'Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: Meiotic studies in a man with a deletion of distal Xp', *American Journal of Human Genetics*, 51(3), pp. 526–533.

Mruk, D. D. and Cheng, C. Y. (2004) 'Sertoli-Sertoli and Sertoli-Germ Cell Interactions and Their Significance in Germ Cell Movement in the Seminiferous Epithelium during Spermatogenesis', *Endocrine Reviews*, 25(5), pp. 747–806. doi: 10.1210/er.2003-0022.

Mumtaz, Z., Shahid, U. and Levay, A. (2013) 'Understanding the impact of gendered roles on the experiences of infertility amongst men and women in Punjab.', *Reproductive health*, 10, p. 3. doi: 10.1186/1742-4755-10-3.

Nadeem, F., Fahim, A. and Bugti, S. (2012) 'Effects of cigarette smoking on male fertility', *Turk J Med Sci*, 42(2), pp. 1400–1405. doi: 10.3906/sag-1107-25.

Nakagawa, H. *et al.* (2015) 'Cancer whole-genome sequencing: present and future', *Oncogene*, 34(49), pp. 5943–5950. doi: 10.1038/onc.2015.90.

Nakano, K. *et al.* (2000) 'NESK, a member of the germinal center kinase family that activates the c-Jun N-terminal kinase pathway and is expressed during the late stages of embryogenesis', *Journal of Biological Chemistry*, 275(27), pp. 20533–20539. doi: 10.1074/jbc.M001009200.

Neto, F. T. L., Bach, P. V., Najari, Bobby Baback, *et al.* (2016) 'Genetics of Male Infertility', *Current Urology Reports*, 17(10), p. 70. doi: 10.1007/s11934-016-0627-x.

Neto, F. T. L., Bach, P. V., Najari, Bobby B., *et al.* (2016) 'Spermatogenesis in humans and its affecting factors', *Seminars in Cell and Developmental Biology*. Academic Press, pp. 10–26. doi: 10.1016/j.semcdb.2016.04.009.

O'Bryan, M. K. *et al.* (2013) 'RBM5 Is a Male Germ Cell Splicing Factor and Is Required for Spermatid Differentiation and Male Fertility', *PLoS Genetics*, 9(7). doi: 10.1371/journal.pgen.1003628.

O'Hara, L. and Smith, L. B. (2015) 'Androgen receptor roles in spermatogenesis and infertility', *Best Practice and Research: Clinical Endocrinology and Metabolism*. Bailliere Tindall Ltd, pp. 595–605. doi: 10.1016/j.beem.2015.04.006.

O'Hara, L. and Smith, L. B. (2017) 'The Genetics of Androgen Receptor Signalling in Male Fertility', in *Monographs in Human Genetics*. S. Karger AG, pp. 86–100. doi: 10.1159/000477280.

Olayemi, F. O. (2010) 'A review on some causes of male infertility', *African Journal of Biotechnology*, 9(20), pp. 2834–2842.

Olesen, I. A. *et al.* (2017) 'Clinical, genetic, biochemical, and testicular biopsy findings among 1,213 men evaluated for infertility', *Fertility and Sterility*, 107(1), pp. 74-82.e7. doi: 10.1016/j.fertnstert.2016.09.015.

Oud, M. S. *et al.* (2019) 'A systematic review and standardized clinical validity assessment of male infertility genes', *Human Reproduction*, 34(5), pp. 932–941. doi: 10.1093/humrep/dez022.

Oud, M S, Smits, R. M., *et al.* (2021) 'A de novo paradigm for male infertility', *bioRxiv*, p. 2021.02.27.433155. doi: 10.1101/2021.02.27.433155.

Oud, M S, Houston, B. J., *et al.* (2021) 'Exome sequencing reveals variants in known and novel candidate genes for severe sperm motility disorders', *Human Reproduction*, (deab099). doi: 10.1093/humrep/deab099.

Oud, M. S. *et al.* (2021) 'Lack of evidence for a role of PIWIL1 variants in human male infertility', *Cell*. Elsevier B.V., pp. 1941–1942. doi: 10.1016/j.cell.2021.03.001.

Palermo, G. *et al.* (1992) 'Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte', *The Lancet*, 340(8810), pp. 17–18. doi: 10.1016/0140-6736(92)92425-F.

Pasqualotto, F. F. *et al.* (2004) 'Risks and benefits of hormone replacement therapy in older men.', *Revista do Hospital das Clinicas*, 59(1), pp. 32–38.

Patat, O. *et al.* (2016) 'Truncating Mutations in the Adhesion G Protein-Coupled Receptor G2 Gene ADGRG2 Cause an X-Linked Congenital Bilateral Absence of Vas Deferens', *American Journal of Human Genetics*, 99(2), pp. 437–442. doi: 10.1016/j.ajhg.2016.06.012.

Patrizio, P. *et al.* (1993) 'Andrology: Cystic fibrosis mutations impair the fertilization rate of epididymal sperm from men with congenital absence of the vas deferens', *Human Reproduction*, 8(8), pp. 1259–1263. doi: 10.1093/oxfordjournals.humrep.a138237.

Pavone, P. *et al.* (2019) 'Microcephaly/trigonocephaly, intellectual disability, autism spectrum disorder, and atypical dysmorphic features in a boy with Xp22.31 duplication', *Molecular Syndromology*, 9(5), pp. 253–258. doi: 10.1159/000493174.

Pedersen, B. S. and Quinlan, A. R. (2017) 'Who's Who? Detecting and Resolving

Sample Anomalies in Human DNA Sequencing Studies with Peddy', *American Journal of Human Genetics*, 100(3), pp. 406–413. doi: 10.1016/j.ajhg.2017.01.017.

Pfundt, R. *et al.* (2017) 'Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders', *Genetics in Medicine*, 19(6), pp. 667–675. doi: 10.1038/gim.2016.163.

Picard Toolkit (no date). Available at: http://broadinstitute.github.io/picard/.

Pönighaus, C. *et al.* (2007) 'Human xylosyltransferase II is involved in the biosynthesis of the uniform tetrasaccharide linkage region in chondroitin sulfate and heparan sulfate proteoglycans', *Journal of Biological Chemistry*, 282(8), pp. 5201–5206. doi: 10.1074/jbc.M611665200.

Popli, K. and Stewart, J. (2007) 'Infertility and its management in men with cystic fibrosis: Review of literature and clinical practices in the UK', *Human Fertility*. Hum Fertil (Camb), pp. 217–221. doi: 10.1080/14647270701510033.

Poultney, C. S. *et al.* (2013) 'Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder', *American Journal of Human Genetics*, 93(4), pp. 607–619. doi: 10.1016/j.ajhg.2013.09.001.

Pruitt, K. D. *et al.* (2009) 'The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes', *Genome Research*, 19(7), pp. 1316–1323. doi: 10.1101/gr.080531.108.

Punab, M. *et al.* (2017) 'Causes of male infertility: A 9-year prospective monocentre study on 1737 patients with reduced total sperm counts', *Human Reproduction*, 32(1), pp. 18–31. doi: 10.1093/humrep/dew284.

Qian, M. X. *et al.* (2013) 'Acetylation-mediated proteasomal degradation of core histones during DNA repair and spermatogenesis', *Cell*, 153(5), p. 1012. doi: 10.1016/j.cell.2013.04.032.

Qin, J. *et al.* (2015) 'Assisted reproductive technology and risk of congenital malformations: a meta-analysis based on cohort studies', *Archives of Gynecology and Obstetrics*. Springer Verlag, pp. 777–798. doi: 10.1007/s00404-015-3707-0.

Quill, T. A. *et al.* (2003) 'Hyperactivated sperm motility driven by CatSper2 is required for fertilization', *Proceedings of the National Academy of Sciences of the United States of America*, 100(25), pp. 14869–14874. doi: 10.1073/pnas.2136654100.

Quinque, D. *et al.* (2006) 'Evaluation of saliva as a source of human DNA for population and association studies', *Analytical Biochemistry*, 353(2), pp. 272–277. doi: 10.1016/j.ab.2006.03.021.

Rajagopalan, R. *et al.* (2020) 'A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data', *Genome Medicine*, 12(1), pp. 1–11. doi: 10.1186/s13073-020-0712-0.

Rehm, H. L. *et al.* (2013) 'ACMG clinical laboratory standards for next-generation sequencing', *Genetics in Medicine*, 15(9), pp. 733–747. doi: 10.1038/gim.2013.92.

Reuter, M. *et al.* (2009) 'Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile', *Nature Structural and Molecular Biology*, 16(6), pp. 639–646. doi: 10.1038/nsmb.1615.

Riggs, E.R. et al. (2020) 'Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen)', Genetics in Medicine, 22(2), pp. 245–257. doi:10.1038/s41436-019-0686-8.

Ruderfer, D. M. *et al.* (2016) 'Patterns of genic intolerance of rare copy number variation in 59,898 human exomes', *Nature Genetics*, 48(10), pp. 1107–1111. doi: 10.1038/ng.3638.

Saadi, I. *et al.* (2011) 'Deficiency of the cytoskeletal protein SPECC1L leads to oblique facial clefting', *American Journal of Human Genetics*, 89(1), pp. 44–55. doi: 10.1016/j.ajhg.2011.05.023.

Safarinejad, M. R., Azma, K. and Kolahi, A. A. (2009) 'The Effects of intensive, long-term treadmill running on reproductive hormones, hypothalamus-pituitary-testis axis, and semen quality: A randomized controlled study', *Journal of Endocrinology*, 200(3), pp. 259–271. doi: 10.1677/JOE-08-0477.

Sandin, S. *et al.* (2013) 'Autism and mental retardation among offspring born after in vitro fertilization', *JAMA - Journal of the American Medical Association*, 310(1), pp. 75–84. doi: 10.1001/jama.2013.7222.

Sazci, A. *et al.* (2005) 'Male factor infertility associated with a familial translocation t(1;13)(q24;q10)', *Fertility and Sterility*, 83(5), p. 1548.e19-1548.e21. doi: 10.1016/j.fertnstert.2004.10.055.

Schreml, J. *et al.* (2014) 'The missing "link": An autosomal recessive short stature syndrome caused by a hypofunctional XYLT1 mutation', *Human Genetics*, 133(1), pp. 29–39. doi: 10.1007/s00439-013-1351-y.

Schultz, N., Hamra, F. K. and Garbers, D. L. (2003) 'A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets', *Proceedings of the National Academy of Sciences*, 100(21), pp. 12201–12206. doi: 10.1073/pnas.1635054100.

Schwarze, K. *et al.* (2018) 'Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature', *Genetics in Medicine*. Nature Publishing Group, pp. 1122–1130. doi: 10.1038/gim.2017.247.

Seabra, C. M. *et al.* (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. doi: 10.1016/j.rbmo.2014.04.017.

Seabra, C. M. *et al.* (2015) 'The mutational spectrum of WT1 in male infertility', *Journal of Urology*, 193(5), pp. 1709–1715. doi: 10.1016/j.juro.2014.11.004.

Sebat, J. *et al.* (2007) 'Strong association of de novo copy number mutations with autism', *Science*, 316(5823), pp. 445–449. doi: 10.1126/science.1138659.

Sedjaï, F. *et al.* (2010) 'Control of ciliogenesis by FOR20, a novel centrosome and pericentriolar satellite protein', *Journal of Cell Science*, 123(14), pp. 2391–2401. doi: 10.1242/jcs.065045.

Sha, Y. W. *et al.* (2017) 'A homozygous CEP135 mutation is associated with multiple morphological abnormalities of the sperm flagella (MMAF)', *Gene*, 633, pp. 48–53. doi: 10.1016/j.gene.2017.08.033.

Sha, Y.-W. *et al.* (2017) 'Novel Mutations in *CFAP44* and *CFAP43* Cause Multiple Morphological Abnormalities of the Sperm Flagella (MMAF)', *Reproductive Sciences*, p. 193371911774975. doi: 10.1177/1933719117749756.

Shaffer, L. G. (2005) 'American College of Medical Genetics guideline on the cytogenetic evaluation of the individual with developmental delay or mental retardation', *Genetics in Medicine*. Nature Publishing Group, pp. 650–654. doi: 10.1097/01.gim.0000186545.83160.1e.

Sharma, R. and Agarwal, A. (2011) 'Spermatogenesis: An Overview', in *Sperm Chromatin*. New York, NY: Springer New York, pp. 19–44. doi: 10.1007/978-1-4419-6857-9_2.

Shaw-Smith, C. *et al.* (2004) 'Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features', *Journal of Medical Genetics*, 41(4), pp. 241–248. doi: 10.1136/jmg.2003.017731.

Shevell, M. *et al.* (2003) 'Practice parameter: Evaluation of the child with global developmental delay: Report of the quality standards subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society', *Neurology*, 60(3), pp. 367–380. doi: 10.1212/01.WNL.0000031431.81555.16.

Skaletsky, H. *et al.* (2003) 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes', *Nature*, 423(6942), pp. 825–837. doi: 10.1038/nature01722.

Smith, K. D. *et al.* (1965) 'A Familial Centric Chromosome Fragment', *Cytogenetic and Genome Research*, 4(4–5), pp. 219–226. doi: 10.1159/000129858.

Sokol, R. Z. (2001) 'Infertility in men with cystic fibrosis', *Current Opinion in Pulmonary Medicine*. Lippincott Williams and Wilkins, pp. 421–426. doi: 10.1097/00063198-200111000-00011.

Song, H. *et al.* (2020) 'The function of pre-mRNA alternative splicing in mammal spermatogenesis', *International Journal of Biological Sciences*. Ivyspring International Publisher, pp. 38–48. doi: 10.7150/ijbs.34422.

de Souza, D. A. S. *et al.* (2018) 'Congenital bilateral absence of the vas deferens as an atypical form of cystic fibrosis: reproductive implications and genetic counseling', *Andrology*. Blackwell Publishing Ltd, pp. 127–135. doi: 10.1111/andr.12450.

Van Steirteghem, A. (2012) 'Celebrating ICSI's twentieth anniversary and the birth of more than 2.5 million children-the "how, why, when and where"', *Human Reproduction*. Oxford University Press, pp. 1–2. doi: 10.1093/humrep/der447.

Steptoe, P. C. and Edwards, R. G. (1978) 'Birth after the reimplantation of a human embryo', *Lancet*. Elsevier, p. 366. doi: 10.1016/s0140-6736(78)92957-4.

Stouffs, K. *et al.* (2012) 'Array comparative genomic hybridization in male infertility', *Human Reproduction*, 27(3), pp. 921–929. doi: 10.1093/humrep/der440.

Sun, C. *et al.* (1999) 'An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y', *Nature Genetics*, 23(4), pp. 429–432. doi: 10.1038/70539.

Sun, F. *et al.* (2007) 'Abnormal progression through meiosis in men with nonobstructive azoospermia', *Fertility and Sterility*, 87(3), pp. 565–571. doi: 10.1016/j.fertnstert.2006.07.1531.

Suntharalingham, J. P. *et al.* (2015) 'DAX-1 (NR0B1) and steroidogenic factor-1 (SF-1, NR5A1) in human disease', *Best Practice and Research: Clinical Endocrinology and Metabolism*. Bailliere Tindall Ltd, pp. 607–619. doi: 10.1016/j.beem.2015.07.004.

Tan, R. *et al.* (2014) 'An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data', *Human Mutation*, 35(7), pp. 899–907. doi: 10.1002/humu.22537.

Tan, Y.-Q. *et al.* (2019) 'Loss-of-function mutations in TDRD7 lead to a rare novel syndrome combining congenital cataract and nonobstructive azoospermia in humans', *Genetics in Medicine*, 21(5), pp. 1209–1217. doi: 10.1038/gim.2017.130.

Tang, S. *et al.* (2017) 'Biallelic Mutations in CFAP43 and CFAP44 Cause Male Infertility with Multiple Morphological Abnormalities of the Sperm Flagella.', *American journal of human genetics*, 100(6), pp. 854–864. doi: 10.1016/j.ajhg.2017.04.012.

Teich, N. *et al.* (2006) 'Mutations of human cationic trypsinogen (PRSS1) and chronic pancreatitis', *Human Mutation*. NIH Public Access, pp. 721–730. doi: 10.1002/humu.20343.

Ten, J. *et al.* (2008) 'Occupational and Lifestyle Exposures on Male Infertility : A Mini Review', *The Open Reproductive Science Journal*, 1, pp. 16–21. doi: 10.2174/1874255600801010016.

Teo, S. M. *et al.* (2012) 'Statistical challenges associated with detecting copy number variations with next-generation sequencing', *Bioinformatics*. Oxford Academic, pp. 2711–2718. doi: 10.1093/bioinformatics/bts535.

Thomas, N. S. and Hassold, T. J. (2003) 'Aberrant recombination and the origin of Klinefelter syndrome', *Human Reproduction Update*. Hum Reprod Update, pp. 309–317. doi: 10.1093/humupd/dmg028.

Thorvaldsdottir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration', *Briefings in Bioinformatics*, 14(2), pp. 178–192. doi: 10.1093/bib/bbs017.

Tong, Z. *et al.* (2020) 'Whole-exome sequencing reveals potential mechanisms of drug resistance to FGFR3-TACC3 targeted therapy and subsequent drug selection: Towards a personalized medicine', *BMC Medical Genomics*, 13(1), p. 138. doi:

10.1186/s12920-020-00794-x.

Touré, A. *et al.* (2020) 'The genetic architecture of morphological abnormalities of the sperm tail', *Human Genetics*. Springer. doi: 10.1007/s00439-020-02113-x.

Tournaye, H., Krausz, C. and Oates, R. D. (2017) 'Novel concepts in the aetiology of male reproductive impairment', *The Lancet Diabetes and Endocrinology*. Lancet Publishing Group, pp. 544–553. doi: 10.1016/S2213-8587(16)30040-7.

Tremblay, M. S., Copeland, J. L. and Van Helder, W. (2004) 'Effect of training status and exercise mode on endogenous steroid hormones in men', *Journal of Applied Physiology*, 96(2), pp. 531–539. doi: 10.1152/japplphysiol.00656.2003.

Tsuchida, N. *et al.* (2018) 'Detection of copy number variations in epilepsy using exome data', *Clinical Genetics*, 93(3), pp. 577–587. doi: 10.1111/cge.13144.

Tüttelmann, F. *et al.* (2011) 'Clinical experience with azoospermia: Aetiology and chances for spermatozoa detection upon biopsy', *International Journal of Andrology*, 34(4 PART 1), pp. 291–298. doi: 10.1111/j.1365-2605.2010.01087.x.

Tüttelmann, Frank *et al.* (2011) 'Copy Number Variants in Patients with Severe Oligozoospermia and Sertoli-Cell-Only Syndrome', *PLoS ONE*. Edited by L. Orban, 6(4), p. e19426. doi: 10.1371/journal.pone.0019426.

Tüttelmann, F., Ruckert, C. and Röpke, A. (2018) 'Disorders of spermatogenesis: Perspectives for novel genetic diagnostics after 20 years of unchanged routine'. doi: 10.1007/s11825-018-0181-7.

Tyler-Smith, C. and Krausz, C. (2009) 'The Will-o'-the-Wisp of Genetics — Hunting for the Azoospermia Factor Gene', *New England Journal of Medicine*, 360(9), pp. 925–927. doi: 10.1056/nejme0900301.

Uhlén, M. *et al.* (2015) *Proteomics. Tissue-based map of the human proteome., Science (New York, N.Y.)*. American Association for the Advancement of Science (6220). doi: 10.1126/science.1260419.

Untergasser, A. *et al.* (2007) 'Primer3Plus, an enhanced web interface to Primer3', *Nucleic Acids Research*, 35(SUPPL.2), p. W71. doi: 10.1093/nar/gkm306.

Vagin, V. V. *et al.* (2009) 'Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members', *Genes and Development*, 23(15), pp. 1749–1762. doi: 10.1101/gad.1814809.

Veltman, J. A. and Brunner, H. G. (2012) 'De novo mutations in human genetic disease', *Nature Publishing Group*. doi: 10.1038/nrg3241.

Verploegen, S. *et al.* (2005) 'Characterization of the role of CaMKI-like kinase (CKLiK) in human granulocyte function', *Blood*, 106(3), pp. 1076–1083. doi: 10.1182/blood-2004-09-3755.

Viñas-Jornet, M. *et al.* (2018) 'High Incidence of Copy Number Variants in Adults with Intellectual Disability and Co-morbid Psychiatric Disorders', *Behavior Genetics*, 48(4), pp. 323–336. doi: 10.1007/s10519-018-9902-6.

Vincent, M. C. *et al.* (2002) 'Cytogenetic investigations of infertile men with low sperm counts: A 25-year experience', *Journal of Andrology*. American Society of Andrology Inc., pp. 18–22. doi: 10.1002/j.1939-4640.2002.tb02597.x.

Vissers, L. E., Gilissen, C. and Veltman, J. A. (2016) 'Genetic studies in intellectual disability and related disorders', *Nat Rev Genet*, 17(1), pp. 9–18. doi: 10.1038/nrg3999.

Vissers, L. E. L. M., Gilissen, C. and Veltman, J. A. (2016) 'Genetic studies in intellectual disability and related disorders', *Nature Reviews Genetics*, 17(1), pp. 9–18. doi: 10.1038/nrg3999.

Vockel, M. *et al.* (2019) 'The X chromosome and male infertility', *Human Genetics*. Springer, pp. 203–215. doi: 10.1007/s00439-019-02101-w.

Vogt, P. H. *et al.* (1996) 'Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11.', *Human molecular genetics*, 5(7), pp. 933–43.

Wang, Y. *et al.* (2016) 'Genomic copy number variation association study in Caucasian patients with nonsyndromic cryptorchidism', *BMC Urology*, 16(1), p. 62. doi: 10.1186/s12894-016-0180-4.

Wickham, H. *et al.* (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*, 4(43), p. 1686. doi: 10.21105/joss.01686.

Wischmann, T. and Thorn, P. (2013) '(Male) infertility: What does it mean to men? New evidence from quantitative and qualitative studies', in *Reproductive BioMedicine Online*. Elsevier, pp. 236–243. doi: 10.1016/j.rbmo.2013.06.002.

Wong, M. L. *et al.* (2017) 'The PHF21B gene is associated with major depression and modulates the stress response', *Molecular Psychiatry*, 22(7), pp. 1015–1025. doi: 10.1038/mp.2016.174.

World Health Organization (2019) 'WHO | International Classification of Diseases, 11th Revision (ICD-11)', *WHO*.

World Health Organization, D. of R. H. and R. (2010) 'WHO laboratory manual for the examination and processing of human semen', *WHO*.

Xavier, M. J. *et al.* (2020) 'Disease gene discovery in male infertility: past, present and future', *Human Genetics*. Springer, pp. 1–13. doi: 10.1007/s00439-020-02202-x.

Xiao, Y. *et al.* (2012) 'Cre-Mediated Stress Affects Sirtuin Expression Levels, Peroxisome Biogenesis and Metabolism, Antioxidant and Proinflammatory Signaling Pathways', *PLoS ONE*. Edited by M. G. Bonini, 7(7), p. e41097. doi: 10.1371/journal.pone.0041097.

Xu, J. *et al.* (2017) 'A novel functional variant in Wilms' Tumor 1 (WT1) is associated with idiopathic non-obstructive azoospermia', *Molecular Reproduction and Development*, 84(3), pp. 222–228. doi: 10.1002/mrd.22768.

Yang, F. *et al.* (2015) ' TEX 11 is mutated in infertile men with azoospermia and regulates genome-wide recombination rates in mouse ', *EMBO Molecular Medicine*, 7(9), pp. 1198–1210. doi: 10.15252/emmm.201404967.

Yang, Y. *et al.* (2013) 'Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders', *New England Journal of Medicine*, 369(16), pp. 1502–1511. doi: 10.1056/nejmoa1306555.

Yao, R. *et al.* (2017) 'Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data', *Molecular Cytogenetics*, 10(1), pp. 30–30. doi: 10.1186/s13039-017-0333-5.

Yatsenko, A. N. *et al.* (2015) 'X-Linked TEX11 Mutations, Meiotic Arrest, and Azoospermia in Infertile Men', *New England Journal of Medicine*, 372(22), pp. 2097–2107. doi: 10.1056/NEJMoa1406192.

Zenteno-Ruiz, J. C., Kofman-Alfaro, S. and Méndez, J. P. (2001) '46, XX sex reversal', *Archives of Medical Research*. Elsevier, pp. 559–566. doi: 10.1016/S0188-4409(01)00322-8.

Zhang, X. *et al.* (2019) 'A novel homozygous CFAP65 mutation in humans causes male infertility with multiple morphological abnormalities of the sperm flagella', *Clinical Genetics*, 96(6), pp. 541–548. doi: 10.1111/cge.13644.

Zhang, Y. *et al.* (2007) 'Sensorineural deafness and male infertility: a contiguous gene deletion syndrome', *Journal of Medical Genetics*, 44(4), pp. 233–240. doi: 10.1136/jmg.2006.045765.

Zhao, H. *et al.* (2012) 'A genome-wide association study reveals that variants within the HLA region are associated with risk for nonobstructive azoospermia', *American Journal of Human Genetics*, 90(5), pp. 900–906. doi: 10.1016/j.ajhg.2012.04.001.

Zhao, M. *et al.* (2013) 'Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives', *BMC Bioinformatics*, 14(SUPPL11), p. S1. doi: 10.1186/1471-2105-14-S11-S1.

Zhu, Q. *et al.* (2015) 'The impact of DNA input amount and DNA source on the performance of whole-exome sequencing in cancer epidemiology', *Cancer Epidemiology Biomarkers and Prevention*, 24(8), pp. 1207–1213. doi: 10.1158/1055-9965.EPI-15-0205.

Zhu, X.-B. *et al.* (2016) 'Association of a TDRD1 variant with spermatogenic failure susceptibility in the Han Chinese', *Journal of Assisted Reproduction and Genetics*, 33(8), pp. 1099–1104. doi: 10.1007/s10815-016-0738-9.

Zhuang, J. *et al.* (2019) 'A prenatal diagnosis and genetics study of five pedigrees in the Chinese population with Xp22.31 microduplication', *Molecular Cytogenetics*, 12(1). doi: 10.1186/s13039-019-0461-1.

Zou, S. W. *et al.* (2003) 'Expression and localization of VCX/Y proteins and their possible involvement in regulation of ribosome assembly during spermatogenesis.', *Cell research*, 13(3), pp. 171–177. doi: 10.1038/sj.cr.7290161.